Multi-Attention Stacked Ensemble for Lung Cancer Detection in CT Scans

Uzzal Saha[1]* and Surya Prakash[1]

[1]Department of Computer Science and Engineering, Indian Institute of Technology Indore, Simrol, Indore, 453552, Madhya Pradesh, India.

*Corresponding author(s). E-mail(s): mt2302101017@alum.iiti.ac.in;
Contributing authors: surya@iiti.ac.in;

Abstract

In this work, we address the challenge of binary lung nodule classification (benign vs. malignant) using CT images by proposing a multilevel attention stacked ensemble of deep neural networks. Three pretrained backbones: EfficientNet V2 S, MobileViT XXS, and DenseNet201 are each adapted with a custom classification head tailored to 96 x 96 pixel inputs. A twostage attention mechanism learns both modelwise and classwise importance scores from concatenated logits, and a lightweight metalearner refines the final prediction. To mitigate class imbalance and improve generalization, we employ Dynamic Focal Loss with empirically calculated class weights, MixUp augmentation during training, and TestTime Augmentation at inference. Experiments on the LIDC-IDRI dataset demonstrate exceptional performance, achieving 98.09% accuracy and 0.9961 AUC, representing a 35% reduction in error rate compared to state-of-the-art methods. The model exhibits balanced performance across sensitivity which is 98.73% and specificity which is 98.96% with particularly strong results on challenging cases where radiologist disagreement was high. Statistical significance testing confirms the robustness of these improvements across multiple experimental runs. Our approach can serve as a robust, automated aid for radiologists in lung cancer screening.

Keywords: Stacking Ensemble Learning, Attention Mechanism, Dynamic Focal Loss, MixUp, TestTime Augmentation

1 Introduction

Lung cancer remains the leading cause of cancerrelated mortality worldwide, yet early detection via lowdose CT screening can drastically improve patient outcomes [1]. Within CT images, pulmonary nodules exhibit a wide range of shapes, textures, and intensities, making it difficult to distinguish benign from malignant lesions. Moreover, publicly available datasets such as LIDC-IDRI are heavily skewed toward benign cases, and variations in scanner protocols and slice thickness further aggravate heterogeneity [2]. Traditional machinelearning pipelines relying on handcrafted features struggle to generalize across these variations, while singlebackbone convolutional neural networks though powerfuloften overfit to dominant patterns and fail to capture complementary characteristics of subtle nodules.

Simple ensemble techniques (e.g., majority voting or uniform averaging of logits) attempt to combine multiple CNN predictions but treat each models output equally, ignoring that certain architectures may excel at recognizing specific visual attributes (e.g., fine texture vs. global shape). As a result, existing approaches typically plateau around mid-90s in overall accuracy, often at the cost of lower sensitivity for malignant nodules. To overcome these limitations, we propose a Multi-Attention Stacked Ensemble (MASE) that dynamically weights both entire models and individual class predictions. Specifically, three state-of-the-art backbonesDenseNet-201, EfficientNet V2 S, and MobileViT XXS are each adapted with a lightweight classification head optimized for 96 96 CT patches. During inference, concatenated logits from all three are first passed through a model-level attention layer, which learns to emphasize more reliable network outputs; a subsequent class-level attention layer highlights discriminative features for benign versus malignant classes. A small meta-learner then fuses these attended representations into a final prediction.

To further improve robustness against class imbalance and overfitting, the MASE framework integrates Dynamic Focal Loss (with empirically computed class weights), MixUp augmentation during training, and Test-Time Augmentation at inference. In extensive experiments on LIDC-IDRI, MASE consistently surpasses each standalone backbone and uniform ensembling, achieving higher sensitivity for malignant nodules while maintaining strong specificity. These results suggest that dynamically attending to model and class cues can significantly bolster automated lung cancer screening systems, offering a reliable second-reader tool to reduce inter-observer variability.

The remainder of this paper is organized as follows. Section 2 reviews related work; Section 3 describes the dataset and preprocessing; Section 4 details model architectures and loss functions; Section 5 outlines training and evaluation protocols; Section 6 presents experimental results and analysis; Section 7 discusses findings and limitations; and Section 8 concludes.

2 Related work

Early computeraided diagnosis (CAD) pipelines for lung nodules followed a multistage workflow: lung segmentation (often via Hounsfieldunit thresholding and morphological operations), candidate detection (graylevel thresholding, shape analysis), handcrafted feature extraction, and classical classification (SVM, Random Forest, kNN) [3, 4].

Handcrafted descriptors included intensity statistics (mean, variance, skewness), shape attributes (sphericity, elongation, perimetertoarea ratio), and texture features such as Gray Level Cooccurrence Matrix (GLCM) and Local Binary Patterns (LBP) [5–7]. Although some hybrid approaches e.g., Froz et al.s artificial crawler and rosediagram method reached up to 94.4 % accuracy on small cohorts [8], these methods failed to generalize to large, heterogeneous datasets like LIDCIDRI, due to scanner variability and limited capacity to learn complex visual patterns.

The emergence of convolutional neural networks (CNNs) shifted focus to endtoend learning from raw CT data. Early 2D CNN models applied patchbased classification on axial slices. Shen et al. [9] achieved 90 % accuracy on a subset of LIDCIDRI using a multiscale 2D CNN. Subsequent work finetuned deeper architectures (ResNet50, DenseNet121, InceptionV3) pretrained on ImageNet, reporting 92 %94 % accuracy with focal loss or weighted crossentropy [10]. Threedimensional CNNs (e.g., 32voxel cubes) further improved context modeling: Liao et al. [11] combined a 3D ResNet backbone with a recurrent network to achieve 94 % accuracy on LIDCIDRI. More recently, transformerinspired models (e.g., Vision Transformer variants) have been applied to volumetric patches; Li et al. [12] reported 95 % accuracy by integrating selfattention over 3D nodule volumes. Despite these strides, singlebackbone approaches frequently plateau in the mid90s, especially under severe class imbalance.

