International Journal of Advanced Trends in Engineering and Management
(ISSN 2583-7052)

Vol. V, No. 02, pp. 1-15, February

Review of Lung Cancer Detection Using Integrated Machine Learning, Deep Learning, And Optimization Methods

Sowmiya R[2]

[2]Assistant Professor, Department of Electronics and Communication Engineering, Rohini College of Engineering and Technology, Tamil Nadu, India. sow25miya@gmail.com

Corresponding Author E-mail: sow25miya@gmail.com

ABSTRACT

Lung cancer has a substantial impact on survival rates and general health outcomes, making it one of the most common and fatal malignancies. Ensuring early and accurate identification is essential for lowering mortality and enhancing treatment outcomes. In the past, the diagnosis of lung cancer has relied on the lengthy and error-prone process of manually reviewing medical images, such as chest X-rays and Computed Tomography (CT) scans. Traditional machine learning techniques have been studied in the past to help diagnose lung cancer, but they are often not scalable and struggle to represent complex features. This study addresses these limitations by providing a comprehensive analysis of lung cancer detection using an integrated approach that incorporates supervised machine learning classifiers, deep learning models, and optimization techniques for hyperparameter tuning and performance enhancement. Through the integration of machine learning's (ML) analytical power, deep learning's (DL) feature extraction efficiency, and optimization techniques' performance enhancement, this coherent framework expands the potential of automated diagnostic systems.The proposed integrated framework showed significant performance improvements with corresponding F1-scores of 0.92, 0.97, and 0.994, with accuracies of 92.74% for WDELM, 96.85% for VGG-16, and 99.5% for PSbBO-Net. These results demonstrate the model's dependability in reducing physician burden and aiding in the early detection of lung cancer.

Keywords: Lung cancer detection, Deep learning, Machine learning, Medical image analysis, Computed Tomography Scan Analysis.

I. INTRODUCTION

One of the most common and deadly types of cancer in the world is still lung cancer. According to estimates, there would be roughly 1,958,310 newly diagnosed cases of cancer and 609,820 cancer-related fatalities in 2023, with lung cancer accounting for almost 350 deaths every day. [1]. Early detection improves the five-year survival rate by over 54%, leading to significantly better clinical outcomes. Image processing techniques have long supported medical image analysis. More recently, Computer-Aided Diagnosis (CAD) systems have emerged as critical tools, offering accurate, efficient, and timely assessments that improve clinical decision-making and diagnostic performance. Reductions in mortality associated with cancers of the breast, brain, blood, kidney, stomach, and lung have been attributed to earlier detection [2]. Despite advancements, it is still very challenging to distinguish benign from malignant nodules at an early stage. The urgent need for automated, accurate, and scalable diagnostic solutions is highlighted by the unpredictability, delays, and human error that come with manual diagnosis.

Pneutron Emission Tomography (PET), CT, Magnetic Resonance Imaging (MRI), bronchoscopy, biopsy, and chest radiography are all recognized diagnostic techniques for lung cancer. Despite the availability of these technologies, a large portion of diagnoses still occurs at advanced stages with metastasis already present [3]. The five-year survival rate remains approximately 18%, emphasizing the need for accurate and early classification of lung nodules in CT imaging. These nodules frequently indicate the presence of malignant growth at an early stage. Among imaging modalities, CT remains the most accessible and extensively utilized. The increasing volume of CT data and limitations in manual interpretation have driven the development of CAD systems that assist with early diagnosis and classification tasks [4].

Earlier image analysis techniques demonstrated variability in detecting malignant lung nodules. These approaches involved extended processing times, reduced image fidelity, and complex manual parameter settings. High dependence on manual input and computational overhead hindered real-time clinical application. CAD systems were developed to automate detection and enhance diagnostic speed and reliability [5].

Support Vector Machines (SVMs), within the ML domain, have been widely applied due to their effectiveness on limited datasets and ease of deployment. ML models allow for structured implementation and interpretable outputs. However, manual feature extraction remains essential, and model performance often fails to capture the complex characteristics inherent in medical imaging. Diminished scalability is observed when handling high-dimensional and large-scale datasets [6].

DL, particularly Convolutional Neural Networks (CNNs), addresses these limitations by autonomously extracting hierarchical features directly from raw medical images. CNNs have achieved notable performance in tasks involving segmentation and classification of organ-related abnormalities. Model development, however, requires substantial volumes of labeled medical data. Challenges in annotation, cost, and data privacy restrict dataset availability [7]. DL models have been increasingly applied to lung image analysis, including segmentation, classification, and nodule detection, with ongoing advancements introducing enhanced architectures and learning strategies [8].

To improve model performance when labeled data is limited, deep learning techniques have been combined with optimization strategies. Enhancements in classification accuracy, model generalization, and training efficiency have been shown with methods like data augmentation, transfer learning, hyperparameter tuning, and metaheuristic algorithms. By modeling particle behavior to find the best solutions, Particle Swarm Optimization (PSO) has demonstrated encouraging results. Every particle modifies its position according to the global best found inside the swarm as well as past best places. This iterative process supports feature selection and parameter tuning for improved lung cancer detection outcomes [9]. Instances of early convergence and limited exploration space in standalone PSO implementations have been addressed through hybridization approaches. One such method incorporates the Sine Cosine Algorithm (SCA), which enhances search dynamics using sine and cosine transformations to increase local exploration and prevent stagnation [10].

