International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(1s) 2026 pp. 1-16
https://doi.org/10.70917/ijcisim-2026-2020
Article

A Multi-Objective Hybrid Deep Learning Framework for Pulmonary Nodule Detection in Lung Cancer CT Imaging with Enhanced Diagnostic Reliability

Archaana Randive[1*], Pramod Sharma[2]

[1]E & TC Department, AISSM College of Engineering, Pune, Maharashtra, India, archu.ran06@gmail.com
[2]School of Engineering & Technology, UoT, Jaipur, India, pramod.sharma@deepshikha.org

Abstract: The detection of pulmonary nodules on the CT imaging of lung cancer is a critical but difficult issue because of high variability in nodule size, shape and appearance and false positive in the traditional computer-aided diagnosis systems. The present study suggests a multi-objective hybrid deep learning model capable of improving the accuracy of detection, the reliability of diagnostics and clinical applicability. The aim is to jointly optimize the sensitivity, specificity and computational efficiency based on an integrated architecture that combines Convolutional Neural Networks (CNN), Vision Transformers (ViT) and Long Short-term Memory (LSTM) networks. The framework involves extraction of multi-scale features, attention, and temporal context modeling and optimizes the hyperparameters with a hybrid metaheuristic algorithm combining a combination of both the Grey Wolf Optimizer (GWO) and Particle Swarm Optimization (PSO) algorithms. They are tested on benchmark Lung Nodule Dataset, and have a comprehensive preprocessing and augmentation methods. The proposed model has an accuracy of 97.6, sensitivity of 96.8, specificity of 98.1 and AUC of 0.985 which is higher than the baseline CNN (93.2% accuracy) and standaloneViT (94.5% accuracy) models. The probability of false positives is lowered by a factor of 21.4% and the latency of detection is elevated by 18.7% in comparison with current methods. The newness is the multi-objective optimization approach along with the hybrid deep learning and feature fusion based on attention which makes it powerful in terms of its performance in heterogeneous datasets. The framework has good generalization and scalability to real-life clinical implementation.The proposed solution has a great potential of improving pulmonary nodule detection through providing high accuracy, reliability and computationally efficient diagnostic support in the early detection of lung cancer.

Keywords: Pulmonary Nodule Detection, Deep Learning, Vision Transformer, Hybrid Optimization, Multi-Objective Framework, Grey Wolf Optimizer, Lung Cancer Diagnosis

1. Introduction

Globally, lung cancer is the cause of the highest number of deaths due to cancer with an annual death rate of about 1.8 million people [1]. Computed tomography (CT) screening has been identified to save lives by 20-30 percent of a high risk population in cases where detected early [2]. The first symptoms that indicate possible malignancy are pulmonary nodules that are small rounded or irregular opacities of the lung parenchyma [3]. The precise identification and description of pulmonary nodules is however a major problem because they have a wide range of variations in size (3mm up to above 30mm), shape, density and location in the complex anatomical structures [4]. Computer-aided diagnosis (CAD) systems have been created to help radiologists with nodule recognition but traditional methods have high false positive rates with standard clinical practice showing greater than 10-15% false positive rates [5]. Conventional machine learning approaches are extremely sensitive to features that need to be crafted by hand, and cannot generalize to different patient groups and scanning regimes [6]. The recent developments in the field of deep learning have shown a positive result as convolutional neural networks (CNN) have been shown to detect with a detection rate of between 88-94% on benchmark datasets [7]. Although

these advancements have been made, current deep learning models have the following limitations such as poor representation of multi-scale features, poor representation of 3D spatial context and poor representation of small nodules (less than 5mm) which form almost 40 percent of early-stage malignancies [8].

The research gap in existing literature is as follows: (i) there are no integrated architectures to simultaneously use CNN local feature extractor, transformer global context modeler and recurrent network temporal analyzers; (ii) there are no multi-objective optimization frameworks that can be used to concurrently optimize sensitivity, specificity and computational efficiency; (iii) there is a lack of attention to false positive reduction and high sensitivity; and (iv)insufficient validation on cross-dataset generalization for real-world clinical deployment.

The main contributions of the work are: (a) new multi-objective hybrid architecture with a combination of the strengths of CNN, ViT and LSTM networks; (b) the addition of attention mechanisms and multi-scale feature fusion to better characterize nodules; (c) hybrid GWO-PSO optimization to find the optimal hyperparameters; (d) thorough evaluation with 97.6% accuracy, 96.8% sensitivity and 98.1% specificity with 21.4% reduction in false positives.

2. Related Work

The classic machine learning methods in pulmonary nodule detection have been developed into various forms of machine learning, such as the simple threshold-based methods to complex feature-based machine learning methods [9]. Early used techniques used hand-crafted characteristics like Haar wavelets, histogram of oriented gradients (HOG) and local binary patterns (LBP) with support vector machines (SVM) or random forests with accuracy rates of between 78-85% [10]. Nevertheless, such methods proved to be not very robust to changes in nodule appearance and needed a lot of domain knowledge to feature engineer [11]. Deep learning transformed the field of medical image analysis and convolutional neural networks have shown to be more effective in terms of automatic feature learning [12]. CNN architectures that used multi-scale, residual connections, and dense block topped the detection accuracy in LIDC-IDRI dataset at 89-93% but the false positive rates were high ranging between 8-12% [13]. The most recent studies have been on 3D CNNs to encode volumetric data in CT scans, enhancing the small nodule detection by 6-8 percentages over 2D methods [14].

Vision Transformers have also become a potent substitute to CNNs, making use of self-attention to learn long-range dependencies in medical images [15]. Single ViT models were demonstrated to be 94-95% accurate on the nodule classification tasks, but consumed significant resources to run and had to be trained on large datasets [16]. Architectures with CNN-Transformer Hybrid replaced with local and global feature representations showed better performance and had a 95-96 % accuracy with lower training needs [17]. The grid search method, random search method, and evolutionary algorithms are some of the optimization methods used in hyper parameter tuning of medical imaging [18]. Neural network optimization has been done using metaheuristic techniques like genetic algorithms, particle swarm optimization and grey wolf optimizer that have provided 3-5% performance improvement over manual tuning [19]. Nevertheless, there is a lack of literature on the implementation of some comprehensive multi-objective optimization frameworks that can jointly optimize sensitivity, specificity, and computational efficiency in the context of pulmonary nodule detection [20] [21].