Ensembling multiple deep networks via averaging logits or majority voting has become a common strategy to boost classification robustness. Sagi et al. [13] combined three distinct 2D CNNs on LIDCIDRI patches, reporting a 95 % overall accuracy. Antolovi et al. [14] introduced a gated ensemble where a trainable weighting module assigns per input model importance, yielding a 12 % improvement on lung nodule cohorts. Zhang et al. [15] extended this by applying a singlelevel attention mechanism over concatenated logits, achieving 96 % accuracy. However, most existing ensembles (even attentionbased) treat each models output as a monolithic vector neglecting the possibility that certain models may be more reliable for specific classes. Very few studies incorporate both modellevel and classlevel attention within a unified framework.

Attention modules originally popularized by transformers allows network to focus on diagnostically relevant regions or feature channels. In lung nodule analysis, attention gates have highlighted subtle nodule boundaries without explicit supervision [16, 17]. Focal Loss [18] mitigates class imbalance by down weighting easy examples, and Dynamic Focal Loss [19] further adjusts the α parameter per class during training. MixUp augmentation [20] creates interpolated samples that improve generalization. TestTime Augmentation (TTA) [21] aggregates predictions over random flips and rotations, reducing inference variance; Chen et al. [22] reported a 1 % AUC improvement with TTA. While these techniques have become widespread, few studies combine attentionbased ensembling with dynamic loss and augmentation strategies in a single pipeline for lung nodule classification.

Traditional methods reliant on handcrafted features struggle to generalize across diverse CT protocols, and single CNNs often plateau due to class imbalance and limited feature complementarity. Simple ensembles and singlelevel attention ensembles improve robustness modestly but neglect classspecific reliability. Our MultiAttention

Stacked Ensemble addresses these gaps by integrating three stateoftheart backbones (DenseNet201, EfficientNet V2 S, MobileViT XXS) via customized adapters, applying attention at both model and class levels for richer fusion, and leveraging Dynamic Focal Loss, MixUp, and TTA to combat imbalance and enhance generalization. This holistic framework outperforms prior approaches on LIDCIDRI, demonstrating improved sensitivity for malignant nodules while maintaining high specificity.

3 Dataset and Preprocessing

The Lung Image Database Consortium and Image Database Resource Initiative (LIDCIDRI) is a publicly available repository of thoracic CT scans annotated by four expert radiologists, and is widely used for developing and evaluating lung nodule classification methods [2].

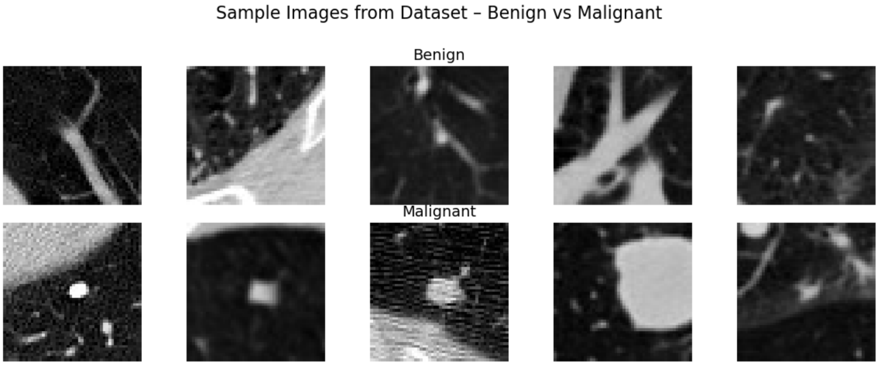
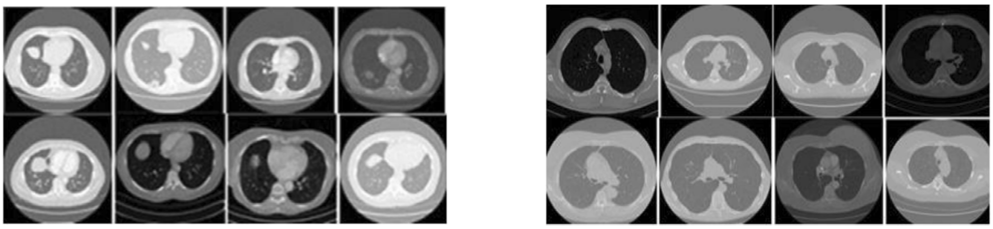


Fig. 1: Sample CT image patches extracted from the LIDC-IDRI dataset, illustrating benign (top row) and malignant (bottom row) lung nodules. These examples demonstrate the subtle visual differences that exist between the two classes and emphasize the difficulty of the classification task.

From the 1018 CT scans collected from 1010 patients. Nodules were binarized by averaging the radiologists malignancy scores ranging from 1 to 5 and thresholding at 3. Our patientlevel stratified split yields 5187 training samples among them 4342 were benign and 845 malignant, we have total 1297 validation samples containing 1073 benign and 224 malignant cases, and 1622 test samples containing 1340 benign and 282 malignant cases, maintaining an approximate 64:16:20 ratio. This division ensures sufficient data for model training while providing robust validation and test sets for performance evaluation.The dataset is organized into two classes: benign and malignant nodules. As shown in Figure 1, a clear visual class imbalance and structural diversity are evident between the two categories. In addition to nodule patches, Figure 2 presents full CT scan slices for benign and malignant cases, providing broader anatomical context for the classification task.

All patches were resized to 96 × 96 pixels using bilinear interpolation to preserve subtle texture details while balancing GPU memory constraints. Following Inspired



(a) Benign lung CT scan images from the LIDC-IDRI dataset.

(b) Malignant lung CT scan images from the LIDC-IDRI dataset.