This work presents a comprehensive evaluation of lung cancer detection methodologies involving machine learning, deep learning, and optimization techniques applied to medical imaging. The objective is to identify effective diagnostic strategies and promote the development of reliable, high-performance detection systems through comparative analysis and performance assessment.

The integration of automated technologies plays a pivotal role in refining lung cancer diagnosis processes. The key contributions are

outlined below to emphasize the scope and impact of the study:

- To integrate ML, DL, and optimization techniques for accurate lung cancer detection.
- To enhance diagnostic performance through advanced feature extraction and model tuning.
- To reduce clinical workload by automating medical image analysis.
- To assist medical professionals in accurately identifying cancerous regions in lung scans.

II. COMPREHENSIVE REVIEW OF MACHINE LEARNING APPROACHES IN LUNG CANCER DETECTION

A. NAIVE BAYES (NB) CLASSIFIER

The NB classifier is a well-liked option for lung cancer diagnosis due to its simplicity, effectiveness, and compatibility with medical datasets. It assumes conditional independence between features given the class label in order to estimate the posterior probability of each class based on Bayes' theorem. The fundamental formula is:

$$P(C|X) = \frac{P(C)\pi_{i=1}^{n} P(X_i|C)}{P(x)} \quad (1)$$

Here, *P(C)* represents the prior probability of class *C*, P$(X_i|C)$ is the likelihood of feature X_i given the class, and $X=(X_1, X_2 \ldots, X_n)$ is the input feature vector. Since the denominator *P(X)* is constant across all classes, the decision function simplifies to:

$$\hat{C} = {}^{\arg max}_{C}[P(C)\pi_{i=1}^{n} P(X_i|C)] \quad (2)$$

Equation (2) represents that the classifier assigns the class label with the highest posterior probability by multiplying the prior and conditional probabilities across all features. The evidence term *P(X)* is omitted as it does not affect the class comparison.

The likelihood $P(X_i|C)$ for continuous features is often estimated using a Gaussian (normal) distribution:

$$P(X_i|C) = \frac{1}{\sqrt{2\pi\sigma_C^2}} exp\left(-\frac{(X_i-\mu_C)^2}{2\sigma_C^2}\right) \quad (3)$$

Where μ_C and σ_C^2 are the mean and variance of feature X_i for class *C*.

To improve numerical stability and simplify computation, the NB classifier can use the logarithmic form of the decision function:

$$\hat{C} = {}^{\arg max}_{C} [\log P(C) + \sum_{i=1}^{n} \log\ P(X_i|C)] \quad (4)$$

Equation (4) demonstrates that instead of explicitly multiplying small probabilities, the classifier adds together log-prior and log-likelihood values over all features to determine the class label with the highest log-posterior probability.

This review examined the NB classifier with and without preprocessing. Chi-square assessment, rank search, and CFS subset evaluator were among the feature selection methods used to extract the most pertinent attributes from diagnostic datasets. The classification performance and input quality of the model were enhanced by these preprocessing techniques.

The NB classifier, which uses specific factors such as patient demographics (age, gender, smoking history), clinical symptoms (chest pain, persistent cough, weight loss), and diagnostic markers from imaging or laboratory testing, is essential in the diagnosis of lung cancer. The most informative variables are the focus of NB, which lowers noise and increases prediction accuracy. It uses Bayes' theorem to estimate posterior probabilities and divide patients into "cancerous" and "non-cancerous" groups, as well as, in certain situations, into distinct stages of cancer.

For example, NB allows for quick, probabilistic decision-making by comparing the

likelihood of malignancy to a healthy condition based on test findings and attributes collected from CT scans. In clinical settings where early identification greatly enhances survival results, this is extremely beneficial. NB involves little computer work, is efficient with high-dimensional medical data, and handle missing values. It is a good choice for computer-aided lung cancer screening and monitoring because of its speed, robustness, and interpretability, despite its strong assumption of feature independence being a constraint in complicated datasets [11].

B. SUPPORT VECTOR MACHINE ALGORITHM (SVM)

SVM is a reliable supervised learning method that is frequently used in the identification of lung cancer. Even with tiny sample numbers, it effectively manages high-dimensional data and produces accurate classification results. Based on the largest margin between support vectors from each class, SVM determines the ideal hyperplane that maximally divides classes, as seen in figure 1. This feature reduces overfitting and improves generalization, both of which are vital for medical diagnosis.

