Advancements in machine learning techniques for precise detection and classification of lung cancer

Hamza Abu Owida[1], Areen Arabiat[2], Muhammad Al-Ayyad[1], Muneera Altayeb[2]
[1]Department of Medical Engineering, Faculty of Engineering, Al-Ahliyya Amman University, Amman, Jordan
[2]Department of Communications and Computer Engineering, Faculty of Engineering, Al-Ahliyya Amman University, Amman, Jordan

Article Info

Article history:

Received Apr 17, 2025
Revised Aug 13, 2025
Accepted Sep 11, 2025

Keywords:

Computed tomography
Confusion matrix
Lung cancer
Machine learning
Performance metrices

ABSTRACT

Lung cancer remains one of the most prevalent and lethal malignancies worldwide, necessitating early detection and accurate classification for effective treatment. In this work, we present a unique machine learning (ML) model that uses medical imaging data to detect and classify lung cancer. Utilizing a dataset of 613 images which obtained from Kaggle, our model combines sophisticated feature extraction methods with three essential algorithms: AdaBoost, stochastic gradient descent (SGD), and random forest (RF). Orange3 data mining software was used to classify the model after it was preprocessed and features were extracted using MATLAB. Nonetheless, the model showed good performance in identifying lung cancer lesions in four different categories: squamous cell carcinoma, big cell carcinoma, adenocarcinoma, and normal. With an accuracy of 0.998 and an AUC range of 1.000, AdaBoost notably produced the best results. Overall, ensemble ML techniques demonstrated notable benefits over single classifiers, indicating its potential to aid in the creation of accurate instruments for the diagnosis of lung cancer in its early stages.

Corresponding Author:

Areen Arabiat
Department of Communications and Computer Engineering, Faculty of Engineering
Al-Ahliyya Amman University
Amman, Jordan
Email: a.arabiat@ammanu.edu.jo

1. INTRODUCTION

Lung cancer remains one of the leading causes of cancer-related mortality worldwide, with a high incidence and poor prognosis, necessitating the development of innovative diagnostic and prognostic tools [1], [2]. Early detection and accurate classification of lung cancer are critical for improving patient outcomes, as timely intervention can significantly impact treatment efficacy and survival rates [3], [4]. Lung cancer imaging methods encompass a range of modalities, each contributing unique information to the diagnostic process. While computed tomography (CT) remains the primary imaging modality, positron emission tomography (PET)/CT, magnetic resonance imaging (MRI), and chest X-rays also play essential roles in the comprehensive evaluation and management of lung cancer patients. The selection of imaging modalities depends on clinical indications, tumor characteristics, and patient-specific factors, with the goal of optimizing patient care and treatment outcomes [5].

Lung cancer imaging plays a crucial role in diagnosis, staging, and monitoring, with a variety of modalities offering unique advantages and limitations in visualizing tumors and assessing disease progression [6], [7]. Since CT has a high spatial resolution and can identify small pulmonary nodules, it continues to be the primary imaging modality for evaluating lung cancer. Comprehensive anatomical data from CT scans is useful for identifying lung lesions, figuring out the size and location of tumors, and determining whether lymph nodes are involved. Moreover, contrast-enhanced CT scans make it easier to see blood vessels and discern between benign and malignant tumors [8]. PET imaging, which is frequently used in addition with CT (PET/CT), offers useful data regarding cellular activity and tumor metabolism. PET/CT scans are used to detect metastases in distant organs, evaluate tumor aggressiveness, and identify malignant lesions by identifying regions of enhanced glucose metabolism [9].

MRI is less commonly used for lung cancer imaging but may be employed in specific clinical scenarios, such as evaluating mediastinal invasion or assessing brain metastases. MRI offers superior soft tissue contrast resolution compared to CT and is particularly valuable in cases where CT or PET/CT findings are inconclusive or when there are contraindications to iodinated contrast agents [10]. Chest X-rays remain a valuable initial screening tool for lung cancer, although their sensitivity for detecting small lesions is limited compared to CT imaging. Nevertheless, chest X-rays are readily accessible, cost-effective, and may serve as a first-line imaging modality for patients with suspected lung cancer, guiding subsequent diagnostic workup [11], [12] Conventional diagnostic methods, such as imaging modalities and histopathological analysis, have limitations in terms of accuracy, sensitivity, and speed, prompting the exploration of alternative approaches [9].

With the potential to improve diagnostic efficiency and accuracy, machine learning (ML) approaches have become increasingly attractive in recent years for the detection and classification of lung cancer [13]-[15]. These algorithms can find patterns and features that indicate the existence, subtype, and stage of lung cancer by analyzing vast amounts of medical imaging data, such as chest X-rays, CT scans, and PET images, along with patient demographics and clinical data [16], [17]. The application of ML in lung cancer detection and classification encompasses various approaches, including supervised learning, unsupervised learning, and deep learning (DL). Supervised learning algorithms, such as support vector machines (SVMs) and random forests (RF), utilize labeled training data to build predictive models for distinguishing between different classes of lung cancer and healthy tissue. Unsupervised learning techniques, such as clustering algorithms, enable the identification of hidden patterns and subgroups within lung cancer datasets, facilitating personalized treatment strategies and prognosis prediction [18]-[21]. Even while ML has great promise for diagnosing lung cancer, there are still a number of obstacles and restrictions that must be resolved. In addition to model interpretability, generalizability, and ethical considerations, these also involve data availability, quality, and standards issues. Regulatory approval, thorough validation, and smooth integration with current healthcare procedures are also necessary for the clinical application of ML-based algorithms [22], [23].

Previous studies have demonstrated various ML and image processing techniques for lung cancer detection and classification using CT imaging data. Research by Lin *et al.* [24] suggested a model based on CT images for non-small cell lung cancer (NSCLC) patient clinical staging and histological type. A total of 107 radiomic characteristics were collected from 309 patients, which were split into training and testing sets. The model was built using four classifiers and had two output layers: clinical stage and histology type. With an AUC of 0.700 and 0.881 for histological typing and clinical staging, respectively, the RF model had the best classification performance.

Gupta *et al.* [25] utilized RF classification, where feature extraction evaluated characteristics such as area, perimeter, and eccentricity to derive useful information for lung cancer detection. Sim *et al.* [26] implemented a texture matching process using the local binary pattern (LBP), which proved superior to other available texture patterns. This approach, combined with SVM classification, enhanced the detection accuracy.

Pradhan and Chawla [27] focused on pre-processing techniques to reduce noise and improve CT image quality through various image enhancement methods. They converted grayscale CT images for segmentation and conducted further morphological opening procedures. Pati [28] used an SVM to categorize CT images into normal and abnormal, claiming high accuracy in early-stage cancer detection. They emphasized that image quality and enhancement levels significantly impact the accuracy of the detection process. Vijayalakshmi *et al.* [29] highlighted the significance of classification in digital image analysis, categorizing CT images based on similarities. In traditional systems, histogram equalization (HE) is used for preprocessing CT images, and feature ex0traction is performed using HE. These various methods collectively contribute to the advancement of lung cancer detection and classification, each bringing unique strengths to the diagnostic process.