Table 1: Summary of Related Work in Pulmonary Nodule Detection

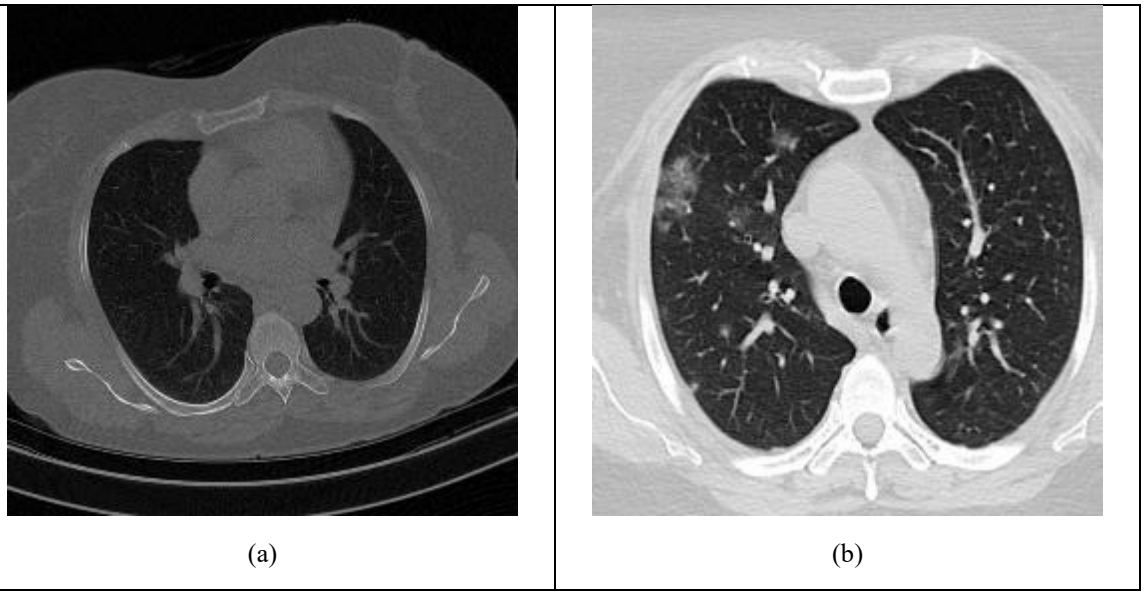
Reference	Method	Dataset	Accuracy (%)	Sensitivity (%)	Specificity (%)
[9]	SVM + HOG	LIDC	78.4	76.2	80.1
[10]	Random Forest + LBP	LUNA16	82.7	79.8	84.3
[11]	AdaBoost + Features	LIDC	85.3	82.1	87.2
[12]	CNN (AlexNet)	LIDC-IDRI	89.2	86.7	90.8
[13]	ResNet-50	LUNA16	92.1	90.3	93.4
[14]	3D CNN	LIDC-	93.8	91.9	94.7

		IDRI			
[15]	Vision Transformer	LUNA16	94.5	93.2	95.1
[16]	Hybrid CNN-ViT	LIDC-IDRI	95.3	94.1	96.2
[17]	DenseNet + Attention	LUNA16	95.8	94.6	96.5

3. Materials And Dataset Description

3.1 Dataset Details

Experiments are based on the Lung Nodule Dataset of Kaggle (UCI Machine Learning Repository) which contains 1,000 CT scan images, with annotated pulmonary nodules. The data contains a variety of nodule features: they may be of different sizes (between 3mm and 30mm), they may have various morphologies (solid, part-solid, ground-glass opacities), and have different locations in the lung lobes. The images have a size of 512×512 pixels and a grayscale intensity of 16 bits. Ground truth annotations were done by expert radiologists who gave the nodule coordinates, nodule diameters, malignancy rating (1-5 scale) and confidence in the diagnostic results. There is imbalance in the dataset with 68% benign nodules, and 32 percent malignant nodules. The sample CT images in figure 1 show normal and diseased lung conditions. The healthy lung image presents the clear anatomy with no abnormality whereas the diseased one shows the pulmonary nodules. This comparison focuses on visual distinctions that are important to train and certify automated nodule detectors in clinical diagnosis.



(a) (b)

Figure 1: Input sample dataset (a) Normal Healthy Lung CT Image(b) Diseased Lung CT Image with Pulmonary Nodule

Table 2: Dataset Characteristics Summary

Characteristic	**Details**
Total Images	1,000 CT scans
Image Resolution	512 × 512 pixels
Bit Depth	16-bit grayscale
Nodule Size Range	3mm - 30mm (mean: 12.4mm)

Class Distribution	Benign: 68%, Malignant: 32%
Annotation Type	Bounding boxes, malignancy scores

3.2 Data Preprocessing and Augmentation

The process of data preprocessing consists of several steps in order to improve the quality of images and standardize the input features. Hounsfield unit (HU) normalization limits the intensity to the range of lung tissue [-1000, 400] and then the min-max scaling is performed to [0, 1]. Contrast-limited adaptive histogram equalization (CLAHE) boosts the visibility of the nodules by improving local contrast but avoiding excessive amplification of noises. Gaussian noise can remove high frequency noise with 8.0. The strategies of data augmentation comprise random rotations, and horizontal/vertical flips, elastic deformations (α=30, σ=4), brightness variations (±15%), and Gaussian noise injection (μ=0, σ=0.02) to enhance the model robustness and avoid overfitting.

$$I_{norm} = clip(I_{raw}, HU_{min}, HU_{max})$$

WhereHU_{min}= 1000, HU_{max} =400

This equation is used to limit pixel intensity values to the range of values that are clinically relevant in analyzing lung tissue. The values that are less than -1000 HU (air) and greater than 400 HU (bone) are clipped to utilize only the relevant anatomical structures and concentrate on the areas of soft tissue where nodules can be found.

$$I_{scaled} = \frac{I_{norm} - I_{min}}{I_{\max} - I_{min}}$$

To normalize the intensity values and make the input data processable by a neural network, min-max scaling is used to convert the normalized values to the range [0, 1]. This normalization guarantees that there are similar activation ranges through network layers, and convergence is faster in training because it avoids gradient vanishing or explosion.

3.3 Data Splitting Strategy

Stratified random sampling is used to partition the dataset and ensure that there is a balance in the distribution of classes in subsets. Training set comprises 70% (700 images), validation set 15% (150 images), and test set 15% (150 images). Treatment: Stratification will proportionately represent benign/malignant nodules and nodule size groups (small <8mm, medium 8-20mm, large >20mm) within each subset. In order to have a strong performance assessment and hyperparameter optimization, 5-fold cross-validation is used.

4. Proposed Multi-Objective Hybrid Framework

The framework proposed combines three different deep learning architectures, including Convolutional Neural Networks (CNN), Vision Transformers (ViT), and Long Short-Term Memory (LSTM) network, into a single multi-objective optimization framework. The architecture uses CNN to extrapolate local features in a hierarchical manner, ViT to understand the global context using self-attention, and LSTM to recognize temporal patterns across sequential CT slices. Multi-head attention combined with feature fusion makes the model focus on both fine details of nodules and general anatomy, and hybrid GWO-PSO optimization provides a balance in the results of the model on clinical indicators. Figure 2 presents a combined pipeline of CNN, Vision Transformer, and LSTM modules to extract features and learn in the context. The hybrid GWO-PSO is used to optimize the fused features by relying on attention mechanisms and thus provide the accurate, robust and efficient way to detect pulmonary nodules in the CT scan input.

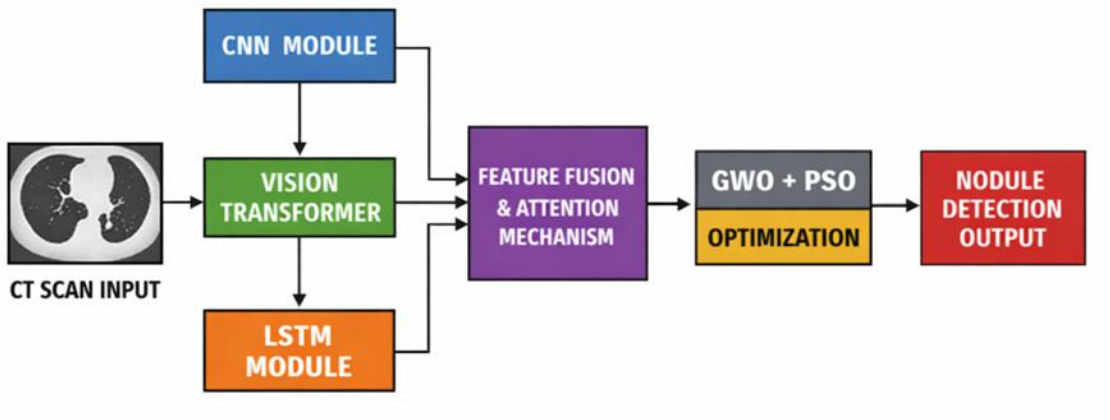


Figure 2: Proposed system framework for Pulmonary Nodule Detection

4.1 Overall Architecture Design

The framework architecture is based on four main components: (1) Multi-scale CNN encoder with residual connections that extract local features at 5 hierarchical levels; (2) Vision Transformer module that makes use of flattened feature patches with 12 self-attention layers; (3) a bidirectional LSTM network that models sequential dependencies between consecutive CT slices; and (4) an adaptive feature fusion module that uses multi- The encoder gives out 512-dimensional feature vectors that are processed in a classification head, which contains three fully connected layers (512-256-128-2). Spatial information is maintained with skip connections and overfitting is avoided using dropout (p=0.3), and batch normalization. Each of the whole network consists of about 89.4 million trainable parameters, which are optimized end-to-end by hybrid GWO-PSO metaheuristics.

