

RESEARCH ARTICLE

Optimizing Bi-LSTM networks for improved lung cancer detection accuracy

Su Diao¹, Yajie Wan², Danyi Huang³, Shijia Huang⁴, Touseef Sadiq^{5*},
 Mohammad Shahbaz Khan⁶, Lal Hussain^{7,8*}, Badr S. Alkahtani⁹, Tehseen Mazhar^{10,11}

1 Department of Industrial & Systems Engineering, Auburn University, Auburn, Alabama, United States of America, **2** Department of Computer Science, Brown University, Providence, RI, United States of America, **3** Department of Chemical Engineering, Columbia University, New York City, NY, United States of America, **4** Fu Foundation School of Engineering and Applied Science, Fu Foundation School of Engineering and Applied Science, Columbia University, New York, NY, United States of America, **5** Department of Information and Communication Technology, Centre for Artificial Intelligence Research (CAIR), University of Agder, Grimstad, Norway, **6** Children's National Hospital, Washington, DC, United States of America, **7** Department of Computer Science and Information Technology, The University of Azad Jammu and Kashmir, Chattar Kalas Campus, Muzaffarabad, Pakistan, **8** Department of Computer Science, Neelum Campus, The University of Azad Jammu and Kashmir, Azad Kashmir, Pakistan, **9** Department of Mathematics, King Saud University, Riyadh, Saudi Arabia, **10** School of Computer Science, National College of Business Administration and Economics, Lahore, Pakistan, **11** Department of Computer Science and Information Technology, School Education Department, Government of Punjab, Layyah, Pakistan

* touseef.sadiq@uia.no (TS); lall_hussain2008@live.com (LH)



OPEN ACCESS

Citation: Diao S, Wan Y, Huang D, Huang S, Sadiq T, Khan MS, et al. (2025) Optimizing Bi-LSTM networks for improved lung cancer detection accuracy. PLoS ONE 20(2): e0316136. <https://doi.org/10.1371/journal.pone.0316136>

Editor: Jawad Rasheed, Istanbul Sabahattin Zaim University: Istanbul Sabahattin Zaim Universitesi, TÜRKİYE

Received: September 30, 2024

Accepted: December 5, 2024

Published: February 24, 2025

Copyright: © 2025 Diao et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The dataset for this study was obtained from a publicly accessible web-based repository maintained by LCA the sole national non-DICOM format images from 76 patients selected image are accessible at <https://www.cancerimagingarchive.net/collection/nsclc-radiogenomics/>. The dataset is utilized and published in our previous studies [46–48].

Funding: The author B.A. extend his appreciation to research Supporting Project number

Abstract

Lung cancer remains a leading cause of cancer-related deaths worldwide, with low survival rates often attributed to late-stage diagnosis. To address this critical health challenge, researchers have developed computer-aided diagnosis (CAD) systems that rely on feature extraction from medical images. However, accurately identifying the most informative image features for lung cancer detection remains a significant challenge. This study aimed to compare the effectiveness of both hand-crafted and deep learning-based approaches for lung cancer diagnosis. We employed traditional hand-crafted features, such as Gray Level Co-occurrence Matrix (GLCM) features, in conjunction with traditional machine learning algorithms. To explore the potential of deep learning, we also optimized and implemented a Bidirectional Long Short-Term Memory (Bi-LSTM) network for lung cancer detection. The results revealed that the highest performance using hand-crafted features was achieved by extracting GLCM features and utilizing Support Vector Machine (SVM) with different kernels, reaching an accuracy of 99.78% and an AUC of 0.999. However, the deep learning Bi-LSTM network surpassed both methods, achieving an accuracy of 99.89% and an AUC of 1.0000. These findings suggest that the proposed methodology, combining hand-crafted features and deep learning, holds significant promise for enhancing early lung cancer detection and ultimately improving diagnosis systems.

(RSPD2025R526), King Saud University, Riyadh, Saudi Arabia.

Competing interests: The authors have declared that no competing interests exist.

1. Introduction

Lung cancer remains the most commonly diagnosed cancer and the leading cause of cancer death globally, particularly among men. According to recent estimates of 2024, approximately 2.5 million cases of lung cancer occur worldwide each year, accounting for roughly one in eight cancers. Tragically, 1.8 million deaths are attributed to lung cancer annually, representing one in five cancer-related deaths with non-small cell lung cancer (NSCLC) constituting roughly 85% of these diagnoses [1, 2]. Early detection of NSCLC is crucial for improved outcomes. Two primary treatment approaches for NSCLC include radiofrequency ablation and stereotactic body radiotherapy (SBRT).

Lung cancer is primarily categorized into two main types: small cell lung carcinoma (SCLC) and non-small cell lung carcinoma (NSCLC). These subtypes differ significantly in their biological behavior, including growth patterns, metastatic potential, and response to treatment. While NSCLC generally progresses more slowly, SCLC is highly aggressive, forming tumors rapidly and spreading to other parts of the body. Cigarette smoking is a well-established primary risk factor for lung cancer, with the cumulative exposure directly correlated to disease risk [3]. Lung cancer remains a leading global cause of cancer-related deaths, claiming over 1.4 million lives annually [4, 5]. Accurate and timely diagnosis is crucial for effective lung cancer treatment. Pathologists play a vital role in this process by carefully examining microscopic tissue samples, known as histopathology slides, to confirm the presence of lung cancer and determine its specific type and subtype [6, 7]. Adenocarcinoma and squamous cell carcinoma are the two primary subtypes of non-small cell lung cancer, distinguished through meticulous pathological analysis [8, 9].

SCLC is an aggressive form of lung cancer strongly linked to cigarette smoking. Its development is influenced by genetic factors such as proto-oncogenes, autocrine growth loops, and tumor-suppressor genes. These biological differences result in distinct treatment approaches and prognoses for SCLC compared to NSCLC. Early detection is crucial for improving NSCLC survival rates, which can range from 35% to 85% depending on tumor stage and type [10–12]. Unfortunately, late diagnosis remains common, leading to a five-year survival rate of only 16% for NSCLC. While initial chemotherapy responses for SCLC exceed 60%, the disease's rapid recurrence results in a dismal five-year survival rate of 6%, with little improvement over the past two decades. Cigarette smoking is the primary risk factor for approximately 90% of lung cancer cases. However, other factors, including exposure to air pollution, radon, asbestos, and chronic infections, also contribute to the development of lung cancer. Additionally, genetic predisposition plays a role, with both inherited and acquired genetic alterations influencing susceptibility to the disease. Lung cancer treatment options are diverse and include surgery, radiation therapy, chemotherapy, and targeted therapies. The optimal treatment approach is determined by individual patient characteristics and disease stage [13].

The discovery of X-rays and radiation in the late 19th century revolutionized medical imaging, enabling non-invasive visualization of internal structures and paving the way for radiation-based cancer treatments. By the 1960s, the collaborative efforts of radiologists and surgeons, combined with advancements in computer technology, led to the accumulation of substantial cancer data. In the subsequent decades, medical imaging has undergone significant advancements. Computed tomography (CT) scans provide detailed cross-sectional images of the lungs and tumors, making them invaluable for treatment planning. Chest CT scans are crucial for staging lung cancer, while abdominal CT scans aid in detecting metastatic disease [14]. Positron emission tomography (PET) scans utilize a radioactive tracer to visualize metabolically active tissues, such as cancer cells. This modality is indispensable for assessing tumor spread to lymph nodes and distant organs [15]. Magnetic resonance imaging (MRI) excels in visualizing soft

tissues, making it particularly useful for evaluating brain metastases in lung cancer patients [16]. These imaging modalities have become essential tools in the diagnosis and management of lung cancer, providing crucial information for treatment decisions and prognosis.

2. Literature background

The advent of powerful computing systems has accelerated the development of deep learning (DL), a subset of artificial intelligence (AI) that employs complex neural networks. Convolutional neural networks (CNNs), a cornerstone of DL, have surpassed traditional neural networks by utilizing significantly deeper architectures [17]. This architecture has enabled remarkable achievements in various domains, including image recognition and segmentation [18, 19], big data analysis [20, 21], speech recognition [22], and complex tasks such as machine translation, autonomous vehicles, and agricultural disease detection [23, 24]. Inspired by the human brain's structure, CNNs consist of interconnected layers of neurons. These models are trained using supervised learning techniques. In this study, we leveraged transfer learning, applying pre-trained CNN models to a new dataset of lung cancer images. Effective feature extraction and data preparation are essential for optimal CNN performance. To this end, we extracted traditional hand-crafted features like GLCM and Haralick textures from lung cancer (NSCLC and SCLC) patient data. These features were optimized and subsequently employed with machine learning algorithms, including Naive Bayes, decision trees, and SVM with various kernels. To explore the potential of deep learning further, we implemented LSTM networks for lung cancer detection.

The discovery of radiation and X-rays in the late 19th century revolutionized medical imaging, enabling physicians to examine the human body and explore non-surgical cancer treatments. Collaborative efforts between radiologists and surgeons, coupled with the advent of computers in 1968, led to the accumulation of significant cancer data. Over the past five decades, substantial research has been dedicated to this field. To assess the stage of lung cancer, various imaging modalities are employed. Computed Tomography (CT) scans, which provide detailed images of the anatomy and lung tumors, are crucial for treatment planning [25]. Chest CT scans are essential for cancer staging, while abdominal CT scans are used to identify secondary tumors and metastases. Positron Emission Tomography (PET) scans, which utilize radioactive sugar to detect rapidly metabolizing cancer cells, are invaluable for identifying the spread of cancer to lymph nodes or other organs [26]. Magnetic Resonance Imaging (MRI) scans are another valuable tool, particularly for examining the brain. Brain MRI scans can help determine the extent of tumor spread within the brain [16].

X-rays provide valuable insights into the functional and structural aspects of the human body. However, the radiation exposure associated with CT scans can compromise image quality. To address this, researchers have explored machine learning (ML) techniques to improve the analysis and interpretation of medical images. ML techniques have demonstrated significant potential in enhancing the accuracy and precision of lung cancer prognosis and prediction, often surpassing traditional expert-based systems and statistical methods [27]. Computer-aided diagnosis (CAD) systems have emerged as powerful tools for characterizing and identifying various lung lesions. These systems can directly extract relevant features from X-ray images, enabling the classification of lung tumors as benign or malignant [28]. Several imaging modalities are employed to assess the stage of lung cancer. CT scans of the chest and abdomen provide detailed images of lung tumors and anatomy, aiding in treatment planning. Chest CT scans are crucial for cancer staging, while abdominal CT scans help identify secondary tumors and metastases. PET scans, which utilize radioactive sugar to detect rapidly

metabolizing cancer cells, are useful for identifying the spread of cancer to lymph nodes or other organs.