Using medical imaging data, namely CT imaging, this work focuses on the creation and assessment of ML algorithms intended to reliably detect and classify lung cancer. Because lung cancer is still the world's top cause of cancer-related death, the key goal is to develop reliable models that can accurately identify lung cancer lesions. This will help address the urgent need for early detection to increase patient survival rates. While they function well, traditional diagnostic techniques are frequently difficult and vulnerable to human error. However, the of this work is to improve the diagnosis process by using cutting-edge ML techniques, which will make it quicker, more accurate, and less dependent on human interpretation. Our method involves the rigorously development of algorithms designed to identify the unique patterns and features of lung cancer lesions in CT images. These algorithms are then trained on large datasets to learn and generalize from a variety of imaging scenarios. The study assesses how well these models identify and categorize lung cancer and investigates how they might be used in medical environments.

To improve classification performance, our novel pipeline incorporates multiple algorithms, supported by a comprehensive feature extraction strategy that combines DL with conventional image processing techniques. Enhanced preprocessing methods ensure high-resolution data quality, while the model's robust validation framework, user-friendly interface, and dynamic adaptability to various imaging protocols facilitate seamless integration into clinical workflows. By comparing outcomes across different ML techniques, we aim to identify the most effective strategies for lung cancer detection.

2. METHOD

This work proposed a ML model for the identification of lung cancer using a dataset of 613 CT images that acquired from Kaggle [30]. The preprocessing stage was deployed using MATLAB in order to enhance image quality and guarantee interoperability with ML algorithms, these steps were crucial through leveling pixel intensity values, boosting contrast by HE, shrinking the images to a consistent size, and using Gaussian filtering to lower noise. Due to the limited number of samples of the dataset, image extraction becomes an essential procedure in computer vision, data mining, and image processing. Using the lightweight convolutional neural network Squeeze Net, which employs Fire modules to extract features from input images, the extracted features are then processed by global average pooling, convolutional, and SoftMax layers, making it suitable for devices with limited resources. In this Python-based model, 1000 features were extracted from images, greatly improving the model's performance and accuracy [31]-[33]. However, this model integrates AdaBoost, SGD, and RF, these classification methods that were applied after feeding the features into the suggested model. While AdaBoost improves weak classifiers, SGD maximizes large datasets and fast convergence to improve the diagnosis of lung cancer from CT images. RF increases prediction accuracy through the use of ensemble learning and robustness against overfitting. By providing information on feature significance, RF helps find CT image features that influence categorization outcomes. By addressing the complex nature of medical imaging problems, this hybrid model enhances diagnostic accuracy in clinical settings. Figure 1 shows the classification model using Orange3.

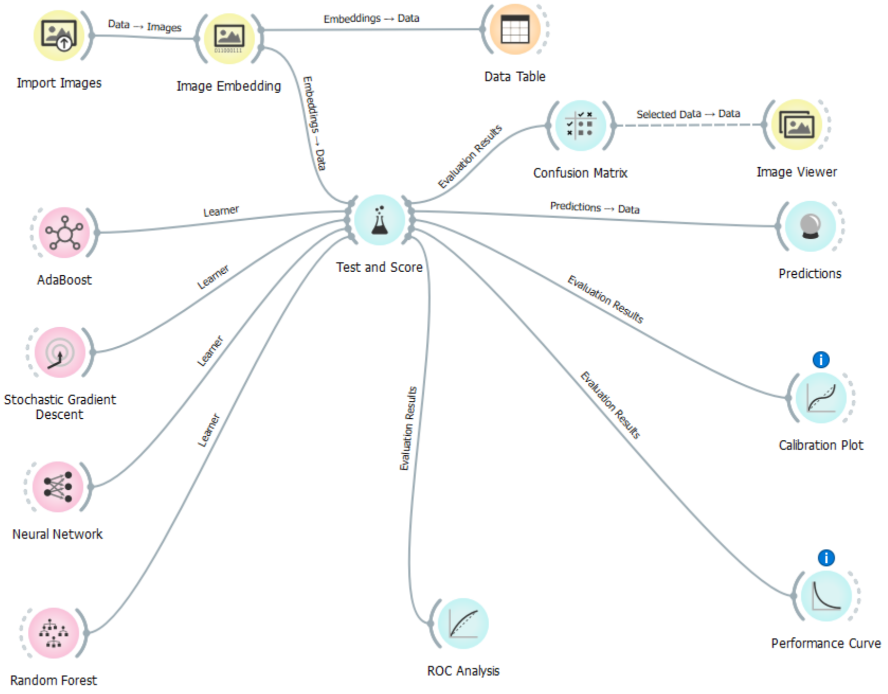
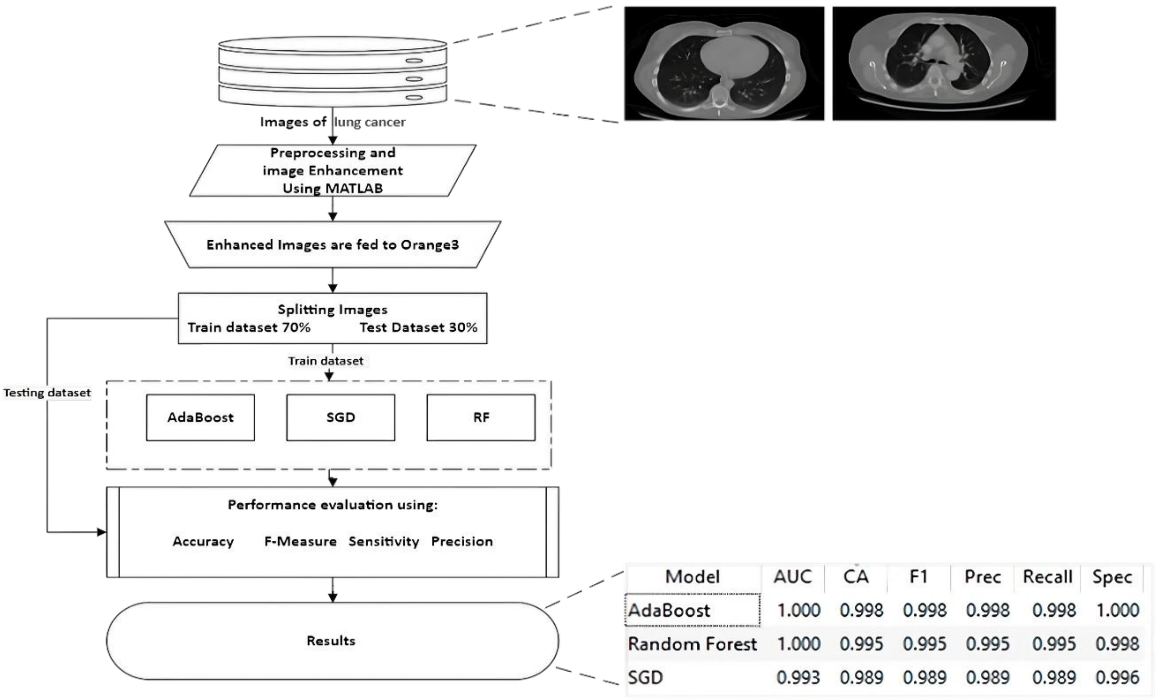


Figure 1. Classification model using Orange3

The model was trained and verified across several subsets of the training data using 10-fold cross-validation, which was used to guarantee model reliability. After training, performance indicators such as accuracy, sensitivity, precision, and F-measure were determined to provide a thorough assessment of the model's efficacy. To help highlight incorrect classifications and areas for improvement, a confusion matrix was also created to show the classification findings. However, the flowchart of lung cancer detection and classification mode is shown in Figure 2. This methodical process made it easier to create a strong model that can reliably identify and categorize lung cancer from CT scan images into four categories: adenocarcinoma, large.cell.carcinoma, normal, and squamous.cell.carcinoma as shown in Figure 3, Figure 3(a) displays adenocarcinoma class, Figure 3(b) displays large cell.carcinoma class, Figure 3(c) displays normal class, and Figure 3(d) displays squamous cell.carcinoma class.