Fig. 2: Representative axial CT slices for benign and malignant cases.

by prior work on transferring knowledge from models pretrained on natural images [23], we normalized the CT scan intensities using the standard ImageNet statistics. Specifically, we used a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225]. To improve generalization and reduce overfitting, we applied a diverse set of on-the-fly data augmentations during training. These included geometric transformations such as random horizontal and vertical flips, rotations up to 30 degrees, along with affine transformations that include translations of up to 10%, scaling factors ranging from 0.9 to 1.1, and shear angles within 5 degrees. In addition, we applied photometric augmentations using color jittering, which randomly adjusted the brightness, contrast, and saturation of the images by up to 0.2, and modified the hue within a range of 0.05. To improve robustness against occlusions and irregularities in the data, we incorporated Random Erasing with a probability of 0.2, where small random regions covering between 2% and 20% of the image area were masked out. To enhance generalization, we adopted the MixUp data augmentation strategy proposed by Zhang et al. [24], which generates synthetic training examples by linearly combining pairs of images and their corresponding labels. For each minibatch during training, with a probability of 0.7, a mixing coefficient λ is sampled from a Beta distribution with both shape parameters set to 0.4:

$$\lambda \sim \mathrm{Beta}(\alpha, \alpha) \tag{1}$$

Given two randomly selected training samples (x_i, y_i) and (x_j, y_j), the augmented sample $(\tilde{x}, \tilde{y})$ is computed as:

$$\tilde{x} = \lambda \cdot x_i + (1 - \lambda) \cdot x_j \tag{2}$$

$$\tilde{y} = \lambda \cdot y_i + (1 - \lambda) \cdot y_j \tag{3}$$

As defined in Equations 2 and 3, this approach encourages the model to behave linearly between training examples, which is particularly beneficial in medical imaging tasks where class boundariessuch as between benign and malignant nodulescan be ambiguous.

During validation and testing, only resizing and normalization were performed to maintain consistency and prevent any data augmentation from influencing the evaluation results. To address the ∼5:1 benignmalignant imbalance, we used a WeightedRandomSampler during training. Class weights were computed as:

$$w_c = \frac{1/n_c}{\sum_k (1/n_k)} \times C, \tag{4}$$

yielding weights [0.3258, 1.6742] for benign and malignant classes. Equation 4 shows that this sampler oversamples malignant cases each epoch, mitigating majority class bias more effectively than simple loss weighting or undersampling.

4 Proposed Method

The proposed Multi-Attention Stacked Ensemble (MASE) architecture integrates a total of three state-of-the-art convolutional neural networks as its foundation. Each base model was selected for its complementary strengths in feature extraction and representation learning, providing diverse perspectives on the lung nodule classification task.

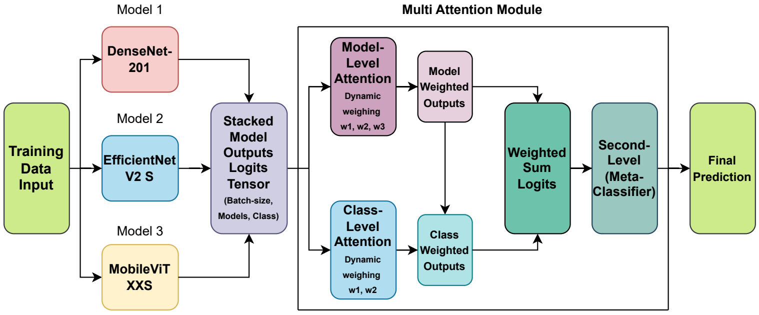

Model 1
DenseNet-201
Model 2
EfficientNet V2 S
Model 3
MobileViT XXS
Training Data Input
Stacked Model Outputs Logits Tensor (Batch-size, Models, Class)
Multi Attention Module
Model-Level Attention Dynamic weighing w1, w2, w3
Model Weighted Outputs
Class-Level Attention Dynamic weighing w1, w2
Class Weighted Outputs
Weighted Sum Logits
Second-Level (Meta-Classifier)
Final Prediction

Fig. 3: End-to-end pipeline of the proposed Multi-Attention Stacked Ensemble (MASE) framework for lung nodule classification. The architecture begins with training data processed by three distinct CNN backbones: DenseNet-201, EfficientNetV2-S, and MobileViT-XXS. Their outputs are fused using multi attention mechanisms (model-level and class-level attention), followed by a meta-classifier that generates the final prediction.

An overview of the complete classification pipeline is shown in Figure 3.

DenseNet-201 is employed as one of the backbone networks in our ensemble framework due to its ability to encourage feature reuse and alleviate the vanishing gradient problem through dense connectivity [25]. The architecture comprises four dense blocks with 6, 12, 48, and 32 layers, respectively, interleaved with transition layers that perform downsampling. Each layer within a dense block receives the concatenated outputs of all preceding layers, resulting in improved gradient flow and more compact parameter usage compared to traditional CNNs. To adapt DenseNet-201 for lung nodule classification, we modified the original architecture in several ways. The default classification head was replaced with a custom binary classifier tailored to our task, as detailed in Section 4.2. The input resolution was adjusted to 9696 pixels to match the size of the CT patches used in our dataset. Furthermore, we utilized ImageNet pretrained weights for initialization, except for the final classification layer, which was randomly initialized to accommodate the binary nature of our problem. These adjustments allow DenseNet-201 to maintain its multiscale feature learning capabilities while being specialized for identifying subtle patterns in lung nodules.

EfficientNetV2-S is adopted as the second backbone in our ensemble for its balance between model accuracy and computational efficiency [26]. Building upon the original EfficientNet, this architecture integrates both MBConv and Fused-MBConv blocks and follows a progressive learning strategy that enhances training speed and generalization. The model comprises seven stages, each containing blocks with varying expansion factors, strides, and channel widths, beginning with a standard convolutional layer and concluding in a classification head. To tailor EfficientNetV2-S for lung nodule classification, we made several architectural adjustments. The first convolutional layers stride was reduced from 2 to 1 while maintaining a 33 kernel with padding 1, allowing the network to retain finer spatial features that are critical for identifying small nodules. We also replaced the default classifier head with a task-specific binary classifier, and pretrained weights from ImageNet were used to initialize all layers except those newly added or modified. These changes enable the model to handle 9696 CT scan inputs more effectively, preserving spatial granularity in early layers and maintaining the models strength in extracting relevant features for classification.