Recent advancements in machine learning have led to significant applications in medical diagnostics, including lung disease prediction. To assess feature importance, researchers have employed various standard toolkits [29]. Additionally, machine learning algorithms have been developed to predict the need for CT exams in emergency departments (EDs), aiding in resource allocation and patient flow management [30]. Electronic nose (e-nose) technology has emerged as a promising tool for distinguishing between healthy individuals and those with chronic obstructive pulmonary disease (COPD) by analyzing volatile organic compounds in exhaled breath [31, 32]. Liquid biopsy, which involves analyzing circulating biomarkers such as cell-free DNA, tumor cells, and microRNAs, has shown potential for diagnosing and detecting lung cancer [33]. Combining multiple biomarkers with computational tools can further enhance the accuracy of lung cancer diagnosis and prognosis. Blood cadmium levels have been identified as a potential biomarker for lung cancer, particularly in former smokers [34]. Machine learning techniques, combined with feature extraction and selection, have been applied to improve the accuracy of lung cancer detection in electronic health records, leading to enhanced diagnosis and treatment [35–37]. Researchers have explored the use of machine learning techniques in conjunction with the Internet of Things (IoT) to improve lung cancer prediction [38]. E-nose systems have also been developed to analyze exhaled breath and classify patients based on their respiratory health status, including COPD, lung cancer, and asthma [38, 39]. Machine learning has a long history of application in medical research. For instance, it has been used to analyze brain levels of polyamines and histamine in individuals exposed to extreme conditions [40].

Chaunzwa et al. developed a supervised CNN predictor based on the VGG network to forecast early-stage ADC and SCC in lung cancer patients [41]. While validated using real-time data from Massachusetts General Hospital, the model achieved a modest 71% AUC, likely due to limitations in preprocessing and image segmentation. Chaturvedi et al. conducted a comprehensive review of contemporary lung cancer detection and classification strategies [42]. Recent approaches commonly employ supervised learning algorithms, such as SVM, KNN, and CNN, often utilizing standard datasets like LIDCIDRI, LUNA 16, and Super Bowl Dataset 2016 [45]. Masood et al. developed a computer-aided prognosis (CAD) system using a 3D-DCNN approach, trained and validated on LUNA16, ANODE09, and LIDC-IDRI datasets [43]. El Nabi et al. proposed a novel deep learning method combining binary particle swarm optimization with a decision tree (BPSO-DT) and a convolutional neural network to classify different cancer types based on tumor RNA sequences [44]. The study evaluated performance metrics such as recall, precision, and F1-score. Abdulghani and Al Ahmad explored label-free approaches for cell categorization, utilizing optical profiles and Prony methodologies [51]. He et al. identified signature genes by enhancing tobacco exposure pattern (TEP) classification models and analyzing their interconnections at various biological stages. TTZ, a novel technique for separating core properties, was integrated into the TEP classification model as an input parameter [45]. Using two independent LUAD datasets, 34 genes were identified as tobacco-related mutation classic genes. The TEP classification model achieved a precision of 94.65% for the training dataset and 91.85% for the testing dataset.

The medical imaging problem requires the better quality images for improved analysis. We first improved the quality of images using robust image enhancement techniques. The machine learning algorithms requires the static features extraction relevant to the problem. Moreover, the performance of machine learning and deep learning LSTM models requires to the hyperparameters optimization. Recent studies utilized fewer texture, statistical and morphological features on this dataset to train the machine learning classifiers. Moreover, the

traditional machine learning algorithms with default parameters were utilized. In this study, we first extracted the most robust GLCM based features to capture the more relevant information from lung cancer imaging types to properly distinguish these types. The CNN models are computationally complex and unable to implement on edge devices. We thus, optimized the parameters of machine learning and deep learning LSTM algorithms using Bayesian optimization a more powerful parameters optimization method than random and grid search method. We also conducted a computational performance analysis. The results demonstrate that our proposed approach achieved the highest detection performance.

3. Materials and methods

3.1. Dataset

The dataset for this study was obtained from a publicly accessible web-based repository maintained by LCA the sole national non-DICOM format images from 76 patients utilized in our previous studies in [46–48]. Of these, 568 images belong to SCLC subjects, while 377 images represent NSCLC subjects from CT images.

3.2. Pre-processing

Following image pre-processing methods were utilized on lung cancer images.

3.2.1. Image resize. We used, "inter area" is a type of interpolation that is used to resize images in a way that produces smooth, accurate results. In computer vision, interpolation is a method of estimating the value of a pixel in an image based on the values of surrounding pixels [49]. The "inter area" option specifies that the interpolation should be performed using the area-based method. In the area-based method, the value of the pixel is calculated based on the average value of the pixels in the area surrounding it. This method is typically used for resizing images, where the goal is to reduce the size of the image by reducing the number of pixels. Because the area-based method considers the values of multiple pixels, it can produce smoother, more accurate results than other interpolation methods.

3.2.2. Data augmentation. Data augmentation is a method of creating additional data samples from existing ones in order to artificially increase the size of a dataset [50]. This can be useful when training machine learning models, especially when the available dataset is small or not representative of the problem being addressed. There are various techniques for data augmentation, including adding noise to the data, applying transformations to the data, and generating synthetic data by combining or modifying existing samples. Data augmentation can improve the generalization of a model by introducing variations in the training data that the model may encounter in the real world and can also help prevent overfitting.

3.3. Hand-crafted features extraction

This study commenced by extracting hand-crafted features from lung cancer images. Specifically, GLCM features were computed, as detailed in references [25–30]. Subsequently, supervised machine learning algorithms were applied for classification. These algorithms included Naive Bayes, decision trees, and SVM with diverse kernels. Comprehensive details of these algorithms can be found in references [51–56].

3.4. Deep learning Long Short-Term Memory (LSTM)

The LSTM networks, introduced by Hochreiter et al. [57] have become a prominent architecture within RNNs. Particularly adept at handling time series data, LSTMs excel at capturing dependencies across both short and long temporal scales [58]. Unlike traditional RNNs, which

suffer from the vanishing gradient problem [59], LSTMs mitigate this issue through the introduction of a cell state—a mechanism for preserving information over extended periods. This cell state, often visualized as a horizontal line in diagrams [60], is regulated by input, forget, and output gates. These gates regulate the flow of information into, out of, and within the cell state, enabling LSTMs to effectively capture long-term dependencies in sequential data. [57, 61]. Consequently, LSTMs are widely adopted for time series forecasting due to their capacity to capture both short-term and long-term patterns. Fig 1 illustrates the LSTM architecture.

Memory cell's role: "The memory cell, considered LSTM's core innovation, stores past information relevant to the current state."

Forget gate: "The forget state f_t first identifies irrelevant information from the previous state. A sigmoid function is then applied to these values, essentially determining how much of this information to forget (values closer to 0 are forgotten, while values closer to 1 are retained).

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

The second step focuses on identifying which new data is crucial to store in the cell state. An additional sigmoid layer, called the "input gate layer," determines the importance of the current input. This information is then combined with the forget gate's output to control how much new data gets added to the cell. Then, a \tanh function is used to create a vector \tilde{c}_t of novel values that ought to be updated embarked upcoming state.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (3)$$

The new candidate values for updating are represented by

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \quad (4)$$

The final step determines the output of the cell. This step involves two parts:

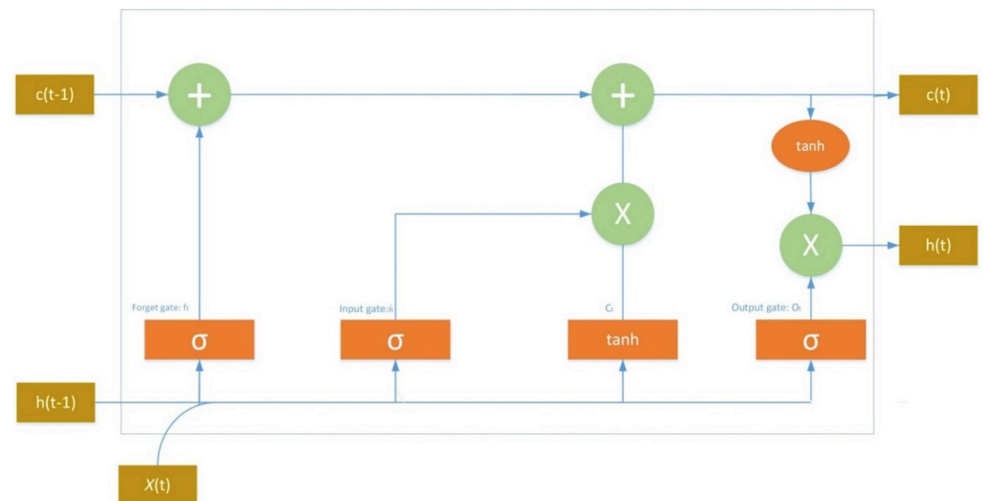


Fig 1. LSTM architecture gates: σ : Sigmoid activation function, \tanh : Hyperbolic tangent activation function, Operations: +: Addition, \times : Multiplication, States: $x(t)$: input state, $c(t)$: Current cell state, $c(t-1)$: previous cell state, $h(t)$: Current hidden state, $h(t-1)$: previous hidden state.

<https://doi.org/10.1371/journal.pone.0316136.g001>

1. A sigmoid function acts as an output gate, filtering the cell state information. This ensures only relevant information contributes to the output.
2. The filtered cell state is then multiplied by the output of another function (potentially a tanh function) to create the final output value.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * \tanh(c_t) \quad (6)$$

In Eqs (1–5), W_b , W_f , W_o , W_c shows how the weight matrices and bias vectors are represented by b_b , b_f , b_o , b_c vectors.

Our LSTM model comprised a single layer with 48 neurons and employed the ReLU activation function [62, 63]. Hyperparameters were optimized to maximize model performance as reflected in the section 3.7.