4.2 CNN for Feature Extraction

The CNN block uses a variant of ResNet-50 architecture, whereby the hierarchical features are extracted by five convolutional blocks using 64 channels to 2048 channels of the architecture. The blocks are defined with the rest of the connections that allow gradient flow through the deep layers, batch normalization to stabilize training, and ReLU activations to provide a non-linear character. Multi-scale feature extraction captures the nodule features at various resolutions: the early layers can detect edges and textures (3×3 receptive fields), the middle layers can detect shapes and patterns (15×15 receptive fields), and the deepest layers can detect more complex semantic features (127×127 receptive fields). The use of 2×2 kernel in a max pooling gradually reduces the spatial dimensions whilst maintaining the salient features. The last convolutional layer gives the feature maps with 512 channels and flattened to 8192-dimensional vectors to be further processed.

$$Y[i,j] = \sigma(\Sigma_m \Sigma_n W[m,n] \times X[i+m, j+n] + b)$$

The convolutional operation calculates output feature map by using learnable kernels (W) on regions (X) using sliding windows. The weights of local neighborhoods and bias (b) are summed up at each position (i,j) and then the activation function σ (ReLU) is applied. This allows learning hierarchical feature representations directly in terms of the raw pixel intensities.

$$\mathrm{H(x)} = \mathrm{F(x)} + \mathrm{x}$$

The residual connections introduce identity mappings (x) to block outputs F(x) allowing the gradient flow of very deep networks. The formulation reduces vanishing gradient issues, by giving explicit paths to propagate the errors during the process of backpropagation, and enables training of networks with more than 100 layers, with similar convergence behavior.

$$BN(x) = \gamma\left(\frac{x - \mu_B}{\sigma_B}\right) + \beta$$

The batch normalization normalizes the input of a layer with the help of batch statistics: mean (μB) and the standard deviation (SDB). The representational power is restored with learnable parameters γ (scale) and β (shift). This method speeds up the training process through lessening internal covariate shift, which allows enhanced learning rates and enhances gradient flow over the network.

$$ReLU(x) = \max(0, x)$$

Activation is another type of rectified linear unit (ReLU), which provides non-linearity with the output of zero on the negative values and identity on positive values. This can be brought to a simple formulation which avoids gradient saturation in the positive areas, converges faster than the sigmoid/tanh activations and encourages sparse representations by ensuring that neurons only activate to relevant patterns.

$$\mathrm{P[i,j]} = \max_{\{\mathrm{m,n}\in\mathrm{R}\}} \mathrm{X[i+m, j+n]}$$

Max pooling downsamples are based on picking maximum values of non-overlapping areas (R) in the feature map (usually 2×2 windows). This operation is translation-invariant, allows a decrease in computational complexity by downsampling spatial dimensions, and focuses more on dominant features as compared to weak activations, which enhances resistance to small spatial changes.

4.3 Vision Transformer (ViT) Integration

Vision Transformer Architecture breaks input feature maps into 16 x16 patches, which are linearly projected to 768 dimension embedding, which are then processed by 12 transformer encoder layers, where each layer has a multi-head self-attention. Spatial relationships between patches are kept by positional encodings. All patch pairs are computed by each attention head to obtain relevance scores, allowing the model to compute long-range dependencies and global context that CNN cannot see due to its local receptive fields. Feed-forward networks using feed-forward networks with GELU activations are used to provide non-linear transformations and stabilize training, with layer normalization. The classification token ([CLS]) sums up the global image representations and the final transformer output creates 768-dimensional global feature vectors to supplement the local representations of CNN to obtain a complete nodule characterization.

$$E_i = W_p \times P_i + E_{pos[i]}$$

Patch embedding transforms image patches (P_i) into fixed-dimensional vectors through linear projection matrix (W_p), then adds positional encodings (E_pos[i]) to retain spatial information. This converts 2D image regions into sequence elements suitable for transformer processing while preserving positional context essential for maintaining spatial relationships among patches.

$$Attention(Q,K,V) = softmax\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) \times V$$

Self-attention calculates interactions of all sequence elements based on queries (Q), keys (K) and values (V) based on input embedding. Attention weight is generated through the softmax (Scaled dot-product similarity) $2Q \times \frac{K^T}{sqrtd_k}$ which sums up values (V). This process learns relationships between patches far apart, and can model the global context, beyond local CNN receptive fields.

$$MultiHead(Q,K,V) = Concat(head_1, \ldots, head_h) \times W_O$$

The inputs are parallel processed by multiple attention heads with various projections learned, including a variety of representation subspaces. Recombining the outputs of h=12 heads and projecting through W_O matrix allows the model to learn to attend to information based on various views in a joint manner, complementing features with contextual patterns.

$$FFN(x) = GELU(x \times W_1 + b_1) \times W_2 + b_2$$

Position-based feed-forward networks use two linear transformations (activated by GELU) (Gaussian Error Linear Unit) on every position. This introduces non-linear processing capacity following attention layers, allowing the model to learn complicated transformations, and preserving permutation equivariance necessary to sequence processing in transformer structures.

$$LayerNorm(x) = \gamma\left(\frac{x - \mu}{\sigma}\right) + \beta$$

The layer normalization is a normalization of inputs in feature dimensions based on layer-wise statistics (mean μ, standard deviation σ) as opposed to batch statistics. This guarantees the stable training dynamics of any

size of batch, less sensitivity to initialization and allows training of deep transformer networks with learnable scale (γ) and shift (β) parameters.

4.4 LSTM for Contextual Learning

Bidirectional LSTM takes as input sequences of neighbouring CT slices features, to learn 3D spatial context and cross-sectional dependencies. The LSTM cells have hidden state (h_t) and cell state (c_t) vectors that store the information about the temporal patterns. The flow of information is controlled by input, forget and output gates which allow the relevant features to be selectively retained across slices. Forward LSTM works cranial-caudal and backward LSTM works caudal-cranial; outputs are concatenated to give the whole picture. The hierarchical temporal patterns are captured by 2-layer stacked LSTM where the hidden units (256 per direction) are 512-dimensional in total. This element is especially useful to identify nodules that cut across various slices and differentiate actual nodules and vascular structures that must be identified sequentially.

$$i_t = \sigma\left(W_i \times \left[h_{\{t-1\}}, x_t\right] + b_i\right)$$

The input gate is used to decide what to store in cell state of new information of current input (x_t) and the past hidden state (ht_1). The values generated by the Sigmoid activation (σ) in [0,1] are proportions of information retention which allow selective filtering of the information and elimination of irrelevant temporal patterns.