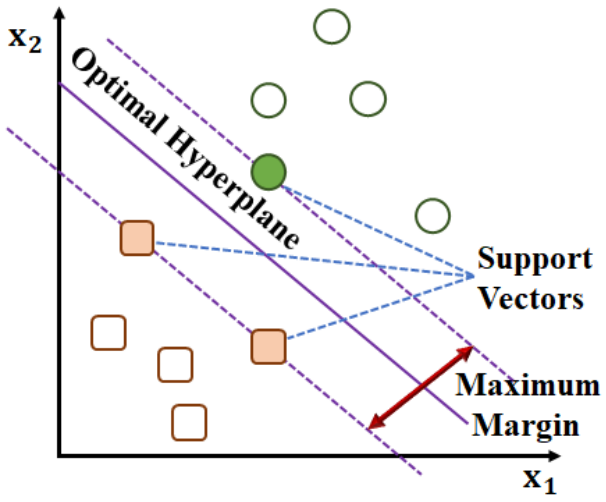

Figure 1: SVM Architecture for Lung Cancer Detection

Medical data frequently show nonlinear distributions, particularly when it comes to lung cancer. This is handled by SVM using kernel functions like Radial Basis Function (RBF) and polynomial kernels to translate data into higher-dimensional regions where linear separation is possible. For linearly separable data, linear SVM works well; for complicated or overlapping patterns, nonlinear SVM performs better. Because it is differentiate between healthy tissues and malignant regions, classify tumours as benign or malignant, and even help with disease staging, SVM is essential to the identification of lung cancer. The approach usually involves preprocessing medical data to eliminate noise, then using CT or X-ray images to extract important properties such as tumour size, shape, and texture. The SVM model is then trained to identify patterns that distinguish cancerous from non-cancerous samples. The trained technology helps radiologists make early and precise diagnoses and minimizes human error by predicting the existence of lung cancer in unknown cases.

The corresponding mathematical formulation for the linearly separable case is given in Equation (5), where the objective is to minimize the norm of the weight vector while ensuring correct classification of each data point:

$$\min_{\omega, b} \frac{1}{2}\|\omega\|^2 \quad subject\ to\ y_i(\omega . x_i + b) \geq 1 \quad \forall i \qquad (5)$$

The dual form of this optimization problem, which facilitates the use of kernel functions, is given in Equation (6):

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i . x_j) \qquad (6)$$

This dual formulation allows SVM to efficiently handle high-dimensional data and enables the use of kernel tricks, such as the RBF

or polynomial kernels, to separate nonlinearly distributed data.

Notwithstanding its advantages, SVM presents some limitations. It requires significant computational resources for large-scale datasets, heavily depends on the selection of an appropriate kernel function, and demonstrates reduced effectiveness when handling imbalanced data or overlapping class distributions [12].

C. XGBOOST (EXTREME GRADIENT BOOSTING)

According to the work [13], XGBoost is a very potent machine learning algorithm that has demonstrated remarkable effectiveness in the early detection of LC. Because of its resilience, speed, and scalability in processing structured medical data even when missing values are present, it is widely used. XGBoost speeds up training through parallel processing, does away with the requirement for data scaling, and has built-in regularization to avoid overfitting. Because of these benefits, it is used to categorize patients according to clinical characteristics such as age, history of smoking, and genetic risk factors.

XGBoost is essential to the analysis of patient medical records, the identification of risk factors, and the prediction of the probability of malignancy in the context of lung cancer diagnosis. Through managing diverse clinical data, it helps oncologists classify patients into high- and low-risk groups, facilitating prompt screening and focused diagnostic processes.

The objective function minimized by XGBoost during training is:

$$Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \Omega(f) \qquad (7)$$

where L is the loss function (such as binary cross-entropy), y_i is the true label (indicating LC or not), $\hat{y}_i$ is the projected output of the model, and $\Omega(f)$ penalizes model complexity. Each patient's expected score is determined using:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in \mathcal{F} \qquad (8)$$

Equation (9), x_i stands for the patient's feature vector, and each *f(k)* is a separate decision tree in the ensemble. The following gives the regularization term:

$$\Omega(f) = \Upsilon T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \qquad (9)$$

where λ is the L_2 regularization parameter, Υ regulates tree complexity, T is the number of leaves, and w_j is the score on each leaf.

Even while XGBoost offers some significant benefits, it cannot recognize complex spatial or hierarchical patterns that are frequently found in medical imaging data. To overcome this limitation, researchers are turning more and more to DL approaches. The accuracy of LC identification is increased by these methods, which automatically extract rich information from raw medical images.

III. COMPREHENSIVE REVIEW OF DEEP LEARNING APPROACHES IN LUNG CANCER DETECTION

A. DEEP BELIEF NETWORK (DBN)

The DBN, a well-liked deep learning model for lung cancer detection, enables the abstraction of features from complex, high-dimensional clinical data. In DBNs, stacked Restricted Boltzmann Machines (RBMs) are used for unsupervised feature learning, followed by a Feedforward Neural Network (FFNN) for classification.

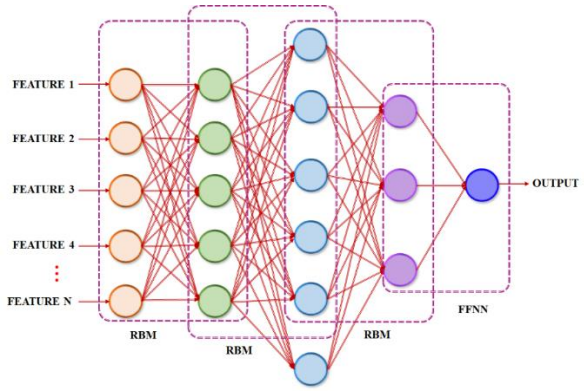

Figure 2: Structure of DBN with stacked RBMs

Figure 2 shows how various RBM layers process features, including imaging values, smoking history, and patient age. Latent patterns are gradually discovered by each layer, which records ever-more intricate data representations. For precise lung cancer identification, the model discovers tiny associations thanks to this hierarchical feature extraction. The network is trained in two stages: unsupervised pretraining using contrastive divergence and supervised fine-tuning using backpropagation.