3.5. Bi-LSTM

Long Short-Term Memory (LSTM) networks are an extension of Recurrent Neural Networks (RNNs) designed to address the vanishing and exploding gradient problems inherent in traditional RNNs. LSTMs employ a sophisticated architecture with three gates (input, forget, and output) and a cell memory state. This structure, illustrated in Fig 3, enables LSTMs to effectively capture long-term dependencies in sequential data. The mathematical expressions governing the behavior of these gates and the cell memory state are as follows

$$F_g = \sigma(W_F \cdot h_{t-1} + R_F K_d + B_F) \quad (7)$$

$$I_g = \sigma(W_I \cdot h_{t-1} + R_I K_d + B_I) \quad (8)$$

$$O_g = \sigma(W_O \cdot h_{t-1} + R_O K_d + B_O) \quad (9)$$

$$C_g = F_g * C_{g-1} + I_g * \tanh(W_C \cdot h_{t-1} + R_C K_d + B_C) \quad (10)$$

Here, W_F , W_I , W_O , W_C and B_F , B_I , B_O , B_C represent the weight matrices and biases for the forget gate, input gate, output gate, and cell memory state, respectively. The σ symbol denotes the sigmoid function, and $*$ represents element-wise multiplication. C_g represents the cell memory gate, and h_{t-1} represents the previous hidden vector. R_F , R_I , R_O , R_C denote the correlation coefficients, and K_d represents the k-space data fed to the LSTM layer.

In our proposed model, we employ a Bidirectional LSTM (Bi-LSTM) as reflected in Fig 2 that combines the information from both forward and backward layers. For each time step, the forward LSTM layer calculates the hidden vector Fh_t based on the previous hidden vector Fh_{t-1} and the input k-space data K_d . Similarly, the backward LSTM layer calculates the hidden vector Bh_t based on the opposite previous hidden vector Bh_{t-1} and the input k-space data K_d . Finally, the forward hidden vector Fh_t and the backward hidden vector Bh_t are fused to obtain the final hidden vector for the Bi-LSTM model. The backward and forward hidden layer vectors are represented as $Bh_1, Bh_2, Bh_3, \dots, Bh_n$ and $Fh_1, Fh_2, Fh_3, \dots, Fh_m$, respectively. The final hidden vector of the Bi-LSTM model is given by Eq (11).

$$h_t = [Fh_t, Bh_t] \quad (11)$$

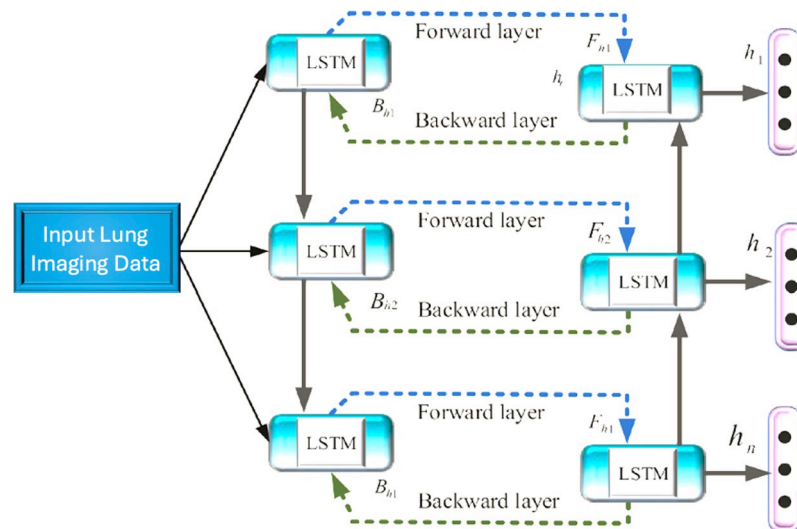


Fig 2. Bi-LSTM architecture.

<https://doi.org/10.1371/journal.pone.0316136.g002>

3.6. Bayesian Optimization (BO)

To optimize the hyperparameters of the Random Forest (RF) and Artificial Neural Network (ANN) models for accurate solar still hourly production forecasting, Bayesian Optimization Algorithm (BOA) was employed. BOA is a powerful technique for finding the global optimum of a black-box function, especially when function evaluations are computationally expensive [10, 64–66].

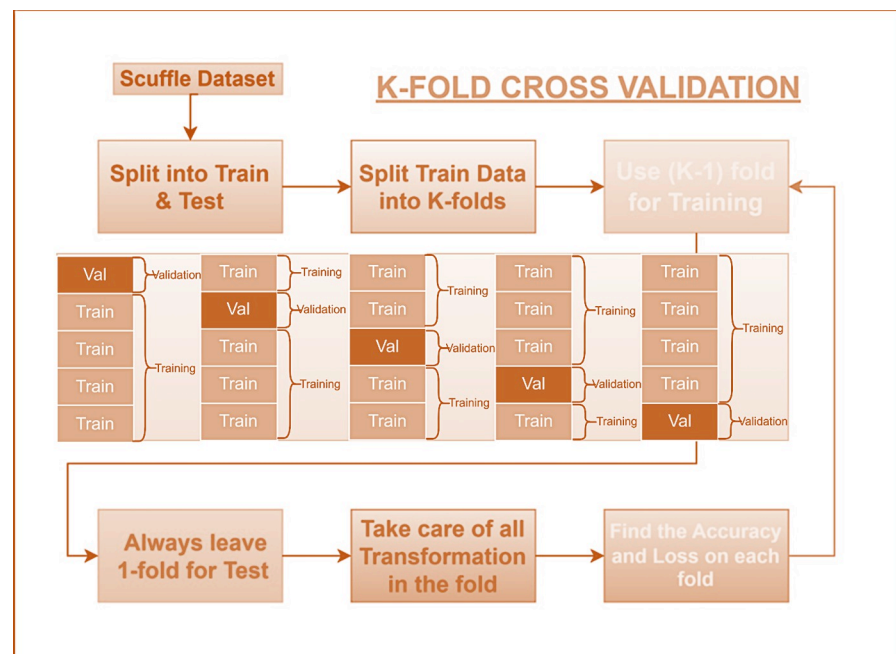


Fig 3. K-fold cross validation method for training/ testing data formulation.

<https://doi.org/10.1371/journal.pone.0316136.g003>

In this study, the 5-fold cross-validation accuracy was used as the objective function for BOA to prevent overfitting. The input variables (X) of BOA were the hyperparameters of the RF and ANN models, including the number of trees, maximum tree depth, learning rate, and number of neurons. The output variable (y) was the accuracy of the 5-fold cross-validation. The Upper Confidence Bound (UCB) acquisition function was selected for BOA, as it effectively balances exploration and exploitation. The BOA algorithm proceeds as follows:

1. In equation (V), the Gaussian Procedure Model GM is built based on the collected data D .

$$GM \sim N(0, C) \quad (12)$$

Where C is the Kernel Matrix

$$C = \begin{bmatrix} C(y_1, y_2) & \cdots & C(y_1, y_t) \\ \vdots & \cdots & \vdots \\ C(y_t, y_1) & \cdots & C(y_t, y_t) \end{bmatrix} \quad (13)$$

$$C(y_j, y_k) = \exp\left(-\frac{1}{2}\|y_j - y_k\|^2\right) \quad (14)$$

2. In the second step of the process, the next position (y_{t+1}) to sample is chosen by the acquisition function. This is the place at which the measured feature (GM) is anticipated to be the best (an accuracy that is achieved by fivefold confirmation.). After that, the Gaussian process model (GM) undergoes modification by incorporating the most recent assessment (y_{t+1}) into its framework.
3. The two processes are continued until the state that is considered to be terminal is attained.

Bayesian optimization is a technique used to find the global maximum of an unknown function. It is particularly useful when function evaluations are expensive, such as in our case where each evaluation involves conducting experiments on a physical system. Bayesian optimization models the objective function as a Gaussian Process (GP). By utilizing the posterior mean and variance of the GP, it can efficiently select the next evaluation point. The GP-Upper Confidence Bound (GP-UCB) acquisition function is a commonly used method for selecting the next sample location. This method balances exploration (trying new regions) and exploitation (focusing on promising areas) to efficiently optimize the function.

$$b_n = \underset{b \in B}{\text{ArgMax}} \mathcal{U}_{n-1}(b) + \mathcal{B}_n^{\frac{1}{2}} \delta_{n-1}(b), \quad (15)$$

Where \mathcal{B}_n is a scalar that varies with each iteration and indicates the level of confidence for the Gaussian Procedure. The Gaussian Process (GP) method selects the next evaluation point based on the uncertainty of the predicted function value. Specifically, it chooses points where the confidence interval of the GP's prediction is the widest. By iteratively evaluating the function at these points, the GP refines its prediction and reduces uncertainty, leading to a more accurate estimation of the global maximum. While the GP-UCB method can be computationally expensive, it's important to note that computational resources are relatively inexpensive. However, Bayesian optimization is most effective for problems with a limited number of parameters. To address this limitation, techniques like dimensionality reduction or incorporating additional structures into the kernel function can be employed to scale the method to larger problems.

3.7. Optimized Bi-LSTM model algorithmic steps

The following algorithmic steps are performed using Bi-LSTM model:

```

1 def objective_function(hyperparameters):
    Create a Bi-LSTM model with the given hyperparameters
    model = create_lstm_model(hyperparameters)
    ii) Compile the model with regularization
    model.compile(optimizer = hyperparameters['optimizer'],
                  loss = hyperparameters['loss'],
                  metrics = ['accuracy'])
    iii) Train the model with data augmentation
    for epoch in range(hyperparameters['epochs']):
        for batch in data_generator:
            X_batch, y_batch = batch
            X_augmented = augment_data(X_batch)
            model.train_on_batch(X_augmented, y_batch)
    iv) Evaluate the model on the validation set
    validation_acc = model.evaluate(X_val, y_val)[1]
    return -validation_acc # Maximize validation accuracy
2 Define the hyperparameter space of Bi-LSTM
hyperparameters = {
    'learning_rate': (1e-4, 1e-2),
    'batch_size': (16, 128),
    'units': (32, 128),
    'dropout': (0.1, 0.5),
    'optimizer': ['adam', 'rmsprop', 'sgd'],
    'loss': ['binary_crossentropy', 'categorical_crossentropy'],
    'epochs': (10, 100)
3 Create a Gaussian Process Regressor
kernel = Matern(length_scale = 1.0, nu = 1.5)
gpr = GaussianProcessRegressor(kernel = kernel)
4 Initialize the Bayesian optimization process
X_init = np.random.uniform(low = 0.0, high = 1.0, size = (10, len
(hyperparameters)))
y_init = np.array([objective_function(hyperparameters) for hyper-
parameters in X_init])
gpr.fit(X_init, y_init)
5 Perform Bayesian optimization
num_iterations = 50
for i in range(num_iterations):
    i) Sample a new hyperparameter configuration
    x_next = gpr.sample_optimization_target(n_samples = 1)[0]
    ii) Evaluate the objective function at the new configuration
    y_next = objective_function(x_next)
    iii) Update the Gaussian Process regressor
    gpr.fit(np.vstack([X_init, x_next]), np.hstack([y_init,
y_next]))
    iv) Update the best hyperparameters
    best_hyperparameters = x_next if y_next < best_y else
best_hyperparameters
return best_hyperparameters

```

3.8. Performance evaluations measures

To assess the performance of our proposed system, we employed standard machine learning classification metrics: precision (positive predictive value), negative predictive value,

specificity, sensitivity, and overall accuracy [67]. These metrics evaluate the system's ability to correctly identify lung cancer cases.