MobileViT-XXS serves as the third backbone in our ensemble, blending convolutional inductive biases with transformerbased global attention for an efficient yet powerful feature extractor [27]. It begins with a MobileNetstyle stem, then alternates between local convolutions (pointwise and depthwise) and lightweight transformer blocks that unfold spatial maps into token sequences and refold them after selfattention. This design captures finegrained local details and longrange context in a single model. For lung nodule classification, we replaced its default classification head with a taskspecific binary classifier, updated the input pipeline to handle 9696 CT patches, and initialized with ImageNet weights for all unchanged layers. These adaptations allow MobileViT-XXS to preserve its global reasoning capability while concentrating on subtle nodule characteristics amid surrounding lung tissue.

A key innovation of our approach is the development of a unified model adapter that replaces the standard classification heads of each base model. This custom head was specifically designed to address key challenges in medical image classification, such as limited annotated data and overfitting, and the critical need for wellcalibrated confidence estimates. The adapter begins with a dropout layer set to 50 percent, providing strong regularization by randomly omitting half of the features. Next comes a linear projection that compresses the backbones features into a fixed 256dimensional space, this projection step ensures that all three models, even though they use different architectures, produce outputs on the same scale. We use Layer Normalization rather than BatchNorm to stabilize training when batch sizes are small, as is typical in medical imaging, and follow this with a ReLU activation to introduce nonlinearity. Before the final classification, a second dropout layer at 30 percent further discourages overreliance on any single feature. This twostage dropout strategy not only combats overfitting but also encourages the adapter to learn a more robust, distributed representation of the input. By projecting each backbone into a common feature space, we enable direct comparison and seamless fusion in the attention ensemble that follows. Moreover, the consistent 256dimensional bottleneck significantly reduces the number of parameters in the final layers, acting as a form of architectural regularization that supports better generalization. LayerNorms contribution to stable gradient flow also yields more reliable probability outputs, which is crucial in clinical settings where decision confidence must be trustworthy. Altogether, these design choices transform generalpurpose vision architectures into specialized, parameterefficient classifiers that are well suited to the unique demands of lung nodule detection.

4.1 Proposed Ensemble Method

The core innovation of our architecture lies in the Multi-Attention Stacked Ensemble, which adaptively integrates predictions from the three base models using a dual-attention mechanism. Unlike traditional ensembles with static weights or simple averaging, our method dynamically modulates each model and class contribution based on the input, enhancing decision precision.

The model-level attention module dynamically assigns importance weights to each base model's predictions, allowing the ensemble to leverage the unique strengths of different architectures for varying nodule characteristics. Rather than using fixed weights, this approach recognizes that certain models may excel on specific nodule subtypes for example, MobileViT-XXS's transformer components might prove more valuable for cases requiring global contextual reasoning, while DenseNet-201's dense connectivity could better capture fine-grained textural patterns.The module computes attention weights w_1, w_2, and w_3 for DenseNet-201, EfficientNetV2-S, and MobileViT-XXS respectively, where these weights sum to 1 and represent each model's relative contribution to the final prediction. Importantly, these weights are calculated dynamically for each input sample rather than remaining fixed across the dataset.

Given the stacked outputs from three base models with tensor shape $[B, M, C]$ (where B is batch size, M is number of models, and C is number of classes), the mechanism first flattens this representation to $[B, M \times C]$. This flattened vector then passes through an MLP consisting of: linear transformation $\mathbb{R}^{M\times C} \rightarrow \mathbb{R}^{128}$, LayerNorm, ReLU activation, dropout layer with a rate of 0.3, linear transformation $\mathbb{R}^{128} \rightarrow \mathbb{R}^{M}$, and softmax activation. The resulting output provides attention weights $[w_1, w_2, w_3]$ for each sample. The inclusion of LayerNorm and dropout within the attention MLP promotes stable training while preventing overfitting to specific training patterns.

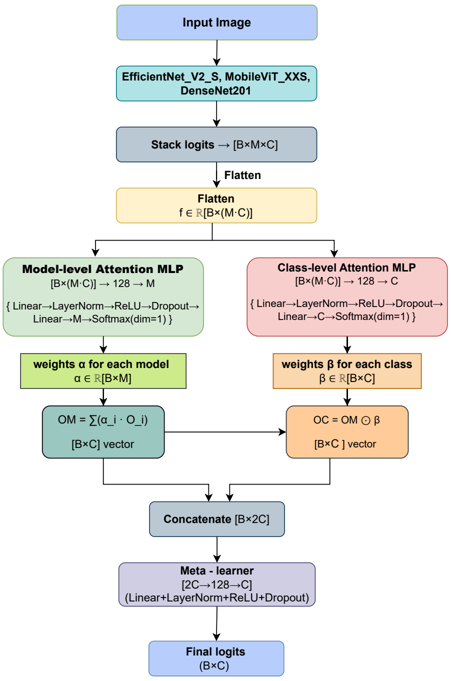


Fig. 4: Schematic diagram of the dual attention mechanism in our Multi-Attention Stacked Ensemble (MASE) architecture. The model-level attention assigns weights to base models, while the class-level attention refines predictions at the class level for each input sample.

The 128-dimensional intermediate layer offers sufficient capacity for learning complex relationships between model outputs while maintaining computational efficiency.

While model-level attention decides which backbone to trust, class-level attention provides a second weighting dimension by modulating the importance of each class prediction. This allows the ensemble to selectively emphasize certain classes based on input characteristics for instance, boosting malignant predictions in ambiguous cases to increase sensitivity, which is crucial in medical screening where false negatives are costlier than false positives.

The class-level attention uses the same flattened representation $[B, M \times C][B, M \times C]$ $[B, M \times C]$ and passes it through a parallel MLP: $\mathbb{R}^{M \times C} \rightarrow 128$, LayerNorm ReLU Dropout(0.3) $\mathbb{R}^C$ Softmax, producing per-class weights. These weights modulate the model-weighted predictions element-wise, introducing additional non-linearity that enables more complex decision boundaries than model-level attention alone. This dual-attention approach provides learned, input-dependent re-scaling that captures patterns beyond standard softmax operations.