Model	AUC	CA	F1	Prec	Recall	Spec
AdaBoost	1.000	0.998	0.998	0.998	0.998	1.000
Random Forest	1.000	0.995	0.995	0.995	0.995	0.998
SGD	0.993	0.989	0.989	0.989	0.989	0.996

Figure 2. Lung cancer detection and classification model architecture

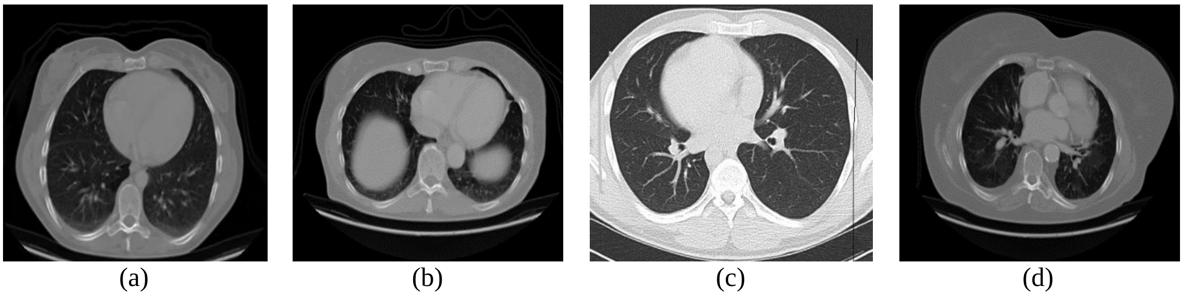


Figure 3. Chest CT-scan images; (a) adenocarcinoma, (b) large.cell.carcinoma, (c) normal, and (d) squamous.cell.carcinoma

3. CLASSIFICATION METHODS

3.1. The random forest

Classifier is a stochastic method that uses a random vector to create numerous decision trees (DTs) to improve accuracy and decrease correlations. Every DT is divided into a subset of features, and the diversity of the tree depends on how many characteristics are considered. The best-split function to encourage tree similarity is found at each split. For a variety of predictions, the objective is to construct an ensemble of several DTs [34]-[36]. Here, RF (1) with equal representation, where y_i is the original value used for feature i and F_i is the value acquired from the system, shows N, the number of characteristics used to find equivalent accounts [37], [38].

$$Random\ forest = \frac{1}{N}\sum_{N=1}^{N}\ (F_i\ -\ y_i\ )2 \quad (1)$$

3.2. AdaBoost

An approach to supervised learning that divides instances into positive and negative groups is called AdaBoost [39]. It works well with imbalanced data since it is based on weighted majority voting standards [40], [41]. Overfitting and generalization are its main problems, though. A unique approach to weak learning is put forward, which increases accuracy by employing numerous thresholds. With this approach, each piece of data is given the same weight, but points that are misclassified are given additional weight. AdaBoost is a simple, adaptable algorithm that can handle unbalanced datasets by changing weights [42], [43]. In (2) demonstrates that when a new tree model is presented, the general tree is eliminated and only the strongest tree is incorporated to the system. With this approach, the model's overall performance constantly becomes better as more simulations are made.

$$Fn(x) = F_{m-1}(x) + argminh_h \sum_{i=1}^{n} L(y_{i,} F_{m-1}(x_i) + h(x_i) \tag{2}$$

when the inserted tree is denoted by h(x$_i$), the freshly inserted tree is represented by y$_i$, the i-th tree prediction result is represented by y$_i$, and the overall model is represented by Fn(x) [44]. The AdaBoost classifier procedure is shown in Figure 4.

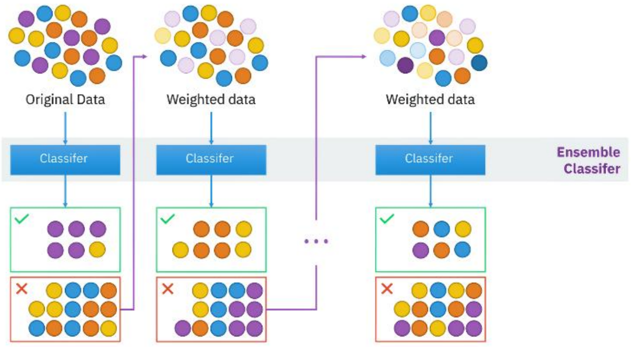


Figure 4. AdaBoost classifier procedure [45]

3.3. Stochastic gradient descent

Strong training techniques for linear, which estimates the cost function's gradient and updates model parameters in response to the addition of fresh training data. Particularly when dealing with enormous datasets, this approach produces substantially better results than traditional techniques [46]. To enhance optimization methods, SGD is a simple and powerful ML model that has seen a lot of effort recently. In (3) illustrates how the optimization problem often looks with i=1,2,...,n and the train data (x_i,y_i).

$$min\ \ell(w) = \frac{1}{n} \sum_{i=1}^{n} \ell_i\ (h\ (x_i; w), yi) \tag{3}$$

where h is prediction function, i is the loss function, and w is the weights in the model.

4. PERFORMANCE EVALUATION

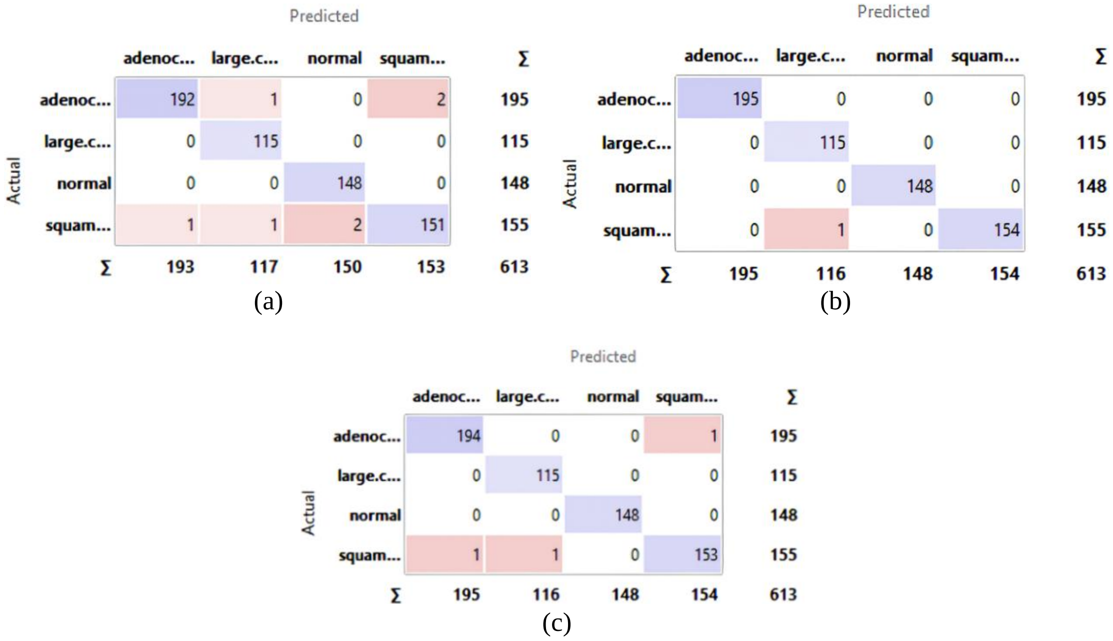
The performance evaluation approach allows for the verification of efficiency and validity. There are several methods to assess a classifier. For this inquiry, two sets of data test sets and a train set represent 30% and 70% of the total dataset, respectively. Using the invisible test set, the data's predicted success is assessed following training on the training set. We used the cross-validation method of 10 folds to further remove the over-fitting problem. Three MLML classification models were utilized in this paper: AdaBoost, RF, and SGD. The performance of each classifier was compared. Using a sample of test data, a ML tool called a confusion matrix indicates which predictions a classification model made correctly and inaccurately. According to [47]-[49], this category includes false positives (FP), true negatives (TN), true positives (TP), and false negatives (FN). Table 1 displays the confusion matrix. As can be shown in Table 2, it generates measures such as accuracy, precision, recall/sensitivity, and F1-score. Orange3, a data mining tool, was used to extract data. Figure 5(a) confusion matrix of the SGD classifier shows the results of testing the proposed model and calculating the performance of each classifier by constructing a confusion matrix. AdaBoost classifier confusion matrix (Figure 5(b)) and RF classifier confusion matrix (Figure 5(c)). Table 2 represents classifier's performance matrices.