3.9. Confusion matrix

This research employs a confusion matrix, a commonly used tabular format for assessing the performance of classification models on a given test dataset. The test data consists of known true positive and true negative values. Although the confusion matrix concept is relatively simple, the associated terminology can be misleading.

A confusion matrix is a visualization tool used in machine learning, particularly in classification problems, to evaluate the performance of a model. It provides a tabular representation of the predicted and actual classes. Here are the key parameters and their definitions:

True Positive (TP): Correctly predicted positive instances.

True Negative (TN): Correctly predicted negative instances.

False Positive (FP): Incorrectly predicted positive instances (type I error).

False Negative (FN): Incorrectly predicted negative instances (type II error).

The confusion matrix offers a detailed assessment of our model's performance by categorizing its predictions. Table 1 presents the confusion matrix, which calculates various performance metrics based on actual and predicted values.

3.9.1. Sensitivity. True positive rate (TPR) or sensitivity quantifies a classifier's ability to correctly identify positive cases. Essentially, it measures the probability that a diseased individual will be accurately diagnosed as such. The following equation calculates sensitivity:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (16)$$

This indicates the likelihood of a correctly positive test result for a patient diagnosed with the disease.

3.9.2. Specificity. True negative rate (TNR) or specificity complements sensitivity by assessing the model's ability to correctly identify negative instances. It measures the proportion of actual negative cases that are accurately classified as negative. The following equation calculates specificity:

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (17)$$

This indicates the likelihood of a correctly negative test result for a truly healthy patient.

Table 1. Lung cancer detection based by utilizing optimized machine learning models and extracting hand-crafted GLCM features using 10-fold CV.

Method	Sensitivity	Specificity	PPV	NPV	Accuracy	F1-Score	AUC
SVM Linear	1.0000	0.9199	0.8620	1.0000	0.9466	0.9259	0.9946
SVM Quadratic	0.9971	0.9877	0.9803	0.9982	0.9913	0.9858	0.9998
SVM Cubic	0.9972	1.0000	1.0000	0.9982	0.9989	0.9958	0.9994

PPV: Positive predictive value, NPV: negative predictive value, AUC: area under the receiver operating characteristic curve

<https://doi.org/10.1371/journal.pone.0316136.t001>

3.9.3. Positive Predictive Value (PPV). Precision, PPV, measures a classifier's ability to correctly predict positive instances. Mathematically, PPV can be calculated as:

$$PPV = \frac{TP}{TP + FP} \quad (18)$$

3.9.4. Negative Predictive Value (NPV). The NPV also known as negative precision, assesses a classifier's ability to correctly predict negative instances. Mathematically, NPV can be calculated using the following equation:

$$NPV = \frac{TN}{TN + FN} \quad (19)$$

3.9.5. Accuracy. Accuracy is a common metric for evaluating a classification model's overall performance. It represents the proportion of correct predictions among all predictions. Mathematically:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (20)$$

3.10. Training and testing data formulation

To evaluate and optimize classifier performance, this study employed 10-fold Jackknife cross-validation, a robust technique suitable for limited datasets. This method involves partitioning the data into ten equal folds. Each fold is sequentially used for testing, while the remaining nine folds serve as training data. This process is repeated ten times, ensuring comprehensive evaluation of model performance.

Cross-validation is a technique used to assess the performance of a machine learning model. In this approach, the dataset is divided into multiple subsets. A portion of the data is used to train the model, while the remaining portion is used to evaluate its performance. A common example is 10-fold cross-validation. Here, the dataset is split into ten equal-sized folds. Nine folds are used for training, and the 10th fold is used for validation. This process is repeated ten times, with each fold serving as the validation set once. The final performance metric is the average of the ten iterations. 10-fold cross-validation offers a more reliable estimate of model performance compared to a single train-test split. It allows the model to be evaluated on a larger portion of the dataset while remaining computationally efficient.

Fig 3 illustrates the 10-fold cross-validation process, a common technique in machine learning to assess model performance. This process involves the following steps:

1. Data Division: The dataset is divided into ten equal-sized folds.
2. Model Training and Evaluation: Nine folds are used to train the model, while the remaining fold is used for validation.
3. Iteration: This process is repeated ten times, with each fold serving as the validation set once.
4. Performance Averaging: The performance metrics from each iteration are averaged to obtain a final estimate.

By training and evaluating the model multiple times, 10-fold cross-validation provides a more reliable assessment of its performance compared to a single train-test split. This is particularly beneficial when dealing with smaller datasets, as it maximizes the utilization of available data for both training and validation.

Given Data:

- NSCLC: 377 images
- SCLC: 563 images

Total Dataset: $377 + 563 = 940$ images

This technique involves dividing the dataset into 10 equally sized folds. In each iteration:

1. Training: 9 folds (approximately 846 images) are used to train the model.
2. Testing: 1 fold (approximately 94 images) is used to evaluate the model's performance.

This process is repeated 10 times, with each fold serving as the testing set once. The final performance metric is the average of the 10 iterations.

4. Results and discussions

This study aimed to enhance lung cancer detection through a dual approach. The first approach involved extracting hand-crafted features, including GLCM to capture distinct image characteristics. These features were subsequently fed into robust machine learning algorithms. The second approach explored the potential of deep learning by employing LSTM algorithms.

The Table 1 and Fig 4 presents the lung cancer detection performance using ML SVM kernels with 10-fold cross validation. The SVM Cubic model demonstrated the highest accuracy and F1-score, making it the most effective classifier for lung cancer detection in this study. It correctly identified 99.72% of true lung cancer cases and 100% of true non-cancer cases. Additionally, it had a perfect positive predictive value, meaning all positive predictions were accurate. While the SVM Quadratic model also performed exceptionally well, with high sensitivity, specificity, and F1-score, the SVM Cubic model's slightly superior performance in terms of accuracy and F1-score suggests it is the most reliable choice for this task.

The Table 2 and Fig 5 shows the performance of various machine learning algorithms in detecting lung cancer using hand-crafted features extracted from GLCM with 5-fold cross-validation. The SVM Cubic consistently outperforms other models, achieving the highest scores in most metrics. It excels in sensitivity (99.72%), specificity (99.82%), PPV (99.72%), NPV (99.82%), accuracy (99.78%), F1-Score (99.44%), and AUC (99.95%). SVM Linear has the highest sensitivity (100%), but its lower specificity and PPV limit its overall performance. Decision Tree and SVM Quadratic both demonstrate excellent specificity and AUC, but their sensitivity and F1-Score are slightly lower than SVM Cubic. Naïve Bayes performs the worst among the models, with lower scores across most metrics. For lung cancer detection based on hand-crafted GLCM features, SVM Cubic is the most recommended model due to its superior performance in all key metrics. It provides a reliable and accurate solution for this task. However, the choice of the best model may depend on specific application requirements and the relative importance of different metrics.

The Table 3 presents the computational performance of various machine learning and deep learning algorithms. The comparative analysis of various machine learning models for lung cancer prediction reveals that SVM Cubic consistently outperforms other models in terms of accuracy, achieving a rate of 99.80%. While LSTM also demonstrates high accuracy (99.87%),

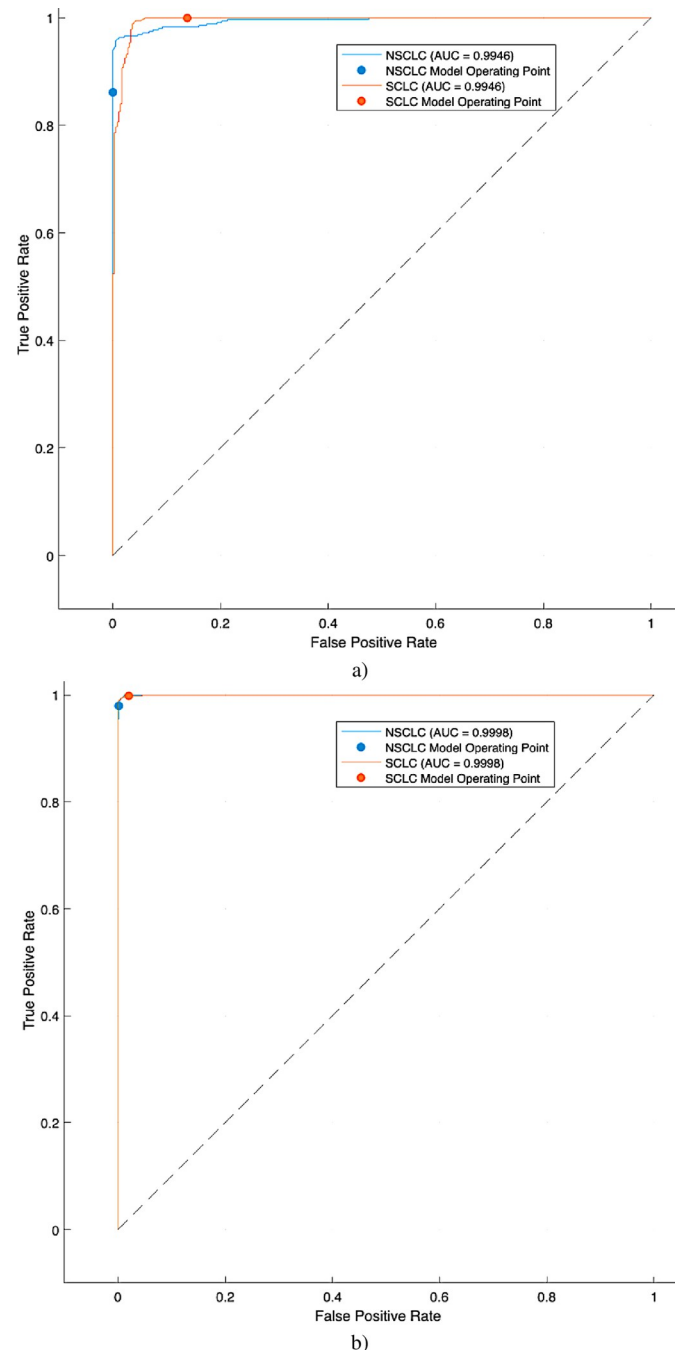


Fig 4. AUC to distinguish SCLC from NSCLC utilizing different machine learning and deep learning algorithms a) SVM Linear, b) SVM Quadratic with 10 fold cross validation (CV).