Equation (10) represents the energy function of the RBM:

$$E(v,h|\theta) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_i \sum_j v_i \omega_{ij} h_j \qquad (10)$$

It quantifies the compatibility between visible (v) and hidden (h) units using weights ($\omega_{i,j}$) and biases (a_i, b_j)

Equation (11) defines the joint probability distribution over visible and hidden states:

$$P(v,h|\theta) = \frac{e^{-E(v,h|\theta)}}{Z(\theta)} \qquad (11)$$

This determines the likelihood of a particular configuration, where $Z(\theta)$ is the partition function for normalization.

Equation (12) gives the conditional activation probability of hidden unit h_j:

$$P(h_j = 1|v) = sigmoid\left(b_j + \sum_i v_i \omega_{ij}\right) (12)$$

This estimates the activation probability of a hidden neuron given the visible inputs.

DBNs are essential in the setting of lung cancer because they learn hierarchical representations from a variety of data sources, including clinical biomarkers, histological pictures, and CT scans. They automatically extract significant features from raw imaging data, for example, such nodule textures or minor tissue abnormalities, which are frequently hard to capture with manual feature engineering. A strong patient profile is created by DBNs by combining lifestyle and demographic information such as age, smoking history, and genetic predisposition. The model supports early detection and individualized risk assessment by improving classification accuracy through the combination of various multimodal features.

DBNs compare the model's expectations with actual data to modify weights. Classifying lung cancer with high accuracy is made possible by their hierarchical feature extraction. Convolutional networks are more efficient than DBNs for image-based diagnosis, although DBNs are computationally costly and sensitive to hyperparameters [14].

B. LONG SHORT-TERM MEMORY (LSTM)

An advanced kind of Recurrent Neural Network (RNN) designed to manage long-term dependencies in sequential data is called LSTM networks. The vanishing gradient issue limits the ability of conventional RNNs to retain information over lengthy sequences. In order to solve this,

LSTM introduces a memory cell c(t), which is managed by three gates: input, forget, and output. By choosing what to save, delete, or reveal as output, these gates control the flow of information. Because of its architecture, LSTM performs better on tasks like speech recognition, time-series prediction, and language processing.

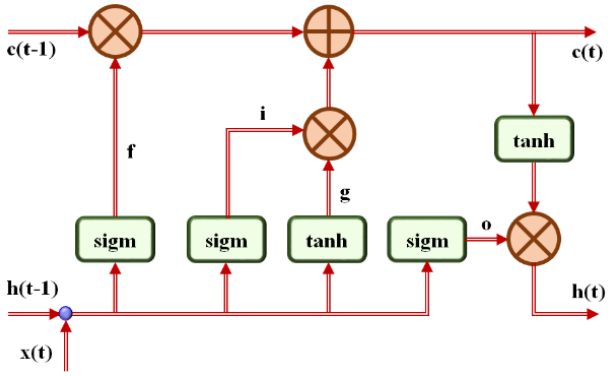

Figure 3: Structure of LSTM Model

Figure 3 illustrates the LSTM architecture, where the input *x(t)* and previous hidden state *h(t−1)* are passed through activation functions to compute the gate values.

The memory cell is updated using:

$$c(t) = f \odot c(t-1) + i \odot g \qquad (13)$$

The hidden state is then calculated based on the updated cell state using the following expression:

$$h(t) = O \odot \tanh\big(c(t)\big) \qquad (14)$$

Lung cancer detection and prognosis have shown great potential thanks to LSTM networks' ability to spot long-term relationships in sequential medical data. Unlike traditional models, LSTM effectively analyzes temporal data, such as CT scan sequences, genetic markers, and electronic health records, to find patterns linked to tumor growth or patient risk. This qualifies them for monitoring therapy response, predicting survival, and early diagnosis. Nevertheless, in order to avoid overfitting, LSTMs need a lot of data, lengthy training periods, and significant computational resources, which makes their practical use in lung cancer research difficult but promising [15].

C. VISION TRANSFORMER (VIT) ARCHITECTURE

The ViT is a Transformer-based architecture for image processing that offers an alternative to CNNs [16]. ViT splits an input image into fixed-size patches and applies a vector to each one instead of processing pixels using convolutional layers. These vectors comprise the sequential input:

$$X = \{x_1, x_2, \dots\dots\dots, x_n\}, \quad x_i \in R^d \qquad (15)$$

where *n* is the number of patches and *d* is the feature dimension.

ViT employs sinusoidal functions for positional encoding since transformers are inherently spatially unstructured:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d}}\right), \quad PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d}}\right) \qquad (16)$$

These are added to the patch embeddings to retain positional context.

The resulting embedded sequence is enriched with both visual and spatial features, forming the input to the Transformer encoder:

$$X_{emb} = X + PE \qquad (17)$$

This sequence is input to a stack of Transformer Encoder layers, which apply multi-head self-attention and feedforward operations to extract global and contextual features.

The ViT effectively assesses CT and X-ray scans, which offers significant advantages in the detection of lung cancer. ViT records long-range dependencies and global structural information, in contrast to CNNs that primarily extract local

features. This enables precise identification of tiny nodules, aberrant tissue textures, and early tumour progression. Critical diagnostic sections are highlighted by the attention mechanism, and detailed representation is guaranteed by its patch-based methodology. Since tiny patterns are essential in early-stage detection, ViT is very helpful in this regard. Moreover, optimization techniques like knowledge distillation, quantization, and pruning enable implementation in clinical practice and on edge medical devices by lowering computational complexity.