Table 1. Confusion matrix [50]

	Predicted	
Actual	TP	FN
	FP	TN

Table 2. Classifier's performance matrices [50]

Performance matrices	Equation	Performance matrices	Equation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Precision	$\frac{TP}{TP+FP}$
Sensitivity	$\frac{TP}{TP+FN}$	F-measure	$\frac{2 * \text{Precision} * \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}}$



	adenoc...	large.c...	normal	squam...	Σ
adenoc...	192	1	0	2	195
large.c...	0	115	0	0	115
normal	0	0	148	0	148
squam...	1	1	2	151	155
Σ	193	117	150	153	613

(a)

	adenoc...	large.c...	normal	squam...	Σ
adenoc...	195	0	0	0	195
large.c...	0	115	0	0	115
normal	0	0	148	0	148
squam...	0	1	0	154	155
Σ	195	116	148	154	613

(b)

	adenoc...	large.c...	normal	squam...	Σ
adenoc...	194	0	0	1	195
large.c...	0	115	0	0	115
normal	0	0	148	0	148
squam...	1	1	0	153	155
Σ	195	116	148	154	613

(c)

Figure 5. Confusion matrix of all classifier; (a) confusion matrix of SGD classifier, (b) confusion matrix of AdaBoost classifier, and (c) confusion matrix of RF classifier

5. RESULTS AND DISCUSSION

The classification analysis's findings show a distinct difference between the AdaBoost, RF, and SGD models' performances on the dataset. AdaBoost outperformed the other models, achieving perfect scores for accuracy, AUC, and sensitivity, with near-perfect results (99.8%) in precision and F-measure as shown in Table 3. This highlights AdaBoost's robustness in handling the dataset, making it the most reliable model for this task. RF also demonstrated strong performance, with a 99.5% classification accuracy, precision, and F-measure, and a perfect AUC score, indicating its effectiveness. However, its slightly lower accuracy compared to AdaBoost suggests that while RF is highly reliable, it may not capture the dataset's nuances as well as AdaBoost. In contrast, SGD showed comparatively lower performance, with a classification accuracy of 98.9%. While this result is still robust, it indicates that SGD is less effective in this context compared to the other models. This difference in performance could be attributed to SGD's sensitivity to the specific characteristics of the dataset, such as its complexity and the presence of noise.

Table 3. Comparison of ML performance classifiers on the training dataset

Model	AUC	Accuracy	F-measure	Precision	Sensitivity
AdaBoost	1.000	0.998	0.998	0.998	1.000
RF	1.000	0.995	0.995	0.995	0.998
SGD	0.993	0.989	0.989	0.989	0.996

Overall, the results underscore the superiority of AdaBoost and RF in this classification task, with AdaBoost emerging as the most precise and consistent model. The lower performance of SGD suggests that it may not be the best choice for datasets with similar characteristics, particularly when high accuracy and reliability are crucial.

The SGD, AdaBoost, and RF classifiers' confusion matrices are analyzed to show important details about each of their particular characteristics. The SGD classifier's confusion matrix (Figure 5(a)) shows that it can correctly identify a significant number of scenarios, although it might have some issues with managing misclassifications across different categories. According to this, SGD may be more vulnerable to errors in some situations than the other models, even though it performs well overall. With fewer misclassifications across categories, the confusion matrix of the AdaBoost classifier (Figure 5(b)) shows its exceptional classification performance. This suggests that AdaBoost performs exceptionally well at the higher levels, which makes it a dependable option for applications demanding high recall and precision.

The accuracy and balanced error distribution of the RF classifier are demonstrated by the confusion matrix (Figure 5(c)). RF is a strong model that can handle a variety of complex datasets, demonstrated by its ability to maintain low error rates across all categories. Nevertheless, depending on the particulars of the dataset, it can perform slightly differently than AdaBoost. While all three models are successful, AdaBoost, and RF perform better in terms of accuracy and error distribution, according to the confusion matrices, which offer a thorough comparison of the classifiers overall. The evaluation underscores the importance of choosing the appropriate classifier based on the specific requirements of the task at hand, with AdaBoost and RF emerging as the more reliable options for high-stakes classification tasks.

The comparative analysis of classifier performance, as illustrated in Figure 6, offers a detailed overview of how each model handles the dataset across key metrics such as accuracy, precision, recall, and F1-scores. This visual representation underscores the varying strengths and weaknesses of the classifiers, enabling a deeper understanding of their relative effectiveness.