<https://doi.org/10.1371/journal.pone.0316136.g004>

its significantly longer training time and higher computational complexity may limit its practicality in certain applications. Naïve Bayes, on the other hand, offers a good balance between accuracy (95.20%) and computational efficiency, making it suitable for large datasets. SVM Linear and SVM Quadratic exhibit moderate accuracy and computational complexity, while Tree strikes a balance between the two with reasonable accuracy (98.60%) and low computational cost. Ultimately, the choice of the best model depends on the specific requirements of

Table 2. Lung cancer detection based by utilizing optimized machine learning models and extracting hand-crafted GLCM features using 5-fold CV.

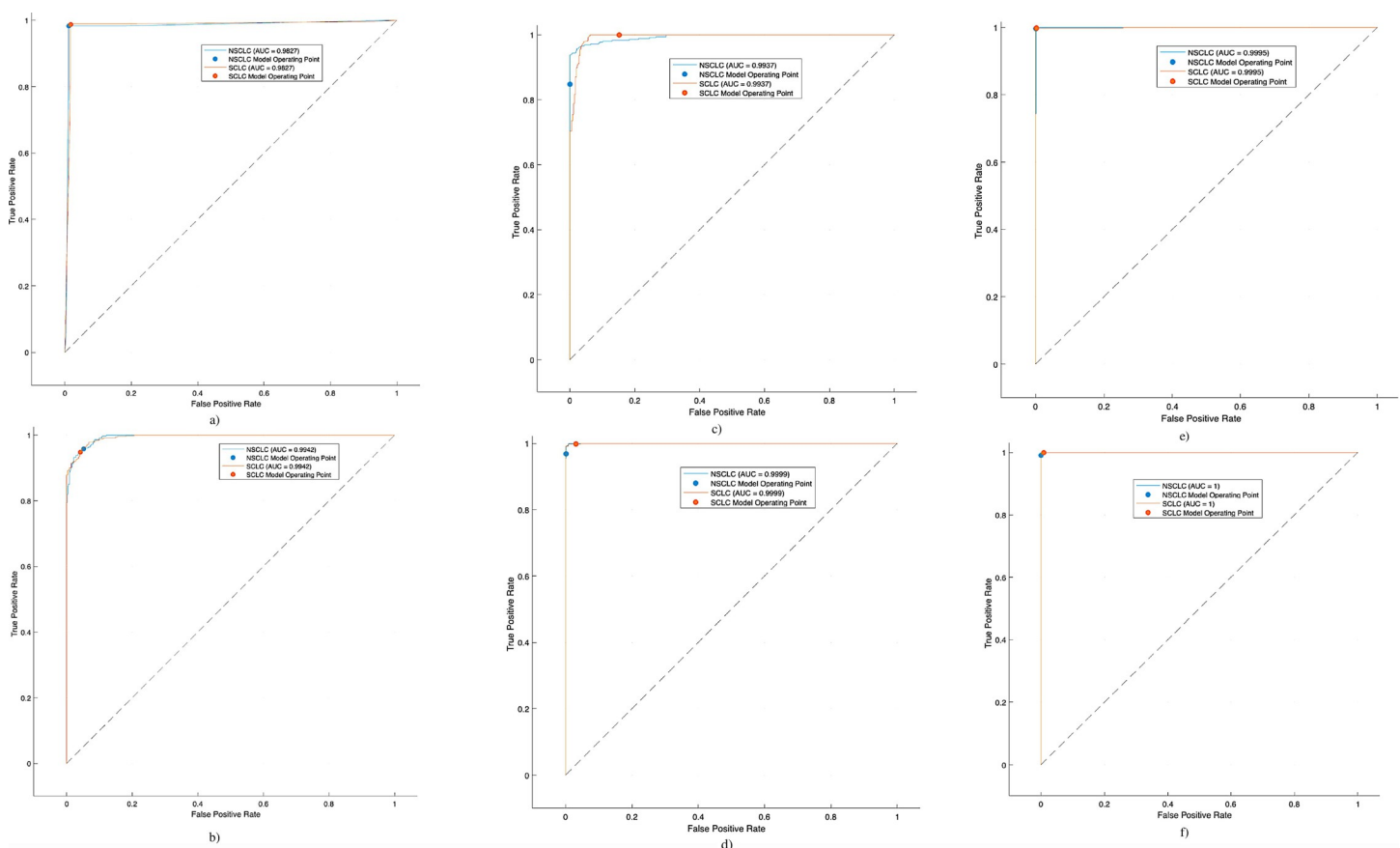
Method	Sensitivity	Specificity	PPV	NPV	Accuracy	F1-Score	AUC
Naïve Bayes	0.9214	0.9727	0.9577	0.9485	0.9521	0.8654	0.9942
Decision Tree	0.9803	0.9893	0.9831	0.9876	0.9858	0.9624	0.9827
SVM Linear	1.0000	0.9125	0.8479	1.0000	0.9412	0.9177	0.9937
SVM Quadratic	0.9971	0.9808	0.9690	0.9982	0.9869	0.9800	0.9995
SVM Cubic	0.9972	0.9982	0.9972	0.9982	0.9978	0.9944	0.9995

PPV: Positive predictive value, NPV: negative predictive value, AUC: area under the receiver operating characteristic curve

<https://doi.org/10.1371/journal.pone.0316136.t002>

the application, such as the size of the dataset, the importance of accuracy, and the available computational resources.

The Table 3 provides insights into the computational efficiency of various machine learning models for lung cancer prediction. Tree-based models, like Decision Trees and Naïve Bayes, exhibit fast training times, small model sizes, and high prediction speeds. However, their accuracy can be lower compared to more complex models. SVM models, both linear and non-linear, offer moderate training times, relatively small model sizes, and reasonable prediction speeds. They generally achieve higher accuracy than simpler models but can be computationally more demanding, especially for non-linear kernels like quadratic and cubic. LSTM, a deep

**Fig 5. AUC to distinguish SCLC from NSCLC utilizing different machine learning and deep learning algorithms a) Decision Tree, b) Naïve Bayes, c) SVM Linear, d) SVM Quadratic, e) SVM Cubic, f) LSTM with 5 folds CV.**

<https://doi.org/10.1371/journal.pone.0316136.g005>

Table 3. Comparative analysis of various performance metrics for predicting lung cancer.

Model Number (Features Selected/total)	Training time	Model size	Prediction speed	Accuracy % (Validation)	Total Cost (Validation)
Tree (22/22)	13.71 sec	8kB	2300 obs/sec	98.60	13
Naïve Bayes (22/22)	7.26 sec	335kB	51000 obs/sec	95.20	44
SVM Linear (22/22)	9.80 sec	46kB	1500 obs/sec	94.10	54
SVM Quadratic (22/22)	9.00 sec	17 kB	1500 obs/sec	98.70	12
SVM Cubic (22/22)	14.56 sec	15 kB	2000 obs/sec	99.80	2
LSTM (22/22)	1138.3 sec	722 kB	3500 obs/sec	99.87	129

<https://doi.org/10.1371/journal.pone.0316136.t003>

learning model, demands significantly more training time, a larger model size, and lower prediction speed compared to traditional machine learning models. However, it often achieves the highest accuracy, making it suitable for complex datasets and scenarios where accuracy is prioritized over computational efficiency. The choice of model depends on the specific requirements of the application, balancing accuracy, computational resources, and real-time constraints.

The Fig 5 presents the AUC to discriminate SCLC from NSCLC utilizing different machine learning and deep learning techniques. The highest AUC of 1.00 was obtained using LSTM followed by SVM Cubic with AUC of 0.9999.

Fig 6 illustrates parallel coordinate plots, a method for visualizing high-dimensional data. Each observation is depicted as a line connecting its values across various variables, revealing patterns and relationships within the data. The plots employ GLCM features extracted from MRI images of lung cancer types. The dataset comprises 356 NSCLC MRI subjects and 563 SCLC MRI subjects. Blue lines represent NSCLC cases, and red lines represent SCLC cases. Solid lines indicate correctly predicted cases, while lines with asterisks denote incorrectly detected cases.

In Fig 6A, SVM Linear achieves a validation accuracy of 98.69%, but with a higher incidence of incorrectly predicted data points, visualized as asterisks. Fig 6B demonstrates the predictions of SVM Cubic, which exhibits fewer incorrect classifications against each computed GLCM feature, also represented by asterisks.

The optimized Deep Learning LSTM model in Table 4 demonstrated exceptional performance in lung cancer detection. It achieved a perfect sensitivity of 100%, indicating that it correctly identified all actual lung cancer cases. Additionally, the model demonstrated a specificity of 99.82%, suggesting that it accurately identified 99.82% of true negative cases. The positive predictive value (PPV) of 99.72% indicates that 99.72% of positive predictions made by the model were indeed correct. Furthermore, the model achieved a perfect negative predictive value (NPV) of 100%, suggesting that all negative predictions made by the model were accurate. The overall accuracy of the model was 99.89%, which is exceptionally high and demonstrates its effectiveness in lung cancer detection. The F1-score of 99.86 reflects a good balance between precision and recall. Finally, the AUC of 1.0000 indicates that the model achieved an area under the ROC curve of 1, which is the maximum possible value and signifies perfect discrimination between lung cancer and non-lung cancer cases.

The Fig 7 presents the accuracy-loss graph to distinguish the NSCLC from SCLC using optimized Bi-LSTM model with Bayesian optimization and using different activation functions. The adam optimizer yielded the highest validation accuracy of 99.90% followed by sgdm with validation accuracy 99.20% and rmsprop with validation accuracy of 97.46%.