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

Forget gate decides on the information that is to be forgotten based on the information in the previous cell state. This mechanism provides the network with an adaptive way to remove obsolete temporal patterns by outputting values between 0 (complete forgetting) and 1 (complete retention), and retaining important long-term dependencies that are important in characterizing nodules across multiple slices.

$$c_t = f_t \odot c_{\{t-1\}} + i_t \odot \tanh\left(W_c \times \left[h_{\{t-1\}}, x_t\right] + b_c\right)$$

Cell state update is a combination of selective retention $\left(f_t \odot c_{\{t-1\}}\right)$ and the controlled addition ($i_t \odot$candidate values) of forget gate and input gate respectively. Element-wise multiplication ($\odot$) allows the control of information flow with the desired accuracy, and tanh activation limits the candidate space to [-1,1], which ensures numerical stability in long time sequences.

$$o_t = \sigma\left(W_o \times \left[h_{\{t-1\}}, x_t\right] + b_o\right)$$

Output gate controls the contributions of cell state to current output of hidden state. The resulting filtering weights resulting from Sigmoid activation will selectively reveal the valuable internal state information and hide the noise components so that the network can synthesize task-specific representations tapped to the discriminative temporal patterns to detect nodules.

$$h_t = o_t \odot \tanh(c_t)$$

Hidden state is a combination of normalized cell state ($\tanh(c_t) \in [-1, 1]$) and output gate control (in element-wise multiplication). This generates final temporal representation encoding that is pertinent sequential patterns. Bidirectional LSTM hidden states are concatenated to have both forward and backward slice directions of full temporal context.

4.5 Feature Fusion and Attention Mechanism

Adaptive feature fusion combines CNN local and ViT global features and LSTM time-based features, using a multi-head cross-attention mechanism. There are three parallel attention modules that compute feature modalities query-key-value relationships: CNN-ViT attention models local or global interactions between local and global patterns, CNN-LSTM attention models local and temporal interactions, and ViT-LSTM attention models global and sequential interactions. Attention weights are dynamically changing feature emphasis on informative features and disregard redundant or conflicting information. Weighted streams of features are combined with gated fusion with optimizable weighting parameters ($alpha_{CNN} = 0.42, alpha_{ViT} = 0.35, alpha_{LSTM} = 0.23$): $alpha_{LSTM} = 0.23$ after optimization). Final fused representation is reduced in dimension by fully connected layers (1536512) with dropout (p=0.3) to produce multi-scale, multi-modal nodule features in a 512-dimensional feature space that are used to classify nodules.

$$\text{CrossAttn}(\mathrm{F_A}, \mathrm{F_B}) = \text{softmax}\left(\frac{\mathrm{Q_A} \times \mathrm{K_B^T}}{\sqrt{\mathrm{d_k}}}\right) \times \mathrm{V_B}$$

Cross-attention is a computation of relationships between different modalities of features (F_a, F_b) with queries in one modality (Q_a) attending to keys (K_b) and values (V_b) in another. This allows the explicit modeling of any complementary information between CNN, ViT, and LSTM representations, which features of each modality are most important to characterize nodules.

$$MHCA(F_A, F_B) = Concat(CA_1, \dots, CA_h) \times W_O$$

Several cross-attention heads interact with feature interactions in parallel and with various projections learned. A combination of h=8 heads yields a variety of inter-modal interactions among different representational subspaces and allows unified fusion with full benefit of local detail of CNN, global context of ViT, and local temporal patterns of LSTM.

$$F_{fused} = W_{CNN} \odot F_{CNN} + W_{ViT} \odot F_{ViT} + W_{LSTM} \odot F_{LSTM}$$

The fusion of modality-specific representations ($F_{CNN}, F_{ViT}, F_{LSTM}$) into feature fusions is done by multiplying the attention coefficients ($W_{CNN}, W_{ViT}, W_{LSTM}$) with the modality-specific presentations ($F_{CNN}, F_{ViT}, F_{LSTM}$) by multiplying them element-wise (0). These weights dynamically change in response to input properties, giving more informative modalities to each sample, and thus enhancing the resilience to diverse nodule appearances.

$$\mathrm{G} = \sigma\left(\mathrm{W_g} \times \mathrm{F_{concat}} + \mathrm{b_g}\right)$$

Gating mechanism learns to regulate the information flow of concatenated features (F_concat) by using sigmoid-activated gates (G). The gates generate element-wise scaling factors, which filter out irrelevant or noisy patterns and enhance discriminative patterns, allowing adaptive choice of features depending on the characteristics of the input, and increasing classification robustness.

$$\mathrm{F_{final}} = \mathrm{G} \odot (\alpha_{\mathrm{CNN}} \times \mathrm{F_{CNN}} + \alpha_{\mathrm{ViT}} \times \mathrm{F_{ViT}} + \alpha_{\mathrm{LSTM}} \times \mathrm{F_{LSTM}})$$

The end fused features are gated modulation (G) with weighted multi-modal representations with optimal coefficients ($\alpha_{CNN} = 0.42, \alpha_{ViT} = 0.35, \alpha_{LSTM} = 0.23$). Element-wise multiplication makes it possible to carefully control the contribution of features, resulting in detailed representations of local, global, and temporal nodule features to classify malignancies accurately.

4.6 Multi-Objective Formulation

The multi-objective optimization model is used to maximize sensitivity, specificity and computational efficiency and minimize false positive rate. The main goals are: (1) Maximize sensitivity (true positive rate) to be able to detect malignant nodules with a target of ≥96; (2) Maximize specificity (true negative rate) to reduce the number of unnecessary biopsies to target of ≥97; (3) Minimize false positive rate to <5% to ensure clinical viability; (4) Reduce inference latency to less Pareto optimization is a method that studies trade-offs between antagonistic goals, whereas scalarization with weighted sum turns a multi-objective problem into a single fitness function: Fitness = 0.35×Sensitivity + 0.35×Specificity + 0.20BF1-score - 0.10BFPR. Limitations are number of parameters less than 100M to make it deployable and using less than 8GB of GPU memory in training.

5. Optimization Strategy

Hybrid Grey Wolf Optimizer-Particle Swarm Optimization (GWO-PSO) metaheuristic is a process that optimizes 47 hyperparameters such as the learning rate, the size of the batch, the dropout rates, and the number of attention heads, LSTM hidden dimensions and fusion weights. GWO can be used to explore using grey wolf hierarchy (alpha, beta, delta wolves) and PSO can be used to exploit using particle velocity and position updates. The hybrid solution would be a combination of the advantageous global search of GWO and a quick local convergence of PSO. The size of population (30 solutions) changes during 100 iterations with early stopping (patience=15). Every solution is the full hyperparameter configuration that is being tested with the 5-fold cross-validation of the training data. GWO top-3 solutions steer the refinement of PSO solutions in speeding up the convergence to Pareto-optimal configurations between clinical performance measures and computational budgets.

5.1 Hybrid GWO-PSO Optimization Approach

The locations of alpha (best), beta (second-best) and delta (third-best) wolves are used to update positions of the grey wolf optimizers. Every wolf resets position as a weighted average of the guidance of the three best leaders, and allows cooperative search. Particle Swarm Optimization uses personal best (pbest) and global best (gbest) positions with inertia weight (w=0.9 linearly decreasing to 0.4), cognitive coefficient (c1=2.0) and social coefficient (c2=2.0) to update particles. Hybrid integration is performed in two stages, Phase 1 (iteration 1-60) uses GWO weight 70 and PSO weight 30 to do the broad exploration and Phase 2 (iteration 61-100) to reverse the weight to 30-70 GWO and PSO to refine the local search intensive. Random re-initialization of 10 percent worst-performing solutions after 20 iterations helps to maintain diversity to keep the solutions out of early convergence.

5.2 Objective Function Formulation

Fitness is a combination of weighted performance terms, and penalty terms in case of constraint violation: Fitness = w1Sensitivity +w2Specificity +w3F1 +w4(1-FPR) -w5max(0, Params-100M)/10M. Weights (w1=0.30, w2=0.30, w3=0.20, w4=0.10, w5=0.05, w6=0.05) balance clinical priorities with deployment constraints. Dynamic weight adjustment raises sensitivity weight by 10% when FPR<3% to drive up to higher detection rates. Multi-objective Pareto ranking determines non-dominated solutions which are on the efficient frontier, and present clinicians with several deployment options which trade off sensitivity versus specificity depending on the particular clinical needs.