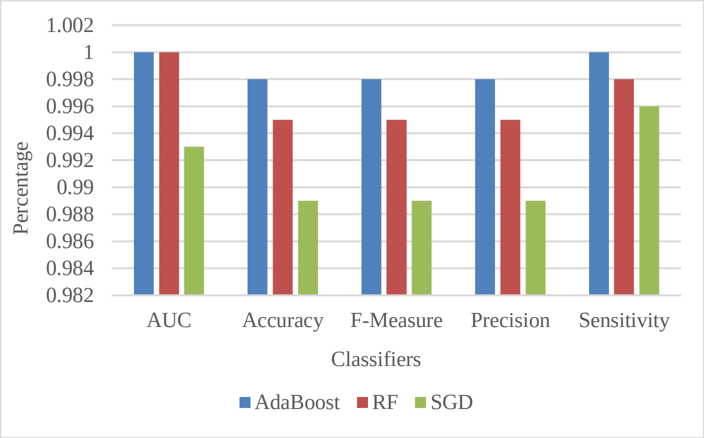


Figure 6. Comparative analysis of different classifiers' performances

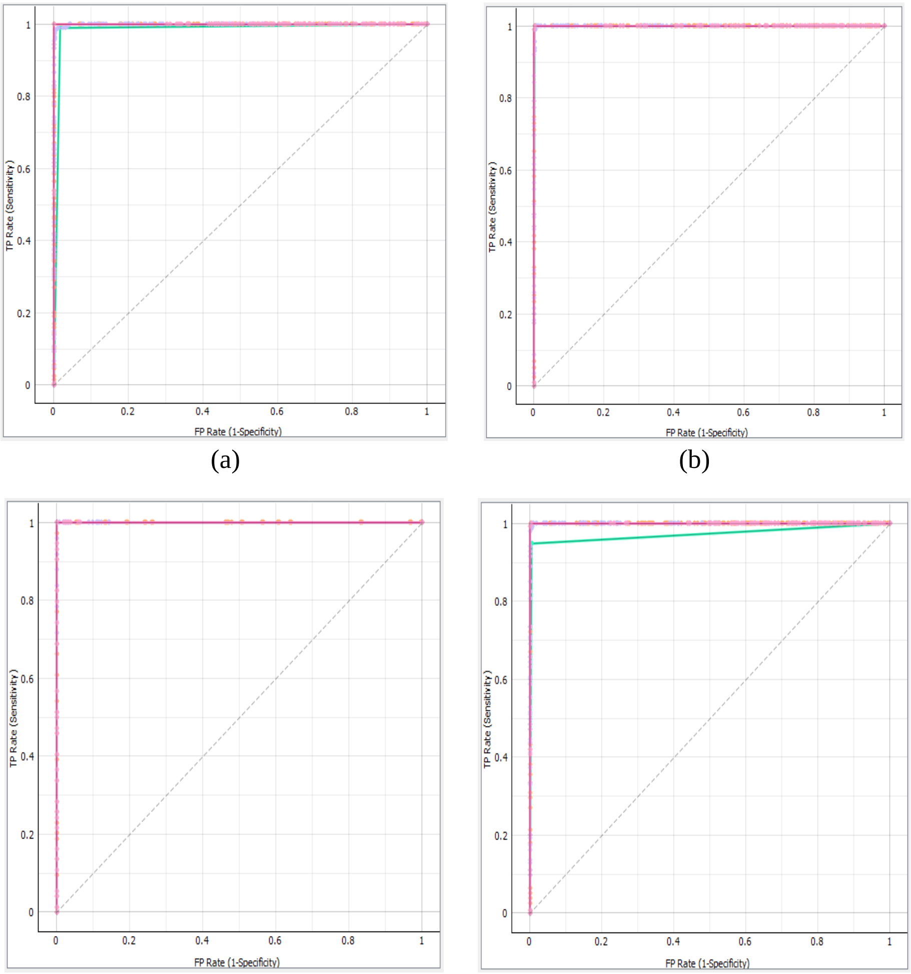
On the other hand, the differences in accuracy among the classifiers suggest that some models are better at correctly predicting outcomes, making them more reliable for tasks where accuracy is paramount. Precision and recall, on the other hand, highlight the classifiers' ability to minimize FP and FN, respectively. The variation in these metrics indicates that certain classifiers may be more suitable for tasks where either precision or recall is particularly critical. Also, the F1-scores, which balance precision and recall, provide a holistic view of each model's performance. The observed differences in F1-scores reflect the trade-offs each classifier makes between precision and recall, offering insights into which model may achieve the best overall balance for specific tasks.

Overall, the results demonstrate that no single classifier universally outperforms the others across all metrics. Instead, the choice of classifier should be guided by the specific requirements of the task at hand. For instance, if the priority is to minimize FP, a model with higher precision may be preferred. Conversely, if the goal is to reduce FN, a model with higher recall would be more appropriate. The analysis in Figure 5 thus provides a valuable tool for selecting the most suitable classifier based on the desired performance outcomes. However, Table 4 depicts a comparison of previous studies with the proposed model.

Table 4. Comparison of previous studies with the proposed model

Study/year	Model	Result
Lin *et al.* [22]	RF	AUC of 0.700.
Saba [23]	RF classification	Utilized feature extraction (area, perimeter, and eccentricity) for effective lung cancer detection.
Lin *et al.* [24]	SVM with texture matching	Implemented LBP for superior texture matching, enhancing detection accuracy.
Gupta *et al.* [25]	Image preprocessing techniques	Focused on noise reduction and morphological opening to improve CT image quality for better segmentation.
Sim *et al.* [26]	SVM	Categorized CT images into normal and abnormal with high accuracy in early-stage cancer detection.
Vijayalakshmi *et al.* [29]	Histogram equalization and HE	Emphasized the significance of classification and feature extraction using HE for CT images.
Proposed	AdaBoost, RF, and SGD	AUC: 1.000, accuracy: 0.998, F-measure: 0.998. Precision: 0.998 and sensitivity: 1.000.

Additionally, the receiver operating characteristic (ROC) curve, which illustrates the trade-offs between TP and FP rates, is used to evaluate the effectiveness of binary classifiers in computational statistics and ML. For assessing the effectiveness of classifiers in a variety of applications, the ROC curve is essential [51]. The analysis provides a better understanding of the advantages and disadvantages of each classifier in real-world situations by examining AI and actual goal results [52]. Figure 7 depicts the ROC curve that is utilized to assess the effectiveness of classifiers: (a) ROC curve analysis according to adenocarcinoma, (b) ROC curve analysis according to large.cell.carcinoma, (c) ROC curve analysis according to normal, and (d) ROC curve analysis according to squamous.cell.carcinoma.



(a)

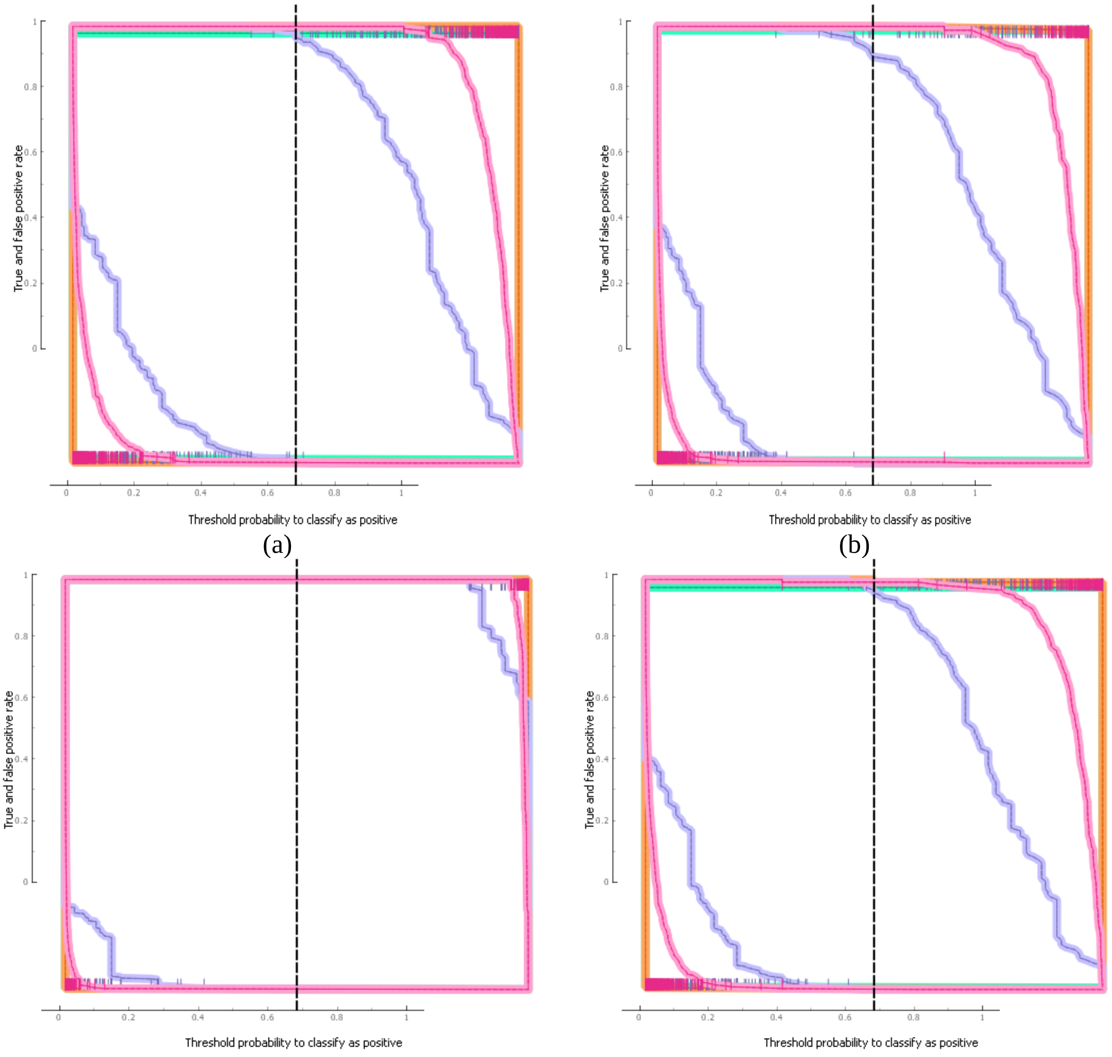
(b)