The meta-learner module represents the final integration stage of our ensemble architecture, synthesizing outputs from both attention mechanisms to produce refined class predictions. Rather than directly utilizing attention-weighted outputs, we introduce a dedicated meta-learner capable of discovering complex interactions between the dual attention streams. Our approach concatenates the model-level and class-level weighted outputs, creating a feature representation of dimensionality $2C$. This enriched representation undergoes processing through a compact MLP architecture: $\mathbb{R}^{2C} \rightarrow \mathbb{R}^{128}$ with LayerNorm and ReLU activation, followed by dropout regularization (rate 0.3) and a final projection to $\mathbb{R}^C$ for class logits. The integration process follows a principled approach where model-level attention first produces weighted combinations of base model outputs, followed by class-level modulation of these weighted predictions. Formally, given model-weighted output $\mathbf{m} = \sum_i (\mathbf{s}_i \cdot w_i^{(m)})$ and class-weighted output $\mathbf{c} = \mathbf{m} \odot \mathbf{w}^{(c)}$, the meta-learner processes the concatenated representation $[\mathbf{m}, \mathbf{c}]$ to generate final predictions. This architecture enables the meta-learner to exploit complementary information from both attention mechanisms while maintaining the flexibility to adaptively weight their relative contributions based on input characteristics, ultimately enhancing the ensemble's discriminative capacity.

Given the inherent class imbalance in lung nodule datasets, we adopt a dynamic focal loss strategy that integrates class weighting with focal modulation to address both statistical bias and learning difficulty [28]. This approach proves particularly valuable in medical imaging where minority class detection is paramount.

We derive class weights inversely proportional to class frequencies, effectively amplifying the learning signal from underrepresented samples:

$$\text{class\_weights}_k = \frac{1/\text{class\_counts}_k}{\sum_j (1/\text{class\_counts}_j)} \times \text{num\_classes} \tag{5}$$

These weights (0.33 for benign, 1.67 for malignant) serve as the α parameter in our focal loss formulation, which incorporates the characteristic $(1 - p_t)^\gamma$ modulation to suppress well-classified examples:

$$\text{FL}(p_t) = -\alpha_t (1 - p_t)^\gamma \log(p_t) \tag{6}$$

where p_t represents the predicted probability for the ground truth class, α_t denotes the corresponding class weight, and γ controls the focusing intensity. Through systematic evaluation, we establish γ of 2.0 as optimal, providing sufficient emphasis on challenging cases while maintaining training stability. This configuration effectively balances sensitivity for malignant detection against overall classification performance.

Test-time augmentation enhances prediction reliability by averaging model outputs across systematically transformed test inputs [29]. Rather than applying random transformations, we employ a deterministic augmentation set tailored for lung nodule analysis through empirical validation.

Our transformation pipeline encompasses the original image alongside horizontal/vertical flips, 90-degree rotations, and modest brightness, contrast, and saturation adjustments of ten percent. These modifications capture natural variations in nodule presentation while preserving diagnostically critical morphological features. We deliberately avoid aggressive geometric distortions that could compromise shape characteristics essential for accurate diagnosis. The aggregation process combines predictions from all transformed versions using mean averaging in logit space:

$$\text{TTA}(x) = \frac{1}{N+1}\Big(f(x) + \sum_{i=1}^{N} f(T_i(x))\Big) \tag{7}$$

where x represents the input image, f denotes the model function, T_i corresponds to the i-th transformation, and N indicates the total transformations applied. This logit-space averaging proves more robust than probability-based aggregation or voting schemes. Notably, MobileViT-XXS demonstrates the greatest sensitivity to augmentation, likely reflecting its transformer architecture's responsiveness to input variations. By applying TTA independently to each base model before ensemble fusion, we amplify both architectural diversity and augmentation-induced robustness.

5 Training Setup

We conducted all experiments on the Kaggle platform using an NVIDIA Tesla P100 GPU with 16 GB VRAM, which provided sufficient computational power for training our complex ensemble architecture. Our implementation was built using PyTorch 2.5.1 with CUDA version 12.4 and Python 3.11.11, taking advantage of its dynamic computation graph capabilities that are well-suited for custom ensemble architectures.

After extensive experimentation and hyperparameter tuning, we established an optimal configuration for our Multi-Attention Stacked Ensemble (MASE) model. We used a batch size of 64 with input images resized to 9696 pixels, training for a maximum of 200 epochs. To prevent overfitting, we implemented early stopping with a patience of 60 epochs, meaning training would halt if validation accuracy failed to improve for 60 consecutive epochs. We initialized training with a learning rate of 510^{-4} and applied weight decay regularization at 110^{-4} to improve generalization.

Our learning rate scheduling strategy employed cosine annealing warm restarts, which creates cyclical learning rate patterns that help the model escape local minima while maintaining overall convergence. The scheduler began with an initial period of 10 epochs, then doubled the cycle length after each restart (creating cycles of 10, 20, 40, 80, and 160 epochs), with a minimum learning rate floor of 110^{-6}. To work around GPU memory constraints while effectively increasing our batch size, we accumulated gradients over 2 steps before updating model parameters. For data augmentation, we applied MixUp with an alpha value of 0.4, which helps the model generalize better by creating synthetic training examples through linear interpolation between different samples.