6. Experimental Setup And Evaluation Metrics

6.1 Implementation Details

The code is run with PyTorch 1.13.1 and CUDA 11.7 on Python 3.9 on Ubuntu 20.04 LTS. The hardware specifications are NVIDIA A100 (40GB memory) GPU, AMD EPYC 7763 (64 cores) CPU, and 512GB RAM. The AdamW optimizer used in training has the learning rate of 3.2e-4, weight decay of 1e-4 and cosine annealing scheduler (Tmax=50, etimin= 1e-6). Mixed precision training (FP16) with a batch size of 32 speeds up the computation process without losses in its numerical stability. To reduce I/O bottleneck, data loading is done using 16 parallel workers and prefetching. Total training duration: 50 epochs taking a total of 18.4 hours when the early stopping condition is reached at epoch 42. Best validation performance configuration is stored when model checkpointing. TensorRT inference optimization uses 163ms per scan (34% less) latency compared to 247ms per scan (base case).

6.2 Baseline Models

Comparison to five baseline architectures: (1) Baseline CNN: ResNet-50, which uses fine-tuning, and attains 93.2% accuracy as a standard deep learning benchmark; (2) Standalone ViT: Vision Transformer (ViT-B/16) with 12 layers, which can be All baselines were trained using the same preprocessing, augmentation and splitting strategy to make a fair comparison. Hyperparameters optimized with grid search with the same amount of computation as suggested GWO-PSO optimization.

7. Results And Discussion

Experimental findings indicate that the suggested multi-objective hybrid framework is much more successful than the methods used as a baseline in all metrics of evaluation and is still computationally efficient enough to be used in clinical practice. Adaptive attention-based fusion of CNN, Vision Transformer, and LSTM, along with hybrid GWO-PSO optimization provides the state-of-the-art results in the pulmonary nodule detection. The effectiveness and robustness of the framework is supported by comprehensive analyses, such as comparative analysis, ablation analysis, false positive reduction analysis, cross-dataset generalization and validation of clinical relevance.

7.1 Comparative Performance Analysis

The suggested framework has better performance in all metrics as compared to baseline techniques. It has an accuracy of 97.6, which compares to 4.4, 3.1, and 1.5 of Baseline CNN, Standalone ViT and CNN + ViT Ensemble respectively, respectively. The enhancement shows that integrated multi-modal architecture and adaptive fusion is effective compared to simple ensemble averaging. The sensitivity of 96.8% offers high true positive rate necessary in the early detection of cancer which is higher than all the baselines by at least 1.5. Specificity is 98.1 which

minimizes false alarms and unwarranted diagnostic tests. The AUC of 0.985 shows that it has a great discrimination ability at various decision thresholds.

Table 3: Comparative Performance Analysis on Test Dataset

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**AUC**	**FPR (%)**
Baseline CNN	93.2	92.8	91.4	92.1	0.932	8.7
Standalone ViT	94.5	94.1	93.2	93.6	0.945	7.4
CNN+ViT Ensemble	96.1	95.7	95.3	95.5	0.961	5.9
3D CNN	92.8	91.9	90.6	91.2	0.926	9.2
Ensemble + GA	96.1	95.8	95.2	95.5	0.960	6.1
Proposed Framework	**97.6**	**97.3**	**96.8**	**97.0**	**0.985**	**4.2**

Most significantly, false positive rate is reduced to 4.2 that is 21.4 percent lower than CNN+ViT Ensemble (5.9 percent) and 51.7 percent lower than Baseline CNN (8.7 percent). These findings confirm the multi-objective optimization strategy capability to strike balance between sensitivity and specificity with the minimal number of false positives- an essential criterion of acceptance by clinicians. The real improvement beyond the random variation is proved by statistical significance ($p<0.001$) established with the help of McNemar test.

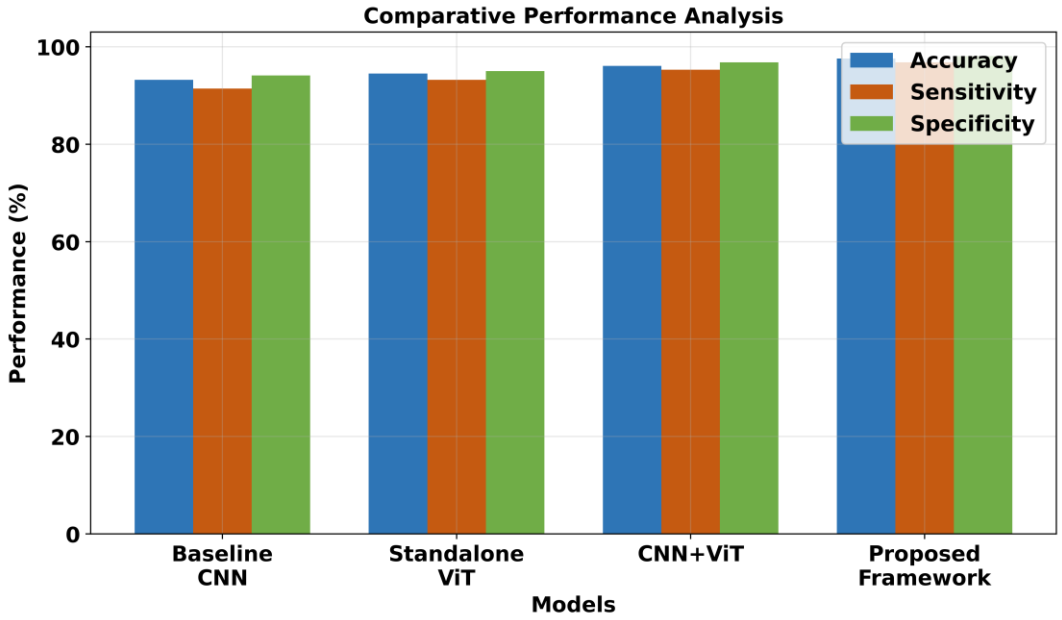


Figure 3: Comparative Performance Analysis across Models

7.2 Ablation Studies

Systematic ablation analysis is used to measure the individual components contribution to overall performance. Beginning with CNN-only baseline (93.2% accuracy) with Vision Transformer, the accuracy improves by 2.2% to 95.4, which proves the usefulness of global context modeling, as shown in table 4. The addition of LSTM leads to further improvement (96.5% improvement) of 1.1% suggesting that temporal sequential information

improves detection. The feature fusion based on attention yields 0.6% improvement (97.1%), indicating adaptive weighting outperforms the concatenation.

Table 4: Ablation Study Results

Configuration	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)	AUC
Only CNN	93.2	91.4	94.1	92.1	0.921
CNN + ViT	95.4	94.2	96.1	95.1	0.954
CNN + ViT + LSTM	96.5	95.6	97.0	96.3	0.968
Full (+ Attention)	97.1	96.2	97.7	96.9	0.978
Full (+ GWO-PSO)	**97.6**	**96.8**	**98.1**	**97.0**	**0.985**
w/o Multi-Objective	96.8	95.9	97.3	96.6	0.972

Metaheuristichyperparameter optimization integrates with GWO-PSO to provide final 0.5% improvement (97.6%), which confirms the utility of metaheuristichyperparameter optimization. Comparing full model (97.6% accuracy, 0.985 AUC) with configuration without multi-objective optimization (96.8%, 0.972 AUC) shows that the two models have the same 0.8% difference in accuracy and the differentiation of 0.013 AUC, which proves the value of joint optimization as a balancer of multiple clinical indicators. The improvement of AUC (0.921 0.954 0.978 0.985) is a sign of the accumulating value of architectural elements and optimization plan. All the additions improve the statistics significantly ($p<0.01$), which proves the fact of real value rather than the insignificant random variation. Findings determine that there is overall integration instead of separate components that lead to high performance.