(c)

(d)

Figure 7. ROC curves analysis of all classifiers according to target; (a) ROC curve analysis according to adenocarcinoma, (b) ROC curve analysis according to large.cell.carcinoma, (c) ROC curve analysis according to normal, and (d) ROC curve analysis according to squamous.cell.carcinoma

In addition, this work assessed the probabilistic classifier performance to make sure the results match reality using calibration charts that show the expected probability against the actual results in Orange data mining tool. An accurate predictions are indicated by points near the diagonal line in a well-calibrated model in order to make well-informed adjustments like Platt scaling or isotonic regression, calibration charts assist in identifying any overconfidence or underconfidence in predictions [50]. However, Figure 8 depicts the calibration plot that is utilized to assess the effectiveness of classifiers: (a) calibration plot analysis according to adenocarcinoma, (b) calibration plot analysis according to large.cell.carcinoma, (c) calibration plot analysis according to normal, and (d) calibration plot analysis according to squamous.cell.carcinoma.



(a) (b)

(c) (d)

Figure 8. Calibration plot analysis of all classifiers according to target; (a) calibration plot analysis according to adenocarcinoma, (b) calibration plot analysis according to large.cell.carcinoma, (c) calibration plot analysis according to normal, and (d) calibration plot analysis according to squamous.cell.carcinoma

6. CONCLUSION

In conclusion, this work highlights the great potential of ML techniques for using medical imaging data to diagnose and categorize lung cancer. According to the results, our suggested model is a potentially very useful tool in clinical settings because it not only outperforms in accuracy but also in sensitivity and specificity. Improving early detection can result in prompt therapies and better patient outcomes, therefore this skill is essential. It is imperative to recognize the limits of the study, though. Future research must concentrate on confirming the model's resilience and generalizability using bigger and more varied datasets. Additionally, incorporating a wider array of clinical variables could further enhance predictive performance, making the model even more reliable for real-world applications. The exceptional results achieved by AdaBoost reflect its potential for flawless identification, while RF and SGD also demonstrated commendable efficacy. These findings highlight the effectiveness of ML in oncology and emphasize the need for ongoing research and refinement. By continuing to integrate diverse data and methodologies, we can fully harness the power of ML in lung cancer detection, ultimately advancing the field and improving patient care.

7. FUTURE WORK

In the future larger datasets from different institutions and imaging devices should be used to validate and expand the suggested ML model. To increase classification accuracy, sophisticated DL architectures such as ResNet, DenseNet, and EfficientNet required to be incorporated. In order to integrate imaging findings and clinical factors, multimodal data fusion should be investigated. In addition of integrating explainable AI methods such as SHAP, LIME, or Grad-CAM is necessary to increase transparency and comprehension of the model's decision-making process and collaboration with hospitals for clinical trials is important for evaluating the accuracy and usability of the model, and it should be tuned for real-time and edge deployment. However, ethical, legal, and regulatory considerations are also essential for the model's safe and ethical deployment. However, these limitations must be reduced to sample size, potential biases, and generalizability in order to improve focus and clarity. Improving the model's accuracy and applicability for a range of clinical scenarios while advancing equitable and inclusive healthcare practices requires addressing these issues and making confident that it is continuously assessed and improved.

8. LIMITATION

The ML model for lung cancer detection faces several limitations, including lengthy processing times, low-quality images, irregular imaging protocols, and dependence on specific feature extraction methods, which can hinder its applicability in clinical settings and affect reliability and flexibility. Additionally, challenges such as small sample size, potential biases, and difficulties with generalizability arise from variations in image quality, limited CT scans, selection bias, labeling bias, and over-reliance on extracted features, all of which impact diagnostic accuracy. Furthermore, healthcare model development necessitates patient privacy, HIPAA compliance, and robust security measures. Deployment barriers, including technological infrastructure, integration challenges, and financial constraints, can complicate model implementation. Biases resulting from skewed training data or algorithmic assumptions may lead to inequitable treatment recommendations. Therefore, addressing these challenges and ensuring continuous assessment and refinement are essential for improving the model's precision and suitability for diverse clinical situations, while promoting fair and inclusive healthcare practices.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Hamza Abu Owida	✓			✓	✓	✓	✓	✓	✓	✓	✓		✓	
Areen Arabiat		✓	✓	✓	✓	✓			✓	✓	✓	✓		
Muhammad Al-Ayyad	✓		✓	✓			✓			✓	✓	✓	✓	
Muneera Altayeb	✓	✓		✓	✓					✓	✓	✓		

C	: **C**onceptualization	I	: **I**nvestigation	Vi	: **Vi**sualization
M	: **M**ethodology	R	: **R**esources	Su	: **Su**pervision
So	: **So**ftware	D	: **D**ata Curation	P	: **P**roject administration
Va	: **Va**lidation	O	: Writing - **O**riginal Draft	Fu	: **Fu**nding acquisition
Fo	: **Fo**rmal analysis	E	: Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The dataset that supports the findings of this study is openly available in Kaggle at https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images, reference number [27].

REFERENCES

[1] V. D. de Jager *et al.*, "Future perspective for the application of predictive biomarker testing in advanced stage non-small cell lung cancer," *The Lancet Regional Health - Europe*, vol. 38, pp. 1-9, Mar. 2024, doi: 10.1016/j.lanepe.2024.100839.

[2] S. J. Adams, E. Stone, D. R. Baldwin, R. Vliegenthart, P. Lee, and F. J. Fintelmann, "Lung cancer screening," *The Lancet*, vol. 401, no. 10374, pp. 390-408, 2023, doi: 10.1016/s0140-6736(22)01694-4.

[3] M. Altayeb, A. Arabiat, and A. Al-Ghraibah, "Detection and classification of pneumonia using the Orange3 data mining tool," *International Journal of Electrical & Computer Engineering*, vol. 14, no. 6, pp. pp. 6894-6903, 2024, doi: 10.11591/ijece.v14i6.pp6894-6903.

[4] M. S. Bhuiyan *et al.*, "Advancements in early detection of lung cancer in public health: a comprehensive study utilizing machine learning algorithms and predictive models," *Journal of Computer Science and Technology Studies*, vol. 6, no. 1, pp. 113-121, 2024, doi: 10.32996/jcsts.2024.6.1.12.