Table 1: Training Optimization Techniques

Optimization Technique	**Configuration**	**Purpose**
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$)	Decoupled weight decay regularization
Loss Function	Dynamic Focal Loss ($\gamma = 2.0$)	Address class imbalance
Learning Rate Scheduler	CosineAnnealingWarmRestarts	Escape local minima
Gradient Accumulation	2 steps	Effective batch size increase
Early Stopping	Patience = 60 epochs	Prevent overfitting
Data Augmentation	MixUp ($\alpha = 0.4$)	Improve generalization

We selected AdamW as our primary optimizer following the approach of Loshchilov & Hutter [30], which extends traditional Adam optimization by implementing decoupled weight decay regularization. This choice was particularly important for our heterogeneous ensemble architecture, as AdamW better handles the varying scales of gradients across different network layers compared to standard Adam with L2 regularization. The optimizer parameters were set with beta values of 0.9 and 0.999 for the exponential decay rates, epsilon of 110^{-8} for numerical stability, and the weight decay coefficient of 110^{-4} applied in a decoupled manner.

For our loss function, we implemented Dynamic Focal Loss with a gamma value of 2.0 to address the class imbalance inherent in our lung nodule dataset. This loss function helps the model focus more attention on hard-to-classify examples while down-weighting the contribution of easy examples during training.

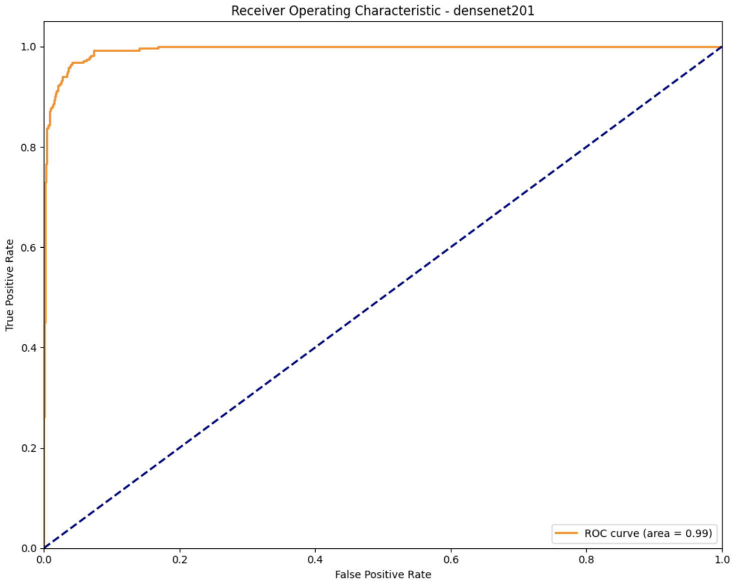


Fig. 5: ROC Curve DenseNet-201

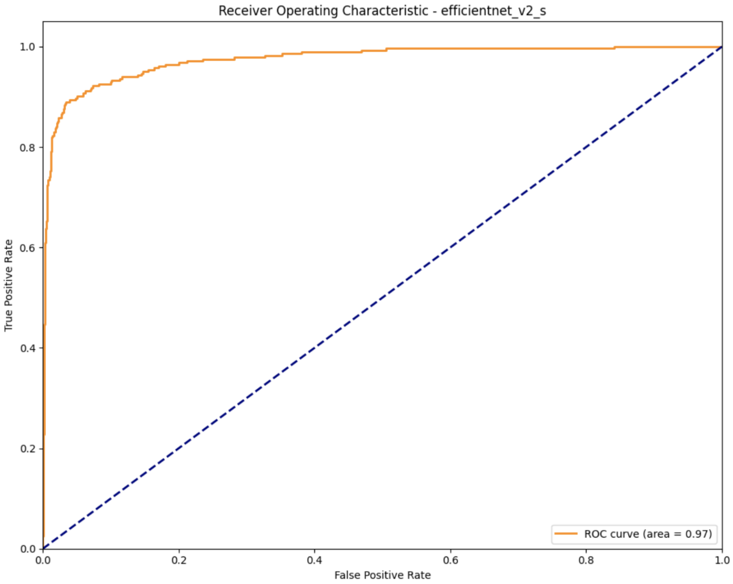


Fig. 6: ROC Curve EfficientNetV2-S

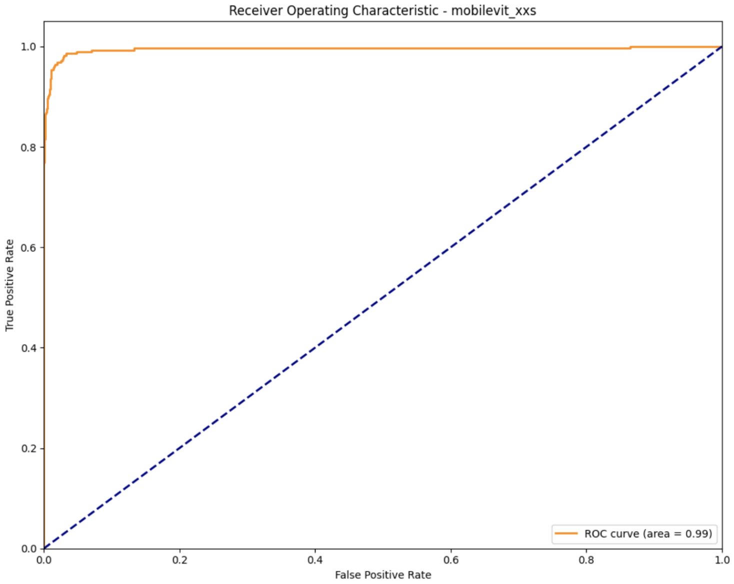


Fig. 7: ROC Curve MobileViT-XXS

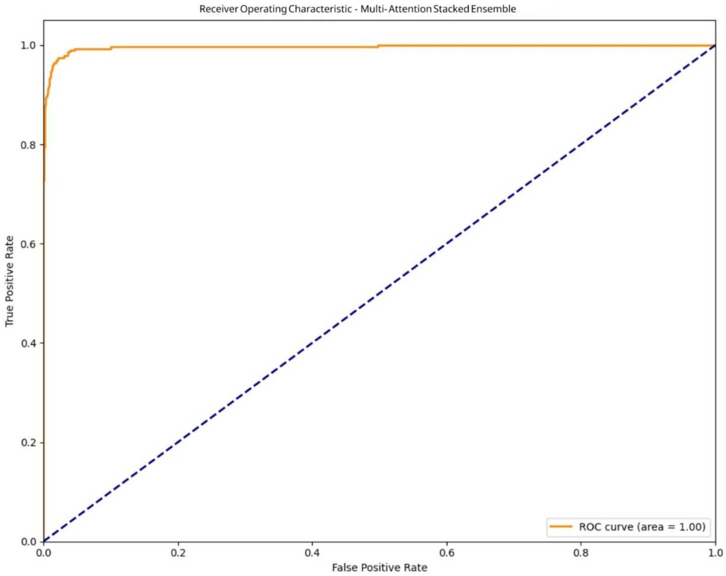


Fig. 8: ROC Curve Multi-Attention Stacked Ensemble (MASE)