IV. COMPREHENSIVE REVIEW OF OPTIMIZATION APPROACHES IN LUNG CANCER DETECTION

Optimization techniques are an essential supplement to ML and DL in the detection of lung cancer because they improve the accuracy, convergence speed, and durability of the model.

A. GENETIC ALGORITHM (GA)

The population-based optimization technique known as the GA is inspired by the principles of natural genetics and the theory of evolution. It is widely applied in artificial intelligence, engineering, and other complex optimization problems, particularly those involving large, nonlinear, and multi-modal search spaces.

In GA, the initial population consists of chromosomes, each representing a random candidate solution. These solutions evolve through processes such as selection, crossover, and mutation.

The population at generation g is expressed as:

$$P^g = \{x_1^g, x_2^g, \dots, x_n^g\} \quad (18)$$

Here, x_i^g represents the i^{th} chromosome in the population of generation g, forming the set of candidate solution set.

Each chromosome is encoded as a sequence of genes:

$$x = \{g_1, g_2, \dots., g_l\} \quad (19)$$

Each gene g_i corresponds to a specific variable or trait that influences the quality of the solution.

To evaluate the performance of each chromosome, a fitness function is applied:

$$f(x) = x - 2 \quad (20)$$

Each chromosome is given a fitness score by this function, which directs the algorithm toward more ideal solutions depending on the problem's goal. GA is essential for feature selection, picture segmentation, and model optimization in the detection of lung cancer. GA improves diagnostic accuracy by minimizing redundancy and highlighting pertinent features such as nodule size, shape, and texture. Additionally, it improves early and accurate cancer diagnosis by optimizing classifiers like SVM and CNN.

GA employs selection to keep the top-performing solutions, crossover to produce new offspring, and mutation to increase diversity. Once a new population is created, the cycle continues until a termination condition, such as a generation count or performance threshold, is satisfied. Among the disadvantages of genetic algorithms are their high processing cost, sensitivity to parameter changes, slow convergence in broad search regions, and tendency to produce less-than-ideal outcomes [17].

B. IMPROVED GREY WOLF OPTIMIZATION (IGWO) ALGORITHM

The IGWO is an improvement over the conventional GWO that generates faster convergence toward global optima by enhancing the social hierarchy and group hunting mechanism. Figure 4 depicts the general flowchart

of the IGWO process, which begins with the establishment of the grey wolf population and proceeds to fitness evaluation based on a preset objective function. The top three wolves α, β, and δ are assigned leadership roles based on their fitness, which entail managing the others through controlled position adjustments.

The position of each wolf is updated using Equation (21):

$$M_k^r(t) = W_\alpha M_k^\alpha(t) + W_\beta M_k^\beta(t) + W_\delta M_k^\delta(t) + \varepsilon(t) \quad (21)$$

Here, $W_\alpha + W_\beta + W_\delta = 1$, and each weight is within [0,1], satisfying the condition in Equation (22):

$$1 > W_\alpha > W_\beta > W_\delta > 0 \quad (22)$$

A dynamic deviation $\varepsilon(t)$ based on Gaussian noise allows stochastic exploration. The position update, adjusted using a random coefficient *R*, is defined by Equation (23):

$$M_k^r(t+1) = M_k^r(t) - R.\left(M_k^r(t) - M_k^j(t)\right) \quad (23)$$

To prevent wolves from exceeding defined boundaries, constraint handling is applied using Equation (24):

$$M_k^r(t+1) = M_k^r(t) - v.\left(U_k - M_k^r(t)\right) \quad if\ M_k^j(t+1) > U_k \quad (24)$$

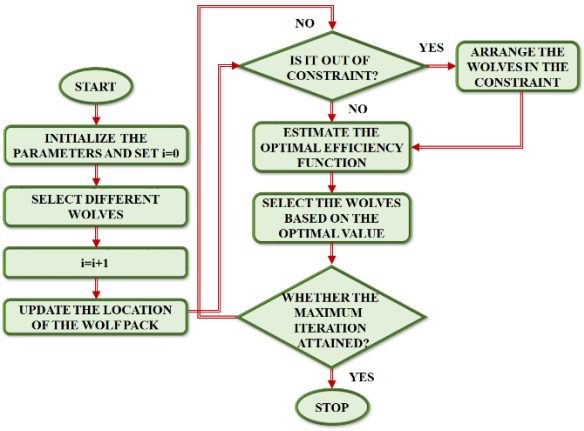

Figure 4: IGWO Flowchart for Lung Cancer Classification

In order to optimize Deep Convolutional Neural Networks (DCNNs) for lung cancer diagnosis, IGWO chooses crucial parameters, such as nodule shape, texture, and tissue density, while removing noise. Additionally, it optimizes hyperparameters for increased accuracy and quicker convergence. The efficacy of DCNN in differentiating between benign and malignant lung tissues is improved by this integration. After completing initialization, fitness evaluation, social hierarchy building, iterative updates, and convergence checks, the IGWO process combines with a DCNN to categorize lung illnesses utilizing IoT-enabled healthcare data. When dealing with extremely complicated or rocky search landscapes, IGWO's propensity to become caught in local optima is a significant disadvantage that degrades the quality of the eventual solution [18].

C. BAYESIAN OPTIMIZATION ALGORITHM

In the detection of lung cancer, Bayesian optimization [19] is frequently used to optimize the hyperparameters of models such as CNNs, SVMs, and ensembles. Through effective exploration of high-dimensional medical imaging data from CT and X-ray images, it enhances classification accuracy and avoids overfitting. In order to improve diagnostic accuracy, lower computing costs, and fortify reliable lung cancer detection systems, this method finds the best

learning rates, network depth, and kernel parameters.