The Table 5 demonstrates the significant performance improvement of the proposed methods, particularly the Bi-LSTM model, compared to previous studies. While prior research has relied on various feature engineering techniques, such as shape, intensity, texture, gradient,

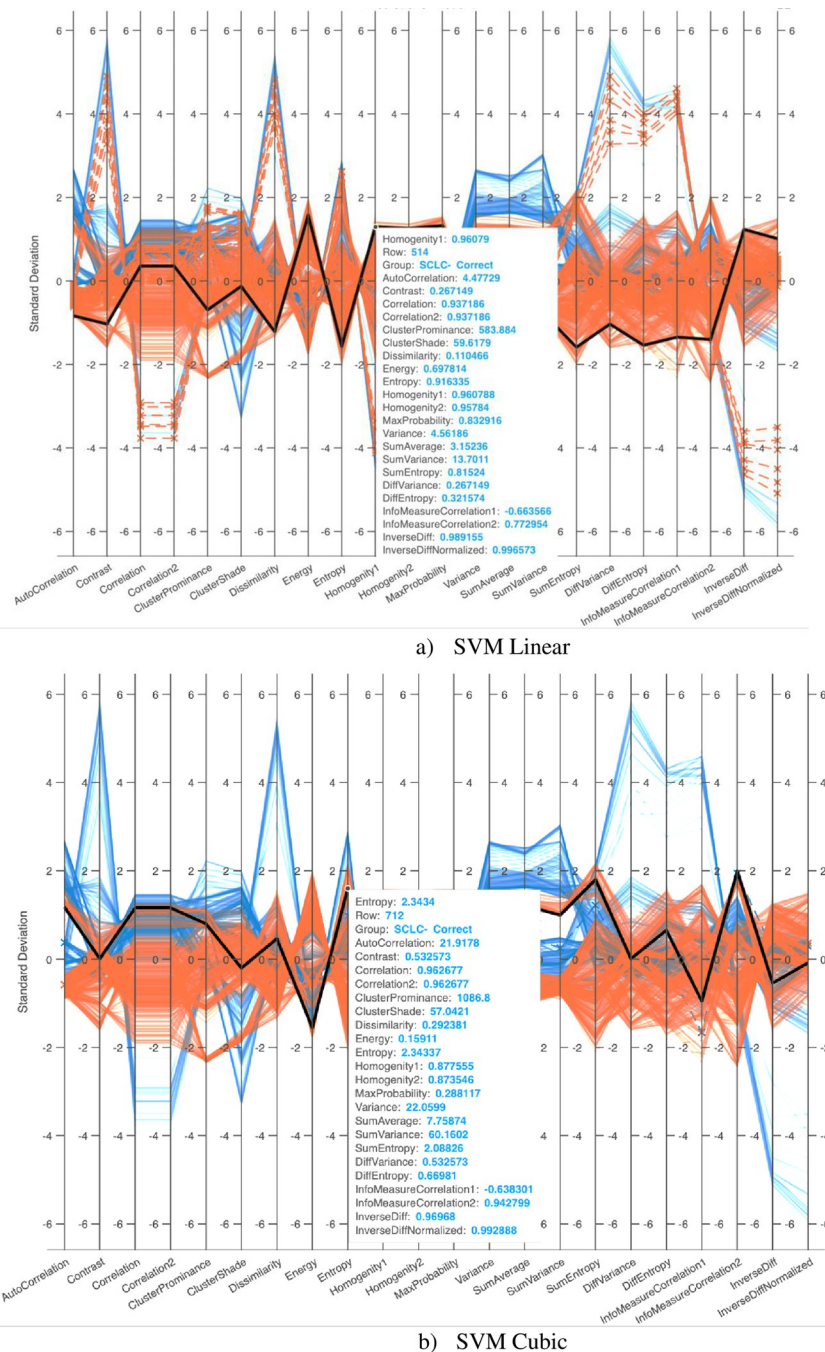


Fig 6. Parallel Coordinate graph to distinguish NSCLC from SCLC by computing GLCM features from lung cancer using a) SVM Linear, b) SVM Cubic.

<https://doi.org/10.1371/journal.pone.0316136.g006>

Table 4. Lung cancer detection using optimized deep learning LSTM with 10-fold cross-validation.

Method	Sensitivity	Specificity	PPV	NPV	Accuracy	F1-score	AUC
Bi-LSTM	1.0000	0.9982	0.9972	1.0000	0.9989	0.9986	1.0000

<https://doi.org/10.1371/journal.pone.0316136.t004>

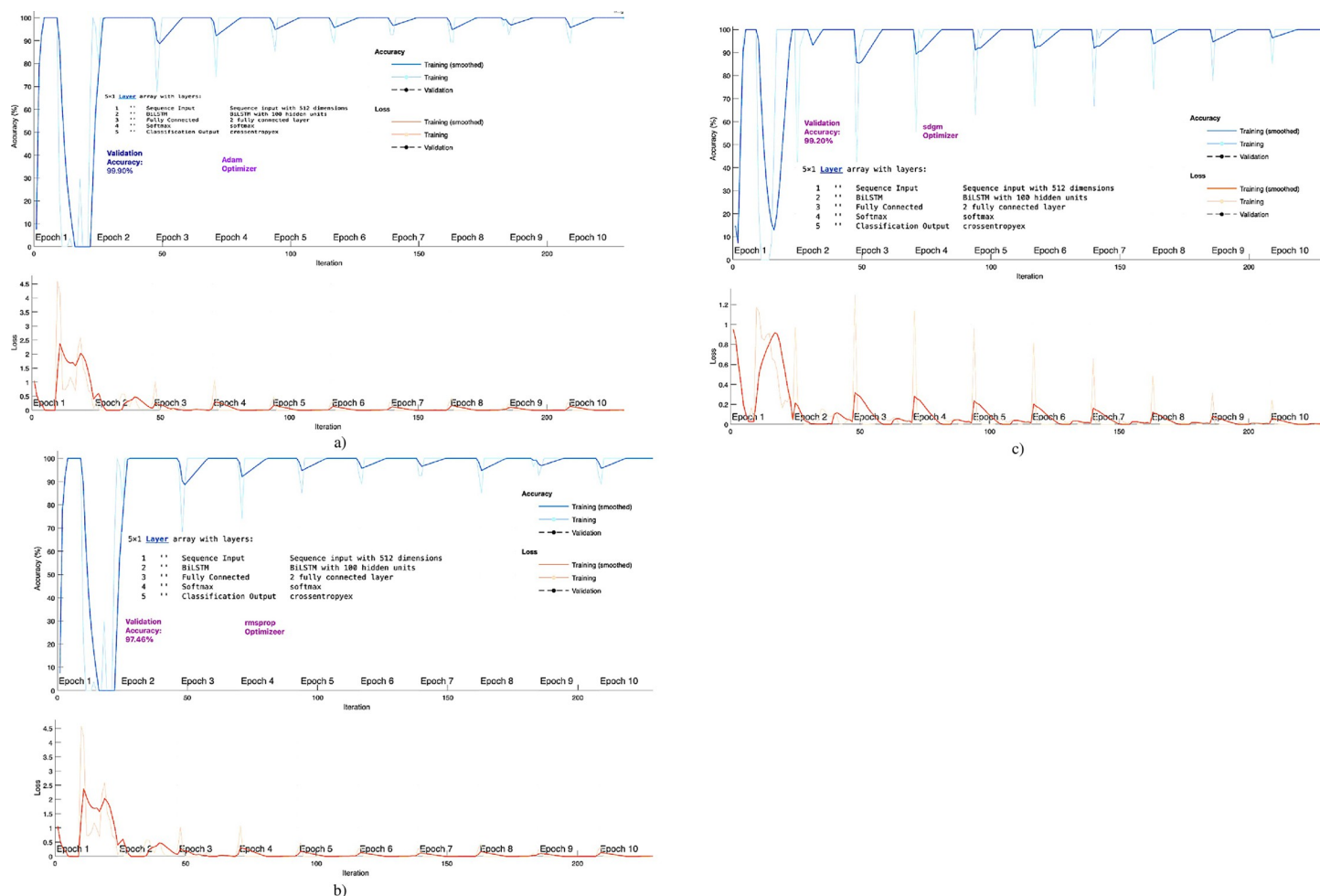


Fig 7. Accuracy-loss graph to detect Lung cancer using Bi-LSTM with different optimizers a) adam, rmsprop, c) sgdm.

<https://doi.org/10.1371/journal.pone.0316136.g007>

and morphology, the proposed approach leverages a combination of GLCM features and a deep learning architecture to capture intricate patterns within the data. This innovative approach has resulted in significantly higher sensitivity and accuracy in detecting lung cancer, positioning it as a promising technique for advancing the field of medical image analysis.

The Table 5 presents the comparative results with existing studies. In this study Teramoto et al. [68] computed the shape and intensity based features and achieved a sensitivity of 83%. Orozco et al. [69] computed texture features and achieved 84%. Guo et al. [70] calculated texture and shape features and obtained 94% of sensitivity. Retico et al. [72] utilized the texture and morphological features and yielded a sensitivity of 72%. This study utilized and optimized machine learning SVM cubic with GLCM features and obtained Sensitivity of 99.72% and accuracy of 99.78% and LSTM with Sensitivity of 100% and accuracy of 99.89%.

LSTMs excel at handling sequential data and capturing long-term dependencies. This makes them particularly suitable for medical image analysis tasks like lung cancer detection, where temporal patterns in image sequences are crucial. Unlike CNNs, which primarily focus on spatial patterns, LSTMs can effectively analyze the evolution of lung nodules or other abnormalities over time. Additionally, LSTMs can automatically learn relevant features from

Table 5. Comparison of results with previous studies.

Author	Features Used	Performance
Teramoto et al. [68]	1. Shape	Sensitivity = 83.00%,
	2. Intensity	
Orozco et al. [69]	1. Texture	Sensitivity = 84.00%,
Guo et al. [70]	1. Texture	Sensitivity = 94.00%,
	2. Shape	
Messay et al. [71]	1. Shape	Sensitivity = 82.00%,
	2. Gradient	
	3. Intensity	
Retico et al. [72]	1. Texture	Sensitivity = 72.00%,
	2. Morphology	
This study	i) GLCM SVM Cubic	i) Sensitivity = 99.72%, Accuracy = 99.78%
	ii) Bi-LSTM	ii) Sensitivity = 100%, Accuracy = 99.89%

<https://doi.org/10.1371/journal.pone.0316136.t005>

the input data, reducing the need for manual feature engineering. This makes them a powerful tool for improving the accuracy and reliability of lung cancer detection.

To enhance medical image analysis, high-quality images are essential. We initially improved image quality using robust enhancement techniques. Subsequently, we extracted relevant static features using machine learning algorithms. To optimize the performance of machine learning and deep learning LSTM models, we employed Bayesian optimization, a more powerful technique than random or grid search. Unlike previous studies that utilized fewer features and default parameters, we extracted a comprehensive set of robust GLCM-based features to capture more relevant information from lung cancer images. This enabled us to effectively distinguish between different image types. Our experimental results demonstrate that our proposed approach achieved superior detection performance compared to existing methods.

The SVM Cubic model demonstrated superior performance in terms of accuracy and F1-score, making it the most effective classifier for lung cancer detection. However, it's important to consider the trade-off between accuracy and computational cost. While LSTM achieved the highest accuracy, its significantly higher training time and model size may limit its practical applicability. Future research should explore techniques to improve the computational efficiency of deep learning models while maintaining high accuracy. Additionally, incorporating more diverse datasets and advanced feature engineering techniques can further enhance the generalizability and robustness of these models. Ultimately, the goal is to develop accurate, efficient, and interpretable AI-driven tools to assist healthcare professionals in early diagnosis and treatment planning for lung cancer.