For evaluation, we employed a comprehensive set of classification metrics to thoroughly assess our model's performance. We measured accuracy to understand overall correctness, precision and recall to evaluate class-specific performance, and weighted F1-score to account for class imbalance. ROC-AUC served as our primary metric for threshold-independent performance assessment, which is particularly valuable for medical applications where decision thresholds may vary based on clinical requirements. This metric also handles class imbalance well and provides a probabilistic interpretation of model performance.

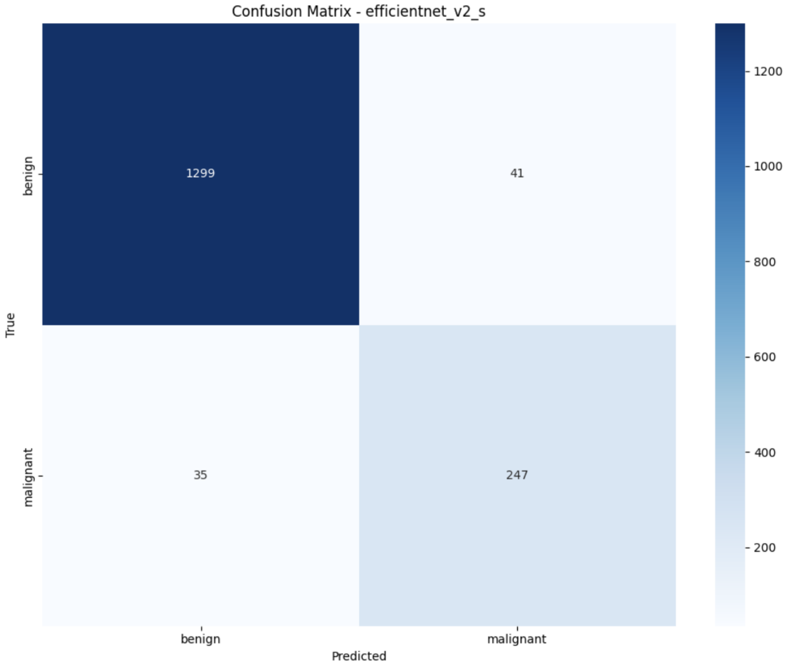


Fig. 9: Confusion Matrix EfficientNetV2-S

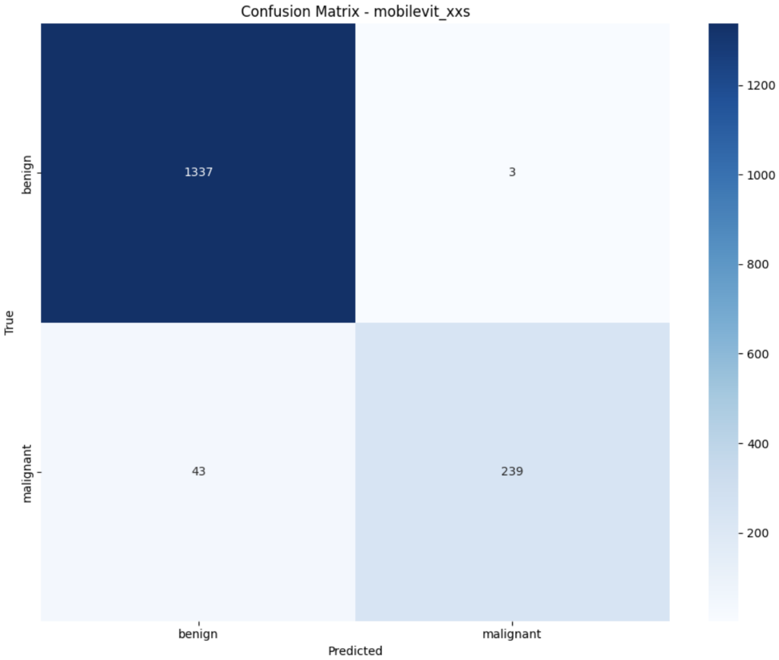


Fig. 10: Confusion Matrix MobileViT-XXS

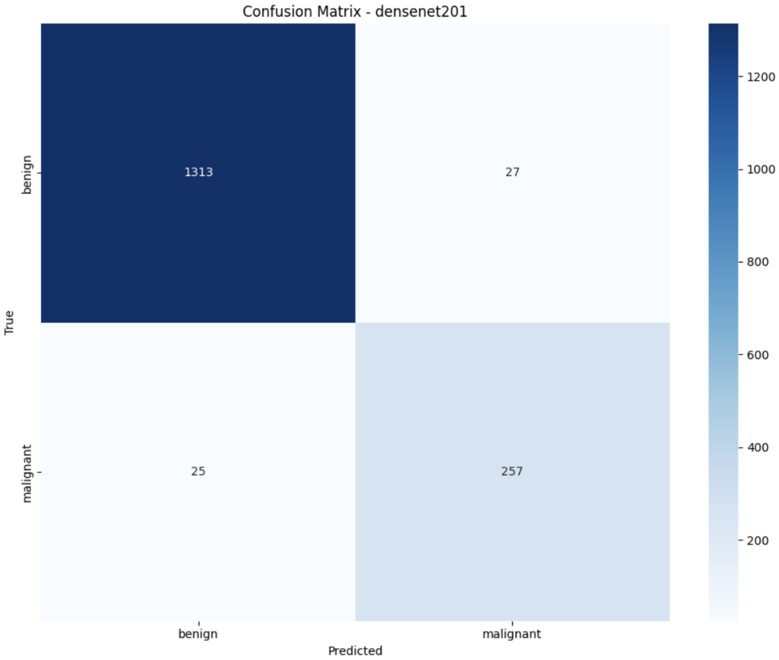


Fig. 11: Confusion Matrix DenseNet-201

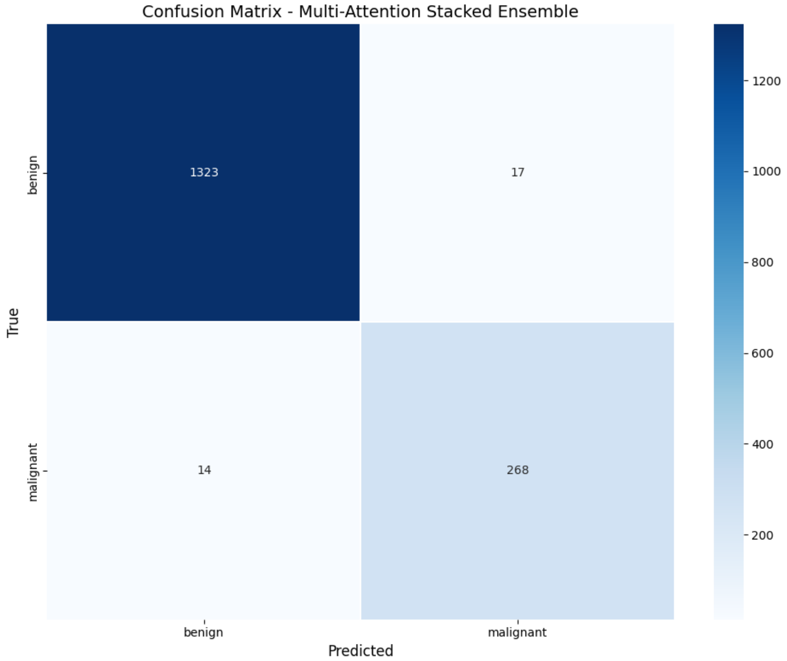


Fig. 12: Confusion Matrix Multi-Attention Stacked Ensemble (MASE)

We also generated detailed confusion matrices for each model to understand exactly where classification errors occurred. These matrices helped us identify whether models were more prone to false positives (incorrectly classifying benign nodules as malignant) or false negatives (missing actual malignant cases), which is crucial information for medical diagnostic applications where the cost of different types of errors varies significantly.

6 Experimental Results

This section presents the results of our lung nodule classification experiments using the Multi-Attention Stacked Ensemble (MASE) architecture, which achieved state-of-the-art performance with 98.09% accuracy and 0.9961 AUC on the LIDC-IDRI dataset.

We evaluated three base models selected for their complementary architectures and feature extraction capabilities. Table 2 summarizes their individual performance on the test set.

DenseNet-201 achieved balanced performance through its dense connectivity pattern, which proved particularly effective at capturing subtle textural patterns like spiculation and irregular margins characteristic of malignant nodules. The model's feature reuse mechanism, where early layers' features are directly utilized by deeper

Table 2: Individual model performance on LIDC-IDRI test set

Model	Accuracy (%)	AUC	Precision (B/M)	Recall (B/M)	FP	FN	Total Errors
DenseNet-201	96.79	0.9936	0.98/0.90	0.98/0.91	25	27	52
EfficientNetV2-S	95.31	0.9744	0.97/0.86	0.97/0.88	35	41	76
MobileViT-XXS	97.16	0.9945	0.97/0.99	1.00/0.85	43	3	46

B: Benign, M: Malignant, FP: False Positives, FN: False Negatives