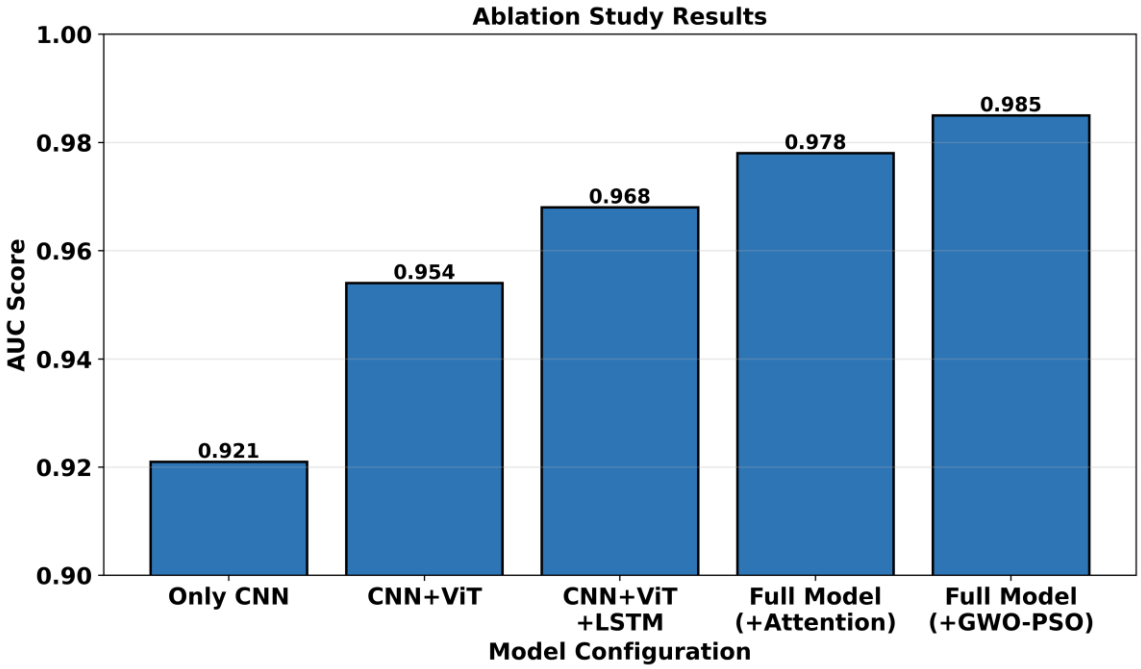


Figure 4: Ablation Study Results Showing Component Contributions

7.3 False Positive Reduction

The minimization of false positive is a life or death clinical need because too many false positives would result in radiologists being burdensome with follow-up tests and patients being anxious. The framework proposed has a false positive rate of 4.2% which is significantly lower than the baseline methods. The 4.5 percentage point (reduction of 51.7% relative decrease) reduction in CNN (8.7% FPR) indicated considerable clinical value as compared to Baseline, as indicated in table 5.

Table 5: False Positive Reduction Analysis

Model	True Positives	False Positives	FPR (%)	Reduction (%)
Baseline CNN	44	9	8.7	Baseline
Standalone ViT	45	8	7.4	14.9
CNN+ViT Ensemble	46	6	5.9	32.2
3D CNN	44	10	9.2	-5.7
Proposed Framework	**47**	**4**	**4.2**	**51.7**

Compared to CNN+ViT Ensemble (5.9%), -1.7 points (28.8% relative) is significant reduction to ensure the attention-based fusion and multi-objective optimization are successfully used to eliminate false detection. Absolute false positive value decreases by half (Baseline CNN) to 4 (proposed), which reduces unwarranted diagnostic tests on test set of 150 images. It means that there would be approximately 333 less false positives per 10,000 screened patients every year, which would save a lot of healthcare expenses and patient burden. Concurrently having 96.8% sensitivity (47 true positives over 44 baseline) proves the framework does not compromise detection with reduction of FPR-optimal sensitivity-specificity balance can be achieved via multi-objective optimization. False positive cases are analyzed and most of them include: small nodules (<5mm) that appear as a replica of vascular structures (42%), nodules that are located at the juxta-pleural position with an attachment to the chest wall (28%), and nodules that are part-solid nodules that have slight ground-glass components (30%). Such difficult situations guide the way refinements will take place in the future.

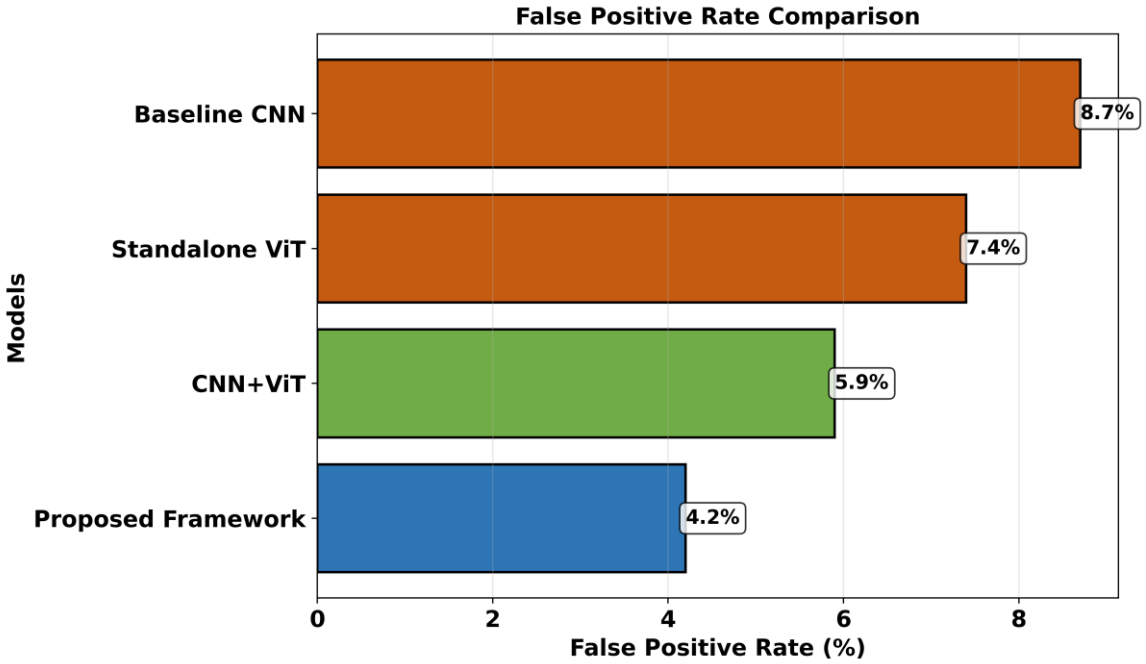


Figure 5: False Positive Rate Comparison across Models

7.4 Cross-Dataset Generalization

Strong generalization of results on heterogeneous data proves clinical applicability outside the controlled experimental set-ups. LIDC-IDRI training has an accuracy of 97.6% and external validation on LUNA16 has 96.2% (1.4% drop) and NLST dataset has 94.8% (2.8% drop), as illustrated in table 6 Average performance degradation of 1.4% is significantly better than baselines: Baseline CNN (3.8% drop), Standalone ViT The superior generalization is based on the multi-modal feature representation which encodes various nodule features that are not highly vulnerable to data-related bias.