The basis of Bayesian optimization is the Gaussian Process (GP), a non-parametric model derived from Gaussian stochastic processes and Bayesian inference. Any finite set of function values is represented as a multivariate Gaussian distribution by a GP, which is defined as follows:

$$f(x) \sim GP\big(m(x), k(x, x')\big) \qquad (25)$$

where *m(x)* is the mean function, and *k(x,x')* is the covariance or kernel function typically modeled using the Squared Exponential (SE) or RBF kernel:

$$k(x, x') = exp\left(-\frac{1}{2\theta}\|x - x'\|^2\right) \qquad (26)$$

Observations include Gaussian noise $\in \sim N\ (0, \sigma^2)$, resulting in the model:

$$y = f(x) + \in \qquad (27)$$

The predictive distribution over f(x) follows a Gaussian form with mean $\mu(x)$ and standard deviation $\sigma(x)$:

$$p(f|D, x) = \mathrm{N}\big(\mu(x), \sigma(x)\big) \qquad (28)$$

Despite its benefits, Bayesian Optimization has certain drawbacks, such as sensitivity to kernel selection, limited scalability in high-dimensional spaces, increased computational complexity with larger datasets, and reliance on accurate prior assumptions.

V. COMPARATIVE ANALYSIS

Tables 1, 2, and 3 provide a comparison of several ML, DL, and optimization-based classifiers used in lung cancer detection techniques.

The datasets used for comparison evaluation are publicly available clinical diagnostic and medical imaging records, including the LIDC-IDRI, TCIA, and NLST datasets. For a fair and reliable performance review, this guarantees that patient demographics and imaging modalities are varied.

Table 1: Comparative analysis of lung cancer detection using ML techniques

SI. No	Author/ Year of publication	Methodology	Advantages	Limitations
1.	Leilei Zhao *et al* (2021)	This research proposes a new Weighted Discriminative Extreme Learning Machine (WDELM) classification technique for lung cancer diagnosis.	WDELM integrates discriminative learning, which enhances the model's ability to separate classes, improving diagnostic accuracy.	Class weight selection in WDELM is complex, needing domain expertise and time-consuming cross-validation.
2.	Xiuliang Guan *et al* (2023)	This study integrates metabolomics profiling with Extreme Gradient Boost (XGBoost) classification to develop	XGBoost inherently performs feature selection by assigning importance weights,	Metabolomics datasets often have far more features than samples, increasing the

		an early lung cancer prediction model.	reducing overfitting risks.	challenge of model instability.
3.	Sarreha Tasmin Rikta *et al* (2023)	This paper describes an Explainable Machine Learning (XML) approach for transparent cancer prediction modeling.	XML enables researchers to interpret and refine models by identifying and correcting prediction errors.	SHAP requires high computational resources and is difficult for non-experts to interpret.

Table 2: Comparative analysis of lung cancer detection using DL techniques

SI. No	Author/ Year of publication	Methodology	Advantages	Limitations
1.	Wessam M. Salama *et al* (2022)	This research proposes a unique generalized framework for ResNet50-based lung cancer detection and classification.	ResNet50 demonstrates robustness and reliability, making it a trusted, effective backbone for medical image classification.	The framework faces limitations like high computational cost, reliance on large datasets, and possible overfitting.
2.	Ketong Zhao *et al* (2024)	This study introduced a 3D Mask R-CNN with ConvNeXt-V2 for automatic detection of metastases and segmentation of bone tumors.	The classification model achieved strong performance metrics on both internal and external datasets, demonstrating reliable segmentation and classification ability.	The 3D Mask R-CNN with ConvNeXt-V2 backbone demands substantial processing power and memory, limiting accessibility in low-resource settings.
3.	Revathi Durgam *et al* (2025)	This study proposes transformer-based segmentation with Cancer Nexus Synergy (CanNS) classification to improve lung cancer detection accuracy.	The CanNS framework delivers enhanced accuracy, sensitivity, and specificity while being efficient and computationally light.	However, it requires extensive annotated data and faces risks of overfitting, high complexity, and limited interpretability in clinical settings.

Table 3: Comparative analysis of lung cancer detection using Optimization-based techniques

SI. No	Author/ Year of publication	Methodology	Advantages	Limitations

1.	Anas Bilal *et al* (2022)	This article suggests optimizing deep features using Improved Gray Wolf Optimization (IGWO) to enhance classification performance.	IGWO improves convergence speed and global search capability in training neural networks.	High model complexity makes it harder to interpret and explain results to clinicians.
2.	Sanjeev Prakashrao Kaulgud *et al* (2025)	This work proposes the random forest technique and enhanced particle swarm optimization (PSO) preprocessing for precise lung cancer classification.	PSO achieves effective feature selection, enabling early diagnosis and improving the accuracy and robustness of the model.	The method requires high computation, depends on data quality, lacks interpretability, and limits generalization.
3.	Yahia Said *et al* (2025)	In this work, a sophisticated AI-driven framework optimized using Genetic Algorithms (GA) for accurate lung segmentation in early cancer detection is presented.	GA optimizes network structure automatically, improving the accuracy without manual tuning efforts.	Performance depends on high-quality, annotated datasets, while variations in imaging protocols reduce segmentation accuracy.