[5] N. M. Batouty *et al.*, "State of the art: Lung cancer staging using updated imaging modalities," *Bioengineering*, vol. 9, no. 10, pp.1-21, 2022, doi: 10.3390/bioengineering9100493.

[6] S. M. Bhavi *et al.*, "Syzygium malaccense leaf extract-mediated silver nanoparticles: synthesis, characterization, and biomedical evaluation in Caenorhabditis elegans and lung cancer cell line," *Green Chemistry Letters and Reviews*, vol. 18, no. 1, pp. 1-15, 2025, doi: 10.1080/17518253.2025.2456624.

[7] A. Panunzio and P. Sartori, "Lung cancer and radiological imaging," *Current Radiopharmaceuticals*, vol. 13, no. 3, pp. 238-242, 2020, doi: 10.2174/1874471013666200523161849.

[8] G. A. Silvestri *et al.*, "Outcomes from more than 1 million people screened for lung cancer with low-dose CT imaging," *Chest*, vol. 164, no. 1, pp. 241-251, 2023, doi: 10.1016/j.chest.2023.02.003.

[9] M. Salehjahromi *et al.*, "Synthetic PET from CT improves diagnosis and prognosis for lung cancer: Proof of concept," *Cell Reports Medicine*, vol. 5, no. 3, 2024, doi: 10.1016/j.xcrm.2024.101463.

[10] Y. Ohno *et al.*, "State of the art MR imaging for lung cancer TNM stage evaluation," *Cancers*, vol. 15, no. 3, pp. 1-22, 2023, doi: 10.3390/cancers15030950.

[11] J. Kufel *et al.*, "Measurement of Cardiothoracic Ratio on Chest X-rays Using Artificial Intelligence—A Systematic Review and Meta-Analysis," *Journal of Clinical Medicine*, vol. 13, no. 16, pp, 1-16, 2024, doi: 10.3390/jcm13164659.

[12] J. Porto-Álvarez *et al.*, "Digital medical x-ray imaging, cad in lung cancer and radiomics in colorectal cancer: Past, present and future," *Applied Sciences*, vol. 13, no. 4, pp. 1-31, 2023, doi: 10.3390/app13042218.

[13] M. Alavinejad, M. Shirzad, M. J. Javid-Naderi, A. Rahdar, S. Fathi-Karkan, and S. Pandey, "Smart nanomedicines powered by artificial intelligence: a breakthrough in lung cancer diagnosis and treatment," *Medical Oncology*, vol. 42, no. 5, p. 134, 2025, doi: 10.1007/s12032-025-02680-x.

[14] L. Hussain, M. S. Almaraashi, W. Aziz, N. Habib, and S. U. R. S. Abbasi, "Machine learning-based lungs cancer detection using reconstruction independent component analysis and sparse filter features," *Waves in Random and Complex Media*, vol. 34, no. 1, pp. 226–251, Jan. 2024, doi: 10.1080/17455030.2021.1905912.

[15] M. Baniata, S. Abuowaida, M. Aljaidi, M. Kharabsheh, A. Alsarhan, and A. A. Alsuwaylimi, "A Multi-Modal Attention-Guided Network for Alzheimer's Disease Classification Using Deep Learning," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 27150-27158, 2025, doi: 10.48084/etasr.12510.

[16] F. A. Altuhaifa, K. T. Win, and G. Su, "Predicting lung cancer survival based on clinical data using machine learning: A review," *Computers in Biology and Medicine*, vol. 165, pp. 1-16, 2023, doi: 10.1016/j.compbiomed.2023.107338.

[17] U. Chandran, J. Reps, R. Yang, A. Vachani, F. Maldonado, and I. Kalsekar, "Machine learning and real-world data to predict lung cancer risk in routine care," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 32, no. 3, pp. 337-343, 2023, doi: 10.1158/1055-9965.EPI-22-0873.

[18] A. Heidari, N. J. Navimipour, M. Unal, and S. Toumaj, "Machine learning applications for COVID-19 outbreak management," *Neural Computing and Applications*, vol. 34, no. 18, pp. 15313-15348, 2022, doi: 10.1007/s00521-022-07424-w.

[19] E. Alpaydin, *Machine learning*, The MIT Press, 2021, doi: 10.7551/mitpress/13811.001.0001.

[20] A. M. Arabiat, "Intelligent Model for Detecting GAN-Generated Images Based on Multi-Classifier and Advanced Data Mining Techniques," *International Journal of Electrical and Electronic Engineering and Telecommunications*, vol. 14, no. 3, pp. 147–157, 2025, doi: 10.18178/ijeetc.14.3.147-157.

[21] M. Arabiat *et al.*, "Enhanced accuracy of deep learning method for fruit images classification," in *2024 25th International Arab Conference on Information Technology (ACIT)*, Zarqa, Jordan, 2024, pp. 1-9, doi: 10.1109/ACIT62805.2024.10877231.

[22] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8-17, 2015, doi: 10.1016/j.csbj.2014.11.005.

[23] T. Saba, "Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges," *Journal of Infection and Public Health*, vol. 13, no. 9, pp. 1274-1289, 2020, doi: 10.1016/j.jiph.2020.06.033.

[24] J. Lin, Y. Yu, X. Zhang, Z. Wang, and S. Li, "Classification of histological types and stages in non-small cell lung cancer using radiomic features based on CT images," *Journal of Digital Imaging*, vol. 36, no. 3, pp. 1029-1037, 2023, doi: 10.1007/s10278-023-00792-2.

[25] S. H. Gupta, S. Goel, M. Kumar, A. Rajawat, and B. Singh, "Design of terahertz antenna to detect lung cancer and classify its stages using machine learning," *Optik*, vol. 249, p. 168271, 2022, doi: 10.1016/j.ijleo.2021.168271.

[26] J.-a. Sim *et al.*, "The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: applications of machine learning," *Scientific Reports*, vol. 10, no. 1, pp. 1-12, Jul. 2020, doi: 10.1038/s41598-020-67604-3.

[27] K. Pradhan and P. Chawla, "Medical Internet of things using machine learning algorithms for lung cancer detection," *Journal of*

Management Analytics, vol. 7, no. 4, pp. 591-623, 2020, doi: 10.1080/23270012.2020.1811789.

[28] J. Pati, "Gene expression analysis for early lung cancer prediction using machine learning techniques: An eco-genomics approach," *IEEE Access*, vol. 7, pp. 4232-4238, 2018, doi: 10.1109/ACCESS.2018.2886604.

[29] D. Vijayalakshmi, M. K. Nath, and M. Mishra, "Novel Pre-processing Stage for Classification of CT Scan Covid-19 Images," in *Proceedings of the 18th International Conference on Signal Processing and Multimedia Applications SIGMA*, 2021, vol. 1, pp. 87-94, doi: 10.5220/0010625200870094.

[30] M. Hany. "Chest CT-Scan images Dataset," Available: https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images.

[31] M. Hao, Q. Sun, C. Xuan, X. Zhang, and M. Zhao, "SqueezeNet: an improved lightweight neural network for sheep facial recognition," *Applied Sciences*, vol. 14, no. 4, pp. 1-13, 2024, doi: 10.3390/app14041399.

[32] M. Altayeb and A. Arabiat, "Crack detection based on mel-frequency cepstral coefficients features using multiple classifiers," *International Journal of Electrical & Computer Engineering*, vol. 14, no. 3, pp. 3332–3341, 2024, doi: 10.11591/ijece.v14i3.pp3332-3341.