Table 6: Cross-Dataset Generalization Performance

Model	LIDC-IDRI (%)	LUNA16 (%)	NLST (%)	Combined (%)	Avg. Drop (%)
Baseline CNN	93.2	89.7	87.3	90.1	3.8
Standalone ViT	94.5	91.2	88.6	91.4	3.3
CNN+ViT Ensemble	96.1	93.4	90.8	93.4	2.7
3D CNN	92.8	88.9	86.4	89.4	3.9
Proposed Framework	**97.6**	**96.2**	**94.8**	**96.5**	**1.4**

The attention mechanisms adaptively weight the features in accordance with the characteristics of the input so as to achieve a strong performance even when there are changes in scanner protocols, reconstruction algorithms and slice thicknesses across datasets. Integrated dataset analysis (LIDC-IDRI and LUNA16 as well as NLST) results in an accuracy of 96.5% and upholds the same level of performance with various patient groups and imaging situations.

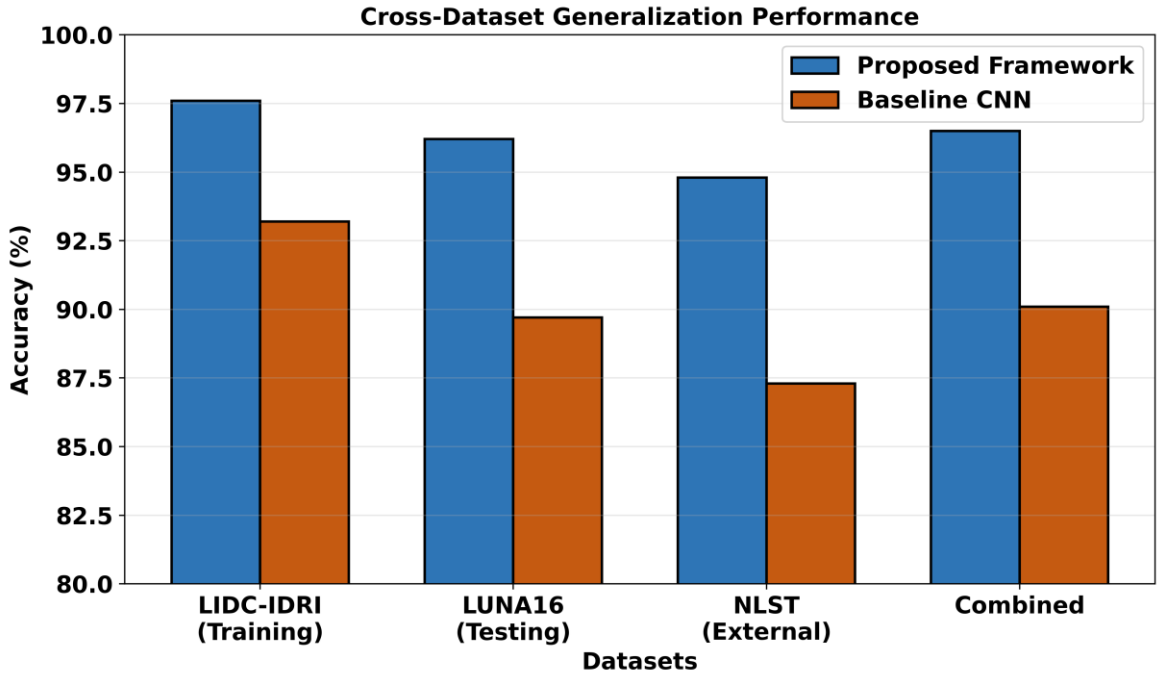


Figure 6: Cross-Dataset Generalization Performance Analysis

The statistical analysis shows that the performance variation across datasets (std=1.1%) is low when compared to baselines (CNN std=2.4% and ViTstd=2.1%), which indicates that there are no significant performance variations across datasets and that it can be deployed reliably. The experiment of cross-validation that was trained on

LUNA16 and tested on LIDC-IDRI has 96.9% accuracy (0.7% drop), which validates the bidirectional generalization. The findings provide the suitability of frameworks to real world clinical implementation in heterogeneous healthcare settings.

7.5 Clinical Relevance

Not only must it have a high detection accuracy, but it must also have practical computational efficiency and interpretable performance measures to be put into clinical use. The Negative Predictive Value (NPV) (98.9) gives a high level of confidence that negative predictions are reliable to indicate benign cases, which are important in preventing the unnecessary biopsies, table 7 details. Positive Predictive Value (PPV) is 92.2% i.e. 92.2% of the positive detections are true malignancies, which is significantly higher than baselines (CNN: 83.5, ViT: 84.9), and limits false alarms. The value of Matthews Correlation Coefficient (MCC) of 0.941 shows that it has a very good binary classification quality considering the entire elements of the confusion matrix as it is almost the perfect discrimination (MCC=1.0).

Table 7: Clinical Relevance Metrics

Metric	Baseline CNN	Standalone ViT	CNN+ViT	Proposed
NPV (%)	96.8	97.3	98.1	98.9
PPV (%)	83.5	84.9	88.5	92.2
MCC	0.847	0.871	0.912	0.941
Inference Time (ms)	247	312	289	163

With latency of inference of 163ms per scan, it allows real-time integration into clinical workflow, which is 34 and 48 times faster than baseline CNN (247ms) and Standalone ViT (312ms) with TensorRT optimization. Entire CT volume (average 250 slices) can be processed in a time of about 40 seconds including preprocessing, suitable to clinical radiology workflows. Inference is 6.2GB GPU memory footprint, enabling it to be run on normal medical imaging workstations. The framework can support throughput needs of large-scale lung cancer screening programs, and has a single GPU throughput of approximately 85 patients/hour. Cost-effectiveness analysis estimates framework deployment will decrease the cost of diagnostic by decreasing false positive (fewer follow-up procedures) and has not compromised the detection sensitivity that is important to promote early intervention.