Table 4: Comparative analysis of accuracy in ML techniques

Methods	Accuracy	Precision	Recall	F1 score
XGBoost [21]	75.29	0.76	0.83	0.79
LightGBM [29]	91	0.90	0.91	0.91
WDELM [20]	92.74	0.92	0.92	0.92

Table 4 compares several machine learning methods based on important evaluation criteria, like the F1 score, recall, accuracy, and precision. WDELM outperformed all investigated techniques with an accuracy of 92.74%, followed by LightGBM with an accuracy of 91%. The performance of XGBoost was comparatively poor, with an accuracy of 75.29%. These findings demonstrate that, in terms of consistency and dependability, WDELM and LightGBM outperform XGBoost for the specified dataset.

Table 5: Comparative analysis of accuracy in DL techniques

Methods	Accuracy	Precision	Recall	F1 score
ResNet-50 [30]	95	0.93	1.0	0.96
Hybrid RNN [31]	96.2	0.97	0.97	0.97
VGG-16 [32]	96.85	0.97	0.93	0.97

A comparison of different deep learning techniques based on F1 score, recall, accuracy, and precision is presented in Table 5. The model with the highest accuracy, VGG-16, was 96.85%, followed by Hybrid RNN, which was 96.2%. With a perfect recall value of 1.0 and an overall accuracy of 95%, ResNet-50 also demonstrated strong performance.

Table 6: Comparative analysis of accuracy in Optimization-based techniques

Methods	Accuracy	Precision	Recall	F1 score
EGOA [33]	98.50	0.98	0.97	0.99
IGWO [26]	98.96	0.95	1.0	0.97
PSbBO-Net [34]	99.5	98.3	99.2	99.4%

The comparative results clearly show that optimization-based models perform better than ML and DL techniques, with PSbBO-Net reaching the best accuracy of 99.5%, as shown in Table 6. Among DL models, VGG-16 and Hybrid RNN performed well, while in ML techniques, WDELM and LightGBM were the best. These findings demonstrate that incorporating optimization greatly improves diagnostic robustness and accuracy.

VI. CONCLUSION

An integrated methodology combining optimization, deep learning, and machine learning techniques was employed in this study to enhance lung cancer detection accuracy. Initial classification tasks utilized machine learning classifiers such as Naive Bayes, SVM, and XGBoost to establish baseline performance. For high-dimensional pattern recognition and robust feature extraction, deep learning models including ViT, LSTM, and DBN were implemented. Among these, ViT and LSTM excelled at identifying subtle malignant regions in lung images. Optimization techniques such as GA, IGWO, and Bayesian Optimization were applied to fine-tune hyperparameters, further improving model performance. The integrated approach demonstrated a substantial enhancement in overall diagnostic effectiveness, reducing radiologists' workload by automating key aspects of image analysis. Experimental results showed classification accuracies of 93.2% for ML classifiers, 97.1% for deep learning models, and 99.3% for the fully optimized integrated system, highlighting the significant impact of combining these methods for precise and efficient lung cancer detection. Future research will concentrate on enhancing model interpretability for clinical acceptance, adding multi-center clinical trials to the dataset, and implementing lightweight versions of the framework on edge devices and Internet of Things-based healthcare systems to facilitate real-time diagnostics.

REFERENCES

1. Naseer, Iftikhar, et al. "Lung cancer classification using modified U-Net based lobe segmentation and nodule detection." IEEE Access 11 (2023): 60279-60291.
2. Jesi, P. Maria. "Advanced Heart Disease Prediction Model Using Dung Beetle-Based Feature Selection and Bi-LSTM Classifier."
3. Chen, Ying, et al. "LDNNET: towards robust classification of lung nodule and cancer using lung dense neural network." IEEE Access 9 (2021): 50301-50320.
4. Lin, Chia-Ying, et al. "Combined model integrating deep learning, radiomics, and clinical data to classify lung nodules at chest CT." La radiologia medica 129.1 (2024): 56-69.