5. Conclusions

Lung cancer remains a major global health crisis, characterized by exceptionally low survival rates. As the most prevalent and deadliest cancer worldwide, its incidence has surged dramatically. This study aimed to differentiate between NSCLC and SCLC. To achieve this, we employed a combination of hand-crafted GLCM features and deep learning techniques, specifically the Bi-LSTM model. The exceptional performance of our Bi-LSTM model, as highlighted in the findings, underscores the potential of these methods for early lung cancer detection. Early detection is crucial for improving patient outcomes and reducing mortality rates.

5.1. Limitations and future directions

To further enhance the methodological rigor of our study, we plan to explore hybrid deep learning architectures and optimize model parameters. We will also expand our evaluation metrics and visualization techniques to provide a more comprehensive assessment of model performance. To address the limitations of our current dataset, we will test these methodologies on larger and more diverse lung cancer datasets. Additionally, we will incorporate clinical information alongside imaging features to improve diagnostic accuracy and predict disease recurrence, survival, and severity more effectively.

The current study employed a limited dataset and a restricted set of machine learning algorithms. Future work will involve exploring a wider range of algorithms and evaluating the clinical significance of the hybrid feature extraction approach. We will also compare our approach to existing clinical tools for lung cancer detection. To address these limitations, we will evaluate the hybrid feature extraction approach on larger and more diverse datasets, incorporating clinical information such as patient demographics, medical history, and symptoms. Furthermore, we will compare our approach to state-of-the-art lung cancer detection methods and explore its applicability to other medical imaging tasks, such as the detection of other cancers, neurodegenerative diseases, and cardiovascular diseases. By expanding the scope of our research, we aim to develop more robust and clinically relevant lung cancer detection models.

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.

Author Contributions

Conceptualization: Su Diao, Yajie Wan, Danyi Huang, Shijia Huang, Touseef Sadiq, Mohammad Shahbaz Khan, Lal Hussain, Badr S. Alkahtani, Tehseen Mazhar.

Data curation: Su Diao, Yajie Wan, Lal Hussain, Badr S. Alkahtani.

Formal analysis: Su Diao, Danyi Huang, Touseef Sadiq, Lal Hussain, Badr S. Alkahtani, Tehseen Mazhar.

Funding acquisition: Yajie Wan, Shijia Huang.

Investigation: Mohammad Shahbaz Khan, Lal Hussain.

Methodology: Mohammad Shahbaz Khan, Lal Hussain, Badr S. Alkahtani, Tehseen Mazhar.

Project administration: Su Diao, Lal Hussain.

Resources: Lal Hussain.

Software: Lal Hussain.

Supervision: Lal Hussain.

Validation: Yajie Wan, Touseef Sadiq, Lal Hussain.

Visualization: Yajie Wan, Danyi Huang, Lal Hussain, Tehseen Mazhar.

Writing – original draft: Shijia Huang, Lal Hussain.

Writing – review & editing: Su Diao, Touseef Sadiq, Mohammad Shahbaz Khan, Lal Hussain, Badr S. Alkahtani.

References

1. Siegel RL, Miller KD, Jemal A (2018) Cancer statistics, 2018. *CA Cancer J Clin* 68:7–30. <https://doi.org/10.3322/caac.21442> PMID: 29313949

2. Oser MG, Niederst MJ, Sequist L V, Engelman JA (2015) Transformation from non-small-cell lung cancer to small-cell lung cancer: molecular drivers and cells of origin. *Lancet Oncol* 16:e165–e172. [https://doi.org/10.1016/S1470-2045\(14\)71180-5](https://doi.org/10.1016/S1470-2045(14)71180-5) PMID: 25846096
3. Krishnaiah V, Narsimha G, Subhash N, #3 C (2013) Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques. *Int J Comput Sci Inf Technol* 4:39–45.
4. González-Forteza C, Torres CS, Tapia AJ, et al (2011) Confiabilidad y validez de la escala de depresión CES-D en un censo de estudiantes de nivel medio superior y superior, en la Ciudad de México. *Salud Ment* 34:53–59. <https://doi.org/10.3322/caac.20107>. Available.
5. Thummajitsakul S, Kaewsri W, Deetae P (2016) Analysis of intraspecific genetic variation, antioxidant and antibacterial activities of Zingiber zerumbet. *Int Food Res J* 23:1552–1557. <https://doi.org/10.3322/caac.21166>.
6. Silvestri GA, Gould MK, Margolis ML, et al (2007) Noninvasive staging of non-small cell lung cancer: ACCP evidenced-based clinical practice guidelines (2nd edition). *Chest* 132:178S–201S. <https://doi.org/10.1378/chest.07-1360> PMID: 17873168
7. Adenocarcinoma L, Travis WD, Brambilla E, et al (2011) American Thoracic Society International Association for the Study of Lung Cancer / American Thoracic Society / European Respiratory Society: International Multidisciplinary Classification of. *Am Thorac Soc* 8:381–385. <https://doi.org/10.1513/pats.201107-042ST>.
8. Loo PS, Thomas SC, Nicolson MC, et al (2010) Subtyping of undifferentiated non-small cell carcinomas in bronchial biopsy specimens. *J Thorac Oncol* 5:442–447. <https://doi.org/10.1097/JTO.0b013e3181d40fac> PMID: 20195168
9. Nicholson AG, Gonzalez D, Shah P, et al (2010) Refining the diagnosis and EGFR status of non-small cell lung carcinoma in biopsy and cytologic material, using a panel of Mucin staining, TTF-1, cytokeratin 5/6, and P63, and EGFR mutation analysis. *J Thorac Oncol* 5:436–441. <https://doi.org/10.1097/JTO.0b013e3181c6ed9b> PMID: 20068475
10. Megyesfalvi Z, Gay CM, Popper H, et al (2023) Clinical insights into small cell lung cancer: Tumor heterogeneity, diagnosis, therapy, and future directions. *CA Cancer J Clin* 73:620–652. <https://doi.org/10.3322/caac.21785> PMID: 37329269
11. Šutić M, Vukić A, Baranašić J, et al (2021) Diagnostic, Predictive, and Prognostic Biomarkers in Non-Small Cell Lung Cancer (NSCLC) Management. *J Pers Med* 11:1102. <https://doi.org/10.3390/jpm11111102> PMID: 34834454
12. Saltos A, Shafique M, Chiappori A (2020) Update on the Biology, Management, and Treatment of Small Cell Lung Cancer (SCLC). *Front Oncol* 10. <https://doi.org/10.3389/fonc.2020.01074> PMID: 32766139
13. Lemjabbar-Alaoui H, Hassan OUUI, Yang Y-WW, Buchanan P (2015) Lung cancer: Biology and treatment options. *Biochim Biophys Acta - Rev Cancer* 1856:189–210. <https://doi.org/10.1016/j.bbcan.2015.08.002> PMID: 26297204
14. Kim JH, Min SJ, Jang HJ, et al (2015) Comparison of RECIST 1.0 and RECIST 1.1 in patients with metastatic cancer: A pooled analysis. *J Cancer* 6:387–393. <https://doi.org/10.7150/jca.11316> PMID: 25767610
15. Berghmans T, Pasleau F, Paesmans M, et al (2012) Surrogate markers predicting overall survival for lung cancer: ELCWP recommendations. *Eur Respir J* 39:9–28. <https://doi.org/10.1183/09031936.00190310> PMID: 21737547
16. Hamdeni T, Fnaiech F, Gasmi S, et al (2018) Overview and definitions on lung cancer diagnosis. *Middle East Conf Biomed Eng MECBME 2018-March*:165–170. <https://doi.org/10.1109/MECBME.2018.8402427>.
17. Lecun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436. <https://doi.org/10.1038/nature14539> PMID: 26017442
18. Pol R, Balague N (2018) From Micro-To Macroscopic Sports Injuries. Applying the Complex Dynamic Systems Approach To From Micro- To Macroscopic Sports Injuries. Applying the Complex Dynamic Systems Approach To Sports Medicine—a.
19. Chen T, Xu R, He Y, Wang X (2017) Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Syst Appl* 72:221–230. <https://doi.org/10.1016/j.eswa.2016.10.065>.
20. Schneider S, Taylor GW, Kremer SC (2018) Deep Learning Object Detection Methods for Ecological Camera Trap Data.
21. Litjens G, Kooi T, Bejnordi BE, et al (2017) A survey on deep learning in medical image analysis. *Med Image Anal* 42:60–88. <https://doi.org/10.1016/j.media.2017.07.005> PMID: 28778026
22. Ferentinos KP (2018) Deep learning models for plant disease detection and diagnosis. *Comput Electron Agric* 145:311–318. <https://doi.org/10.1016/j.compag.2018.01.009>.