[33] I. Bakkouri and S. Bakkouri, "2MGAS-Net: multi-level multi-scale gated attentional squeezed network for polyp segmentation," *Signal, Image and Video Processing*, vol. 18, no. 6, pp. 5377-5386, 2024, doi: 10.1007/s11760-024-03240-y.

[34] P. Palimkar, R. N. Shaw, and A. Ghosh, "Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach," in *Lecture Notes in Networks and Systems*, vol. 218, 2022, pp. 219–244, doi: 10.1007/978-981-16-2164-2_19.

[35] C. B. Pande *et al.*, "Characterizing land use/land cover change dynamics by an enhanced random forest machine learning model: a Google Earth Engine implementation," *Environmental Sciences Europe*, vol. 36, no. 1, pp. 1-23, 2024, doi: 10.1186/s12302-024-00901-0.

[36] A. Niyogi, T. A. Ansari, S. K. Sathapathy, K. Sarkar, and T. Singh, "Machine learning algorithm for the shear strength prediction of basalt-driven lateritic soil," *Earth Science Informatics*, vol. 16, no. 1, pp. 899-917, 2023, doi: 10.1007/s12145-023-00950-8.

[37] M. A. Alhariri, "Early Detection of Similar Fake Accounts on Twitter Using the Random Forest Algorithm," *Academia.Edu*, vol. 11, no. 12, pp. 611–620, 2020, doi: 10.34218/IJARET.11.12.2020.064.

[38] H. A. Salman, A. Kalakech, and A. Steiti, "Random forest algorithm overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69-79, 2024, doi: 10.58496/BJML/2024/007.

[39] W. Wang and D. Sun, "The improved AdaBoost algorithms for imbalanced data classification," *Information Sciences*, vol. 563, pp. 358-374, 2021, doi: 10.1016/j.ins.2021.03.042.

[40] J. Zheng, D. Lin, Z. Gao, S. Wang, M. He, and J. Fan, "Deep Learning Assisted Efficient AdaBoost Algorithm for Breast Cancer Detection and Early Diagnosis," *IEEE Access*, vol. 8, pp. 96946-96954, 2020, doi: 10.1109/ACCESS.2020.2993536.

[41] M. M. Abualhaj, A. Abdurrazaq, O. Almomani, and M. Anbar, "Optimized Feature Selection for Enhanced Network Attack Detection Using Bat Algorithm," in *2025 15th International Conference on Electrical Engineering (ICEENG)*, Cairo, Egypt, 2025, pp. 1-5, doi: 10.1109/ICEENG64546.2025.11031370.

[42] Y. Ding, H. Zhu, R. Chen, and R. Li, "An efficient AdaBoost algorithm with the multiple thresholds classification," *Applied Sciences*, vol. 12, no. 12, p. 5872, 2022, doi: 10.3390/app12125872.

[43] C. Wang, S. Xu, and J. Yang, "Adaboost algorithm in artificial intelligence for optimizing the IRI prediction accuracy of asphalt concrete pavement," *Sensors*, vol. 21, no. 17, pp. 1-16, 2021, doi: 10.3390/s21175682.

[44] S. Gamil, F. Zeng, M. Alrifaey, M. Asim, and N. Ahmad, "An efficient AdaBoost algorithm for enhancing skin cancer detection and classification," *Algorithms*, vol. 17, no. 8, pp. 1-19, 2024, doi: 10.3390/a17080353.

[45] AlmaBetter, "AdaBoost algorithm in Machine Learning". [Online]. Available: https://www.almabetter.com/bytes/tutorials/data-science/adaboost-algorithm, (Accessed: May 5, 2024).

[46] Y. Tian, Y. Zhang, and H. Zhang, "Recent advances in stochastic gradient descent in deep learning," *Mathematics*, vol. 11, no. 3, pp. 1-23, 2023, doi: 10.3390/math11030682.

[47] A. M. Arabiat and Y. G. Eljaafreh, "Intrusion Detection in Wireless Sensor Networks Using ML Based Classification of Denial of Service (DoS) Attacks," *Journal of Communications*, vol. 20, no. 4, 2025, doi: 10.12720/jcm.20.4.501-514.

[48] M. Altayeb and A. Arabiat, "A sustainable system for predicting appliance energy consumption based on machine learning," *Journal of Environmental Management*, vol. 382, p. 125434, 2025, doi: 10.1016/j.jenvman.2025.125434.

[49] M. A. Almaiah, L. M. Saqr, L. A. Al-Rawwash, L. A. Altellawi, R. Al-Ali, and O. Almomani, "Classification of Cybersecurity Threats, Vulnerabilities and Countermeasures in Database Systems," *Computers, Materials & Continua*, vol. 81, no. 2, 2024, doi: 10.32604/cmc.2024.057673.

[50] O. Almomani, A. Alsaaidah, M. A. Almaiah, A. Alzaqebah, M. M. Abualhaj and W. J. Alzyadat, "Evaluating Machine Learning Classifiers for Detecting Distributed Denial of Service Attacks," in *2025 12th International Conference on Information Technology (ICIT)*, Amman, Jordan, 2025, pp. 134-140, doi: 10.1109/ICIT64950.2025.11049130.

[51] E. Richardson, R. Trevizani, J. A. Greenbaum, H. Carter, M. Nielsen, and B. Peters, "The receiver operating characteristic curve accurately assesses imbalanced datasets," *Patterns*, vol. 5, no. 6, 2024, doi: 10.1016/j.patter.2024.100994.

[52] L. Farokhah and S. Y. Riska, "'Analysis and development of eight deep learning architectures for the classification of mushrooms," *Jurnal RESTI*, vol. 8, no. 1, pp. 142–149, Feb. 2024, doi: 10.29207/resti.v8i1.5498.

BIOGRAPHIES OF AUTHORS

Hamza Abu Owida Ph.D. in biomedical engineering, Assistant Professor at the Department of Medical Engineering, Al-Ahliyya Amman University, Jordan. Research interests focused on biomedical sensors, nanotechnology, and tissue engineering. He can be contacted at email: h.abuowida@ammanu.edu.jo.

Areen Arabiat earned her B.Sc. in Computer Engineering in 2009 from al Balqaa Applied University (BAU), and her M.Sc. in Intelligent Transportation Systems (ITS) from Al Ahliyya Amman University (AAU) in 2022. She is currently a computer lab supervisor, researcher and (part-time) lecturer in the Department of Communications and Computer Engineering at Al-Ahliyya Amman University since 2013. Her research interests are focused on the following areas: machine learning, data mining, image processing, internet of things (IoT), cyber security, wireless sensor network (WSN), artificial intelligence, and Intelligent Transportation System (ITS). She can be contacted at email: a.arabiat@ammanu.edu.jo.

Muhammad Al-Ayyad Ph.D. in electrical engineering, biomedical instrumentation, Associate Professor at the Department of Medical Engineering, Al-Ahliyya Amman University, Jordan. His research interests are big home healthcare settings, medical rehabilitation instrumentation, and biomedical measurements. He can be contacted at email: mayyad@ammanu.edu.jo.

Muneera Altayeb earned a bachelor’s degree in computer engineering in 2007, and a master’s degree in communications engineering from the University of Jordan in 2010. She has been working as a lecturer in the Department of Communications and Computer Engineering at Al-Ahliyya Amman University since 2015. She currently holds the position of Assistant Dean at the Faculty of Engineering at Amman Al-Ahliyya University. Her research interests focus on the following areas: digital signals and image processing, machine learning, robotics, and AI. She can be contacted at email: m.altayeb@ammanu.edu.jo.