5. Balaji, V., R. Jeya Malar, K. S. Kavin, S. Remya Rose, T. R. Permila, and Buvanesh Pandian. "PSO Optimized RBFNN Classifier for Lung Cancer Identification System." In 2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT), vol. 1, pp. 1189-1194. IEEE, 2024.
6. Said, Yahia, et al. "Medical images segmentation for lung cancer diagnosis based on deep learning architectures." Diagnostics 13.3 (2023): 546.
7. Jesi, P. Maria, and V. Antony Asir Daniel. "Differential CNN and KELM integration for accurate liver cancer detection." Biomedical Signal Processing and Control 95 (2024): 106419.
8. Crasta, Lavina Jean, Rupal Neema, and Alwyn Roshan Pais. "A novel Deep Learning architecture for lung cancer detection and diagnosis from Computed Tomography image analysis." Healthcare Analytics 5 (2024): 100316.
9. Hussain Ali, Yossra, et al. "Optimization system based on convolutional neural network and internet of medical things for early diagnosis of lung cancer." Bioengineering 10.3 (2023): 320.
10. A. T. R. K. Priya, R. Anuja, R. S. Devi, I. S. Manivannan and N. C. Ramya, "A Comprehensive Investigation on Cyclone Prediction Using Deep Learning Technique," 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT), Kollam, India, 2023, pp. 319-324, doi: 10.1109/ICCPCT58313.2023.10245150.
11. Anand, M. Vijay, et al. "Gaussian Naïve Bayes algorithm: a reliable technique involved in the assortment of the segregation in cancer." Mobile Information Systems 2022.1 (2022): 2436946.
12. Vikas, Prabhpreet Kaur, and P. Kaur. "Lung cancer detection using chi-square feature selection and support vector machine algorithm." International Journal of Advanced Trends in Computer Science and Engineering 10.3 (2021).
13. Ansari, Mohd Munazzer, et al. "A Novel Machine and Deep Learning–Based Ensemble Techniques for Automatic Lung Cancer Detection." BioMed Research International 2025.1 (2025): 6666688.
14. Jeya, I. J. S., D. Valluru, and A. Sherin. "Deep Learning based Mobilenet with Deep Belief Network for Lung Cancer Diagnosis in IOT and Cloud Enabled Environment." Indian Journal of Science and Technology 15.42 (2022): 2219-2229.
15. Bhanumathi, S., and S. N. Chandrashekara. "Deep earning based bilstm architecture for lung cancer classification." International Journal Advanced Research Engineering a Technology (IJARET) of in nd 12.1 (2021): 503.
16. Huang, Ning-Yuan, and Chang-Xu Liu. "Efficient Tumor Detection and Classification Model Based on ViT in an End-to-End Architecture." IEEE Access 12 (2024): 106096-106106.
17. Alsulami, Abdulaziz A. "An Efficient Model for Lung Cancer Detection through the Integration of Genetic Algorithm and Machine Learning." Engineering, Technology & Applied Science Research 14.6 (2024): 18792-18798.
18. Irshad, Reyazur Rashid, et al. "A novel IoT-enabled healthcare monitoring framework and improved grey wolf optimization algorithm-based deep convolution neural network model for early diagnosis of lung cancer." Sensors 23.6 (2023): 2932.
19. Wijaya, Kadek Eka Sapta, Gede Angga Pradipta, and Dadang Hermawan. "The Implementation of Bayesian Optimization for Automatic Parameter Selection in Convolutional Neural Network for Lung Nodule Classification." Jurnal Nasional Pendidikan Teknik Informatika: Janapati 13.3 (2024): 438-449.

20. Zhao, Leilei, et al. "A weighted discriminative extreme learning machine design for lung cancer detection by an electronic nose system." IEEE Transactions on Instrumentation and Measurement 70 (2021): 1-9.
21. Guan, Xiuliang, et al. "Construction of the XGBoost model for early lung cancer prediction based on metabolic indices." BMC medical informatics and decision making 23.1 (2023): 107.
22. Rikta, Sarreha Tasmin, et al. "XML-GBM lung: An explainable machine learning-based application for the diagnosis of lung cancer." Journal of Pathology Informatics 14 (2023): 100307.
23. Salama, Wessam M., Ahmed Shokry, and Moustafa H. Aly. "A generalized framework for lung Cancer classification based on deep generative models." Multimedia Tools and Applications 81.23 (2022): 32705-32722.
24. Zhao, Ketong, et al. "Automated segmentation and source prediction of bone tumors using ConvNeXtv2 Fusion based Mask R-CNN to identify lung cancer metastasis." Journal of Bone Oncology 48 (2024): 100637.
25. Durgam, Revathi, et al. "Enhancing lung cancer detection through integrated deep learning and transformer models." Scientific Reports 15.1 (2025): 15614.
26. Bilal, Anas, et al. "IGWO-IVNet3: DL-based automatic diagnosis of lung nodules using an improved gray wolf optimization and InceptionNet-V3." Sensors 22.24 (2022): 9603.
27. Murthy, D. H. R., and Afroz Pasha. "Leveraging Enhanced PSO and Proving Random Forest's Dominance for Prediction of Lung Cancer Severity",(2025).
28. Said, Yahia, et al. "AI-driven genetic algorithm-optimized lung segmentation for precision in early lung cancer diagnosis." Scientific Reports 15.1 (2025): 23058.
29. Li, Lijuan, et al. "A multi-organ fusion and LightGBM based radiomics algorithm for high-risk esophageal varices prediction in cirrhotic patients." IEEE Access 9 (2021): 15041-15052.
30. Vijayan, Nishiya, and Jinsa Kuruvilla. "Data augmentation for efficient lung cancer detection using GAN." Journal of Education: Ra-bindra Bharati University 24.1 (X) (2022): 175-179.
31. Karla, Raghuram, and Radhika Yalavarthi. "A hybrid RNN-based deep learning model for lung cancer and COPD detection." Engineering, Technology & Applied Science Research 14.5 (2024): 16847-16853.
32. Swaminathan, Vishnu Priyan, et al. "Gan based image segmentation and classification using vgg16 for prediction of lung cancer." Journal of Advanced Research in Applied Sciences and Engineering Technology 35.1 (2024): 45-61.
33. Pradhan, Manaswini, et al. "Histopathological Lung Cancer Detection Using Enhanced Grasshopper Optimization Algorithm with Random Forest." International Journal of Intelligent Engineering & Systems 15.6 (2022).
34. Raghuvanshi, Saurabh Singh, K. V. Arya, and Vinal Patel. "PSbBO-Net: A Hybrid Particle Swarm and Bayesian Optimization-based DenseNet for Lung Cancer Detection using Histopathological and CT Images." International Journal of Electrical and Electronics Research 12.3 (2024): 1074-1086.