23. Yu J, Yang X, Gao F, Tao D (2017) Deep Multimodal Distance Metric Learning Using Click Constraints for Image Ranking. *IEEE Trans Cybern* 47:4014–4024. <https://doi.org/10.1109/TCYB.2016.2591583> PMID: 27529881
24. Jia Y, Shelhamer E, Donahue J, et al (2014) Caffe: Convolutional Architecture for Fast Feature Embedding. <https://doi.org/10.1145/2647868.2654889>.
25. Kamps PG, Bol M, Steller EJA, et al (2021) Colon carcinoma presenting as ovarian metastasis. *Radiol Case Reports* 16:2799–2803. <https://doi.org/10.1016/j.radcr.2021.06.072> PMID: 34367398
26. Zhang Z, Zhao S, Wang K, et al (2021) Identification of biomarkers associated with cervical lymph node metastasis in papillary thyroid carcinoma: Evidence from an integrated bioinformatic analysis. *Clin Hemorheol Microcirc* 78:117–126. <https://doi.org/10.3233/CH-201074> PMID: 33554894
27. Kourou K, Exarchos TP, Exarchos KP, et al (2015) Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J* 13:8–17. <https://doi.org/10.1016/j.csbj.2014.11.005> PMID: 25750696
28. Mridha MF, Prodeep AR, Hoque ASMM, et al (2022) A Comprehensive Survey on the Progress, Process, and Challenges of Lung Cancer Detection and Classification. *J Healthc Eng* 2022:1–43. <https://doi.org/10.1155/2022/5905230> PMID: 36569180
29. Liang P, Wang H, Liang Y, et al (2022) Feature-scML: An Open-source Python Package for the Feature Importance Visualization of Single-Cell Omics with Machine Learning. *Curr Bioinform* 17:578–585. <https://doi.org/10.2174/1574893617666220608123804>.
30. Shahbandegan A, Mago V, Alaref A, et al (2022) Developing a machine learning model to predict patient need for computed tomography imaging in the emergency department. *PLoS One* 17:e0278229. <https://doi.org/10.1371/journal.pone.0278229> PMID: 36520785
31. V A B, Subramoniam M, Mathew L (2021) Noninvasive detection of COPD and Lung Cancer through breath analysis using MOS Sensor array based e-nose. *Expert Rev Mol Diagn* 21:1223–1233. <https://doi.org/10.1080/14737159.2021.1971079> PMID: 34415806
32. Binson VA, Subramoniam M, Mathew L (2021) Discrimination of COPD and lung cancer from controls through breath analysis using a self-developed e-nose. *J Breath Res* 15:046003. <https://doi.org/10.1088/1752-7163/ac1326> PMID: 34243176
33. Freitas C, Sousa C, Machado F, et al (2021) The Role of Liquid Biopsy in Early Diagnosis of Lung Cancer. *Front Oncol* 11:. <https://doi.org/10.3389/fonc.2021.634316> PMID: 33937034
34. Lener MR, Reszka E, Marciniak W, et al (2021) Blood cadmium levels as a marker for early lung cancer detection. *J Trace Elem Med Biol* 64:126682. <https://doi.org/10.1016/j.jtemb.2020.126682> PMID: 33249371
35. Labory J, Njomgue-Fotso E, Bottini S (2024) Benchmarking feature selection and feature extraction methods to improve the performances of machine-learning algorithms for patient classification using metabolomics biomedical data. *Comput Struct Biotechnol J* 23:1274–1287. <https://doi.org/10.1016/j.csbj.2024.03.016> PMID: 38560281
36. Javed R, Abbas T, Khan AH, et al (2024) Deep learning for lungs cancer detection: a review. *Artif Intell Rev* 57:197. <https://doi.org/10.1007/s10462-024-10807-1>.
37. Kim S, Lim JH, Kim C-H, et al (2024) Deep learning–radiomics integrated noninvasive detection of epidermal growth factor receptor mutations in non-small cell lung cancer patients. *Sci Rep* 14:922. <https://doi.org/10.1038/s41598-024-51630-6> PMID: 38195717
38. Pradhan K, Chawla P (2020) Medical Internet of things using machine learning algorithms for lung cancer detection. *J Manag Anal* 7:591–623. <https://doi.org/10.1080/23270012.2020.1811789>.
39. B V.A., Subramoniam M, Mathew L (2021) Detection of COPD and Lung Cancer with electronic nose using ensemble learning methods. *Clin Chim Acta* 523:231–238. <https://doi.org/10.1016/j.cca.2021.10.005> PMID: 34627826
40. Goroshinskaia IA, Tsvetnenko EZ, Frenkel' ML, et al (1987) [Monoamine oxidase activity and brain levels of histamine and polyamines in various extreme exposures]. *Ukr biokhimicheskii zhurnal* 59:84–6.
41. Chaunzwa TL, Hosny A, Xu Y, et al (2021) Deep learning classification of lung cancer histology using CT images. *Sci Rep* 11:5471. <https://doi.org/10.1038/s41598-021-84630-x> PMID: 33727623
42. Chaturvedi P, Jhamb A, Vanani M, Nemade V (2021) Prediction and Classification of Lung Cancer Using Machine Learning Techniques. *IOP Conf Ser Mater Sci Eng* 1099:012059. <https://doi.org/10.1088/1757-899X/1099/1/012059>.
43. Masood A, Sheng B, Li P, et al (2018) Computer-Assisted Decision Support System in Pulmonary Cancer detection and stage classification on CT images. *J Biomed Inform* 79:117–128. <https://doi.org/10.1016/j.jbi.2018.01.005> PMID: 29366586

44. AbdElNabi MLR, Wajeeh Jasim M, El-Bakry HM, et al (2020) Breast and Colon Cancer Classification from Gene Expression Profiles Using Data Mining Techniques. *Symmetry (Basel)* 12:408. <https://doi.org/10.3390/sym12030408>.
45. He Q, Qiu Z, Tong Y, Song K (2020) A New TTZ Feature Extracting Algorithm to Decipher Tobacco Related Mutation Signature Genes for the Personalized Lung Adenocarcinoma Treatment. *IEEE Access* 8:89031–89040. <https://doi.org/10.1109/ACCESS.2020.2993118>.
46. Hussain L, Aziz W, Alshdadi AA, et al (2019) Analyzing the Dynamics of Lung Cancer Imaging Data Using Refined Fuzzy Entropy Methods by Extracting Different Features. *IEEE Access* 7:64704–64721. <https://doi.org/10.1109/ACCESS.2019.2917303>.
47. Shim S-O, Alkinani MH, Hussain L, Aziz W (2022) Feature Ranking Importance from Multimodal Radiomic Texture Features using Machine Learning Paradigm: A Biomarker to Predict the Lung Cancer. *Big Data Res* 29:100331. <https://doi.org/10.1016/j.bdr.2022.100331>.
48. Ahmed S, Raza B, Hussain L, et al (2023) The Deep Learning ResNet101 and Ensemble XGBoost Algorithm with Hyperparameters Optimization Accurately Predict the Lung Cancer. *Appl Artif Intell* 37:. <https://doi.org/10.1080/08839514.2023.2166222>.
49. Hashemzadeh M, Asheghi B, Farajzadeh N (2019) Content-aware image resizing: An improved and shadow-preserving seam carving method. *Signal Processing* 155:233–246. <https://doi.org/10.1016/j.sigpro.2018.09.037>.
50. Shorten C, Khoshgoftaar TM (2019) A survey on Image Data Augmentation for Deep Learning. *J Big Data* 6:60. <https://doi.org/10.1186/s40537-019-0197-0>.
51. Hussain L, Awan IA, Aziz W, et al (2020) Detecting Congestive Heart Failure by Extracting Multimodal Features and Employing Machine Learning Techniques. *Biomed Res Int* 2020:1–19. <https://doi.org/10.1155/2020/4281243> PMID: 32149106
52. Hussain L, Aziz W, Alshdadi AAA, et al (2019) Analyzing the Dynamics of Lung Cancer Imaging Data Using Refined Fuzzy Entropy Methods by Extracting Different Features. *IEEE Access* 7:64704–64721. <https://doi.org/10.1109/ACCESS.2019.2917303>.
53. Hussain L, Huang P, Nguyen T, et al (2021) Machine learning classification of texture features of MRI breast tumor and peri-tumor of combined pre- and early treatment predicts pathologic complete response. *Biomed Eng Online* 20:63. <https://doi.org/10.1186/s12938-021-00899-z> PMID: 34183038
54. Hussain L, Ahmed A, Saeed S, et al (2018) Prostate cancer detection using machine learning techniques by employing combination of features extracting strategies. *Cancer Biomarkers* 21:393–413. <https://doi.org/10.3233/CBM-170643> PMID: 29226857
55. Asim Y, Raza B, Malik AK, et al (2018) A multi-modal, multi-atlas-based approach for Alzheimer detection via machine learning. *Int. J. Imaging Syst. Technol.* 28:113–123.
56. Hussain L (2018) Detecting epileptic seizure with different feature extracting strategies using robust machine learning classification techniques by applying advance parameter optimization approach. *Cogn Neurodyn* 12:271–294. <https://doi.org/10.1007/s11571-018-9477-1> PMID: 29765477
57. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput* 9:1735–1780.
58. Zheng H, Yuan J, Chen L (2017) Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation. *Energies* 10:1168. <https://doi.org/10.3390/en10081168>.
59. Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. *Neural Comput* 9:1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735> PMID: 9377276
60. Bouktif S, Fiaz A, Ouni A, Serhani MA (2018) Optimal deep learning LSTM model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. *Energies* 11:. <https://doi.org/10.3390/en11071636>.
61. Zhao Y, Shen Y, Zhu Y (2018) Forecasting Wavelet Transformed Time Series with Attentive Neural Networks. In: 2018 IEEE International Conference on Data Mining (ICDM). IEEE, pp 1452–1457.
62. Shen K, Guo J, Tan X, et al (2023) A Study on ReLU and Softmax in Transformer.
63. Oh J, Kim S, Lee C, et al (2023) Preventing Vanishing Gradient Problem of Hardware Neuromorphic System by Implementing Imidazole-Based Memristive ReLU Activation Neuron. *Adv Mater* 35:. <https://doi.org/10.1002/adma.202300023> PMID: 36938884
64. Kudela J (2022) A critical problem in benchmarking and analysis of evolutionary computation methods. *Nat Mach Intell* 4:1238–1245. <https://doi.org/10.1038/s42256-022-00579-0>.
65. Omidvar MN, Li X, Yao X (2022) A Review of Population-Based Metaheuristics for Large-Scale Black-Box Global Optimization—Part I. *IEEE Trans Evol Comput* 26:802–822. <https://doi.org/10.1109/TEVC.2021.3130838>.

66. Dahito M-A, Genest L, Maddaloni A, Neto J (2023) A solution method for mixed-variable constrained blackbox optimization problems. *Optim Eng.* <https://doi.org/10.1007/s11081-023-09874-0>.
67. Kashyap KL, Bajpai MK, Khanna P (2015) Breast cancer detection in digital mammograms. In: *IEEE International Conference on Imaging Systems and Techniques*. p 6.
68. Teramoto A, Fujita H, Takahashi K, et al (2014) Hybrid method for the detection of pulmonary nodules using positron emission tomography/computed tomography: a preliminary study. *Int J Comput Assist Radiol Surg* 9:59–69. <https://doi.org/10.1007/s11548-013-0910-y> PMID: 23793722
69. Orozco HM, Villegas OOV, Dominguez H de JO, Sanchez VGC (2013) Lung Nodule Classification in CT Thorax Images Using Support Vector Machines. In: *2013 12th Mexican International Conference on Artificial Intelligence*. IEEE, pp 277–283.
70. Guo Wei, Wei Ying, Zhou Hanxun, Xue DingYe (2009) An adaptive lung nodule detection algorithm. In: *2009 Chinese Control and Decision Conference*. IEEE, pp 2361–2365.
71. Messay T, Hardie RC, Rogers SK (2010) A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. *Med Image Anal* 14:390–406. <https://doi.org/10.1016/j.media.2010.02.004> PMID: 20346728
72. Retico A, Fantacci ME, Gori I, et al (2009) Pleural nodule identification in low-dose and thin-slice lung computed tomography. *Comput Biol Med* 39:1137–1144. <https://doi.org/10.1016/j.compbiomed.2009.10.005> PMID: 19883906