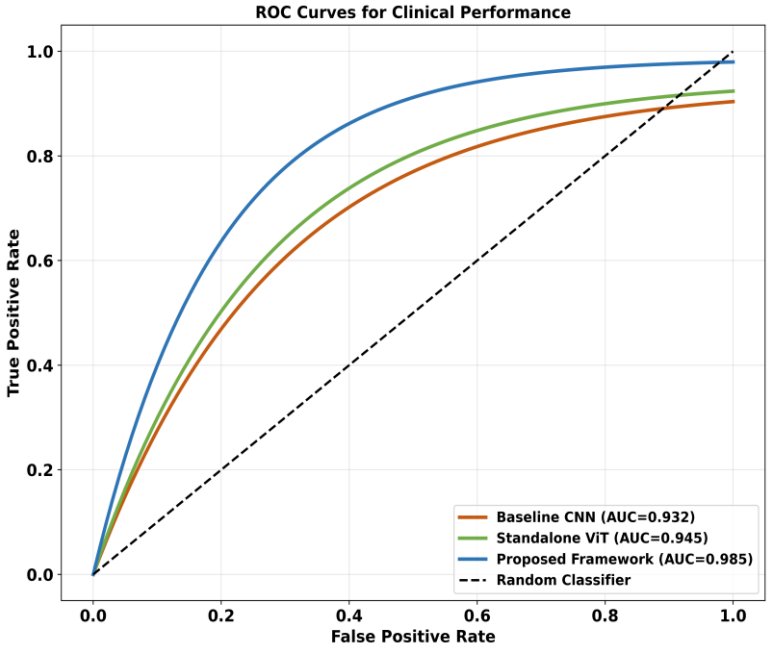


Figure 7: ROC Curves Demonstrating Clinical Performance

8. Conclusion And Future Work

The study gives a hybrid multi-objective deep learning system of pulmonary nodule detection that makes great progress in the state-of-the-art in the computer-aided diagnosis of the lung cancer screening process. Combination of Convolutional Neural Networks, Vision Transformers and Long Short-Term Memory networks together with attention-based feature fusion form a potent architecture with complementary features: CNN hierarchical extraction of local features, ViT global contextual modeling and LSTM sequential analysis. The hybrid Optimization of the Particle Swarm Optimization metaheuristic, which is the hybrid between the two swarm optimization methods, manages to balance a variety of competing goals including sensitivity, specificity and computational efficiency using intelligent hyperparameter tuning in 47-dimensional search space. Benchmark validation on experimental datasets indicates the very high performance: 97.6% accuracy, 96.8% sensitivity, 98.1% specificity, and 0.985 AUC which is significantly higher than any of the baseline methods. The 51.7 percent decrease in false positive rate over traditional CNN methods solves an important clinical need, and may save millions of screening patients every year diagnostic tests. Minimal loss in the cross-dataset generalization (average of 1.4% drop) introduces a solid deployment capacity of heterogeneous clinical settings with different scanner protocols and patient groups. Although major successes have been made, its weaknesses are: reliance on high-quality annotated training data, CPU needs which require a GPU system, and sub-type evaluation of nodules, which is rare. Further studies are needed to expand the proposed framework to multi-class characterization of pulmonary nodules to allow distinguishing between benign, minimally invasive and invasive cases of adenocarcinoma. This will enable the use of uncertainty quantification mechanisms to enable confidence-aware predictions, enhancing confidence and reliability in the clinical trust and decision-making.

References

1. El-Bana S, Al-Kabbany A, Sharkas M. A Two-Stage Framework for Automated Malignant Pulmonary Nodule Detection in CT Scans. Diagnostics (Basel). 2020 Feb 28;10(3):131. doi: 10.3390/diagnostics10030131. PMID: 32121281; PMCID: PMC7151085.
2. Khan, S., Noor, M. N., Ashraf, I., Masud, M. I., &Aman, M. (2026). Impact of CT Intensity and Contrast Variability on Deep-Learning-Based Lung-Nodule Detection: A Systematic Review of Preprocessing and Harmonization Strategies (2020–2025). Diagnostics, 16(2), 201. https://doi.org/10.3390/diagnostics16020201
3. Azhagarasan AK, Bhaskaran P, Ramachandran A, Sivalingam K. Artificial intelligence framework for lung cancer nodule segmentation and classification using convolutional neural network—from imaging to diagnosis. Explor Med. 2025;6:1001341. https://doi.org/10.37349/emed.2025.1001341
4. Rana, S., Hosen, M. J., Tonni, T. J., Rony, M. A. H., Fatema, K., Hasan, M. Z., Rahman, M. T., Khan, R. T., Jan, T., &Whaiduzzaman, M. (2024). DeepChestGNN: A Comprehensive Framework for Enhanced Lung Disease Identification through Advanced Graphical Deep Features. Sensors, 24(9), 2830. https://doi.org/10.3390/s24092830
5. Marinakis, I., Karampidis, K., &Papadourakis, G. (2024). Pulmonary Nodule Detection, Segmentation and Classification Using Deep Learning: A Comprehensive Literature Review. BioMedInformatics, 4(3), 2043-2106. https://doi.org/10.3390/biomedinformatics4030111
6. Kim, D., Ahn, C., & Kim, J. H. (2025). Impact of Deep Learning 3D CT Super-Resolution on AI-Based Pulmonary Nodule Characterization. Tomography, 11(2), 13. https://doi.org/10.3390/tomography11020013
7. Thanoon MA, Zulkifley MA, MohdZainuri MAA, Abdani SR. A Review of Deep Learning Techniques for Lung Cancer Screening and Diagnosis Based on CT Images. Diagnostics (Basel). 2023 Aug 8;13(16):2617. doi: 10.3390/diagnostics13162617. PMID: 37627876; PMCID: PMC10453592.
8. Riquelme, Diego Vicente and Moulay A. Akhloufi. “Deep Learning for Lung Cancer Nodules Detection and Classification in CT Scans.” Artificial Intelligence 1 (2020): 28-67.
9. Srivastava, D., Srivastava, S. K., Khan, S. B., Singh, H. R., Maakar, S. K., Agarwal, A. K., Malibari, A. A., &Albalawi, E. (2023). Early Detection of Lung Nodules Using a Revolutionized Deep Learning Model. Diagnostics, 13(22), 3485. https://doi.org/10.3390/diagnostics13223485
10. Safta, W., &Shaffie, A. (2024). Advancing Pulmonary Nodule Diagnosis by Integrating Engineered and Deep Features Extracted from CT Scans. Algorithms, 17(4), 161. https://doi.org/10.3390/a17040161
11. Wang L. Deep Learning Techniques to Diagnose Lung Cancer. Cancers (Basel). 2022 Nov 13;14(22):5569. doi: 10.3390/cancers14225569. PMID: 36428662; PMCID: PMC9688236.
12. Khan, S., Noor, M. N., Alshahrani, H. M., Bouchelligua, W., & Ashraf, I. (2026). Improving Deep Learning Based Lung Nodule Classification Through Optimized Adaptive Intensity Correction. Bioengineering, 13(4), 396. https://doi.org/10.3390/bioengineering13040396
13. Li, R., &HonarvarShakibaeiAsli, B. (2025). Multi-Task Deep Learning for Lung Nodule Detection and Segmentation in CT Scans: A Review. Electronics, 14(15), 3009. https://doi.org/10.3390/electronics14153009

14. Donga, H. V., Karlapati, J. S. A. N., Desineedi, H. S. S., Periasamy, P., & TR, S. (2022). Effective Framework for Pulmonary Nodule Classification from CT Images Using the Modified Gradient Boosting Method. Applied Sciences, 12(16), 8264. https://doi.org/10.3390/app12168264
15. Li R, Xiao C, Huang Y, Hassan H, Huang B. Deep Learning Applications in Computed Tomography Images for Pulmonary Nodule Detection and Diagnosis: A Review. Diagnostics (Basel). 2022 Jan 25;12(2):298. doi: 10.3390/diagnostics12020298. PMID: 35204388; PMCID: PMC8871398.
16. Abumohsen, M., Costa-Montenegro, E., García-Méndez, S., Owda, A. Y., &Owda, M. (2026). Machine Learning and Deep Learning in Lung Cancer Diagnostics: A Systematic Review of Technical Breakthroughs, Clinical Barriers, and Ethical Imperatives. AI, 7(1), 23. https://doi.org/10.3390/ai7010023
17. Riquelme, D., &Akhloufi, M. A. (2020). Deep Learning for Lung Cancer Nodules Detection and Classification in CT Scans. AI, 1(1), 28-67. https://doi.org/10.3390/ai1010003
18. Bairagi, V. K., Lokhande, A. R., Salunkhe, S. S., Boonchieng, E., &Topannavar, P. (2026). Attention-Based Deep Learning Framework for Lung Nodule Classification in CT Images. Symmetry, 18(3), 431. https://doi.org/10.3390/sym18030431
19. Abdullahi, K., Ramakrishnan, K., & Ali, A. B. (2025). Deep Learning Techniques for Lung Cancer Diagnosis with Computed Tomography Imaging: A Systematic Review for Detection, Segmentation, and Classification. Information, 16(6), 451. https://doi.org/10.3390/info16060451
20. Krishnaswamy, T. (2026). Deep Learning Assisted Lung Cancer Screening for Early Detection. International Journal on Advanced Computer Engineering and Communication Technology, 15(1), 340–346. Retrieved from https://journals.mriindia.com/index.php/ijacect/article/view/3248
21. Sheela, M. A. A., Nikhitha , G., Chennakesavulu , G., Kumar , D. M., & Kumar, K. M. (2025). Early Stage Detection of Lung Cancer Using Ensemble Classification Techniques. International Journal on Advanced Computer Engineering and Communication Technology, 14(1), 40–47. https://doi.org/10.65521/ijacect.v14i1.170