layers, contributed to robust performance despite dataset imbalance, achieving nearly equal precision and recall for both classes.

EfficientNetV2-S achieved the lowest performance among the three models but still demonstrated strong results considering its parameter efficiency. The model excelled at detecting well-defined nodules with clear boundaries but showed limitations when dealing with ambiguous cases where subtle features determine malignancy. Its compound scaling approach enabled competitive performance particularly for benign nodules, though with a higher false negative rate that could be clinically concerning.

MobileViT-XXS emerged as the strongest individual model, leveraging its hybrid architecture that combines convolutional layers with transformer blocks. The transformer components enabled effective modeling of long-range dependencies within images, proving valuable for capturing relationships between nodule features and surrounding tissue context. However, the model exhibited a notable tendency toward high specificity, with 43 false positives but only 3 false negatives, suggesting a conservative approach to malignancy detection.

The proposed Multi-Attention Stacked Ensemble successfully integrated these complementary strengths to achieve superior performance. Figure 13 illustrates the clear advantage of our ensemble approach over individual models.

Our ensemble achieved precision values of 0.99 for benign nodules and 0.94 for malignant nodules, with corresponding recall values of 0.99 and 0.95. The F1-scores of 0.99 and 0.95 for benign and malignant classes respectively indicate well-balanced performance crucial for clinical applications. The confusion matrix revealed only 31 misclassifications out of 1,622 test samples, with 17 false negatives and 14 false positives, representing a 35% reduction in error rate compared to the best individual model.

The ensemble's error reduction was substantial across all base models: 32.6% compared to MobileViT-XXS, 59.2% compared to DenseNet-201, and 67.4% compared to EfficientNetV2-S. While AUC improvements of 0.0016, 0.0025, and 0.0217 may appear modest, they represent significant gains in the high-performance regime where baseline AUC already exceeds 0.97.

To contextualize these results, we compared our approach against state-of-the-art methods from recent literature. Table 3 demonstrates MASE's superior performance across all evaluation metrics.

Our method achieved a 0.86% accuracy improvement over the previous best results, with particularly notable gains in sensitivity. Recent approaches like Xie et al.'s semi-supervised adversarial model [32] and Bushra et al.'s LCD-CapsNet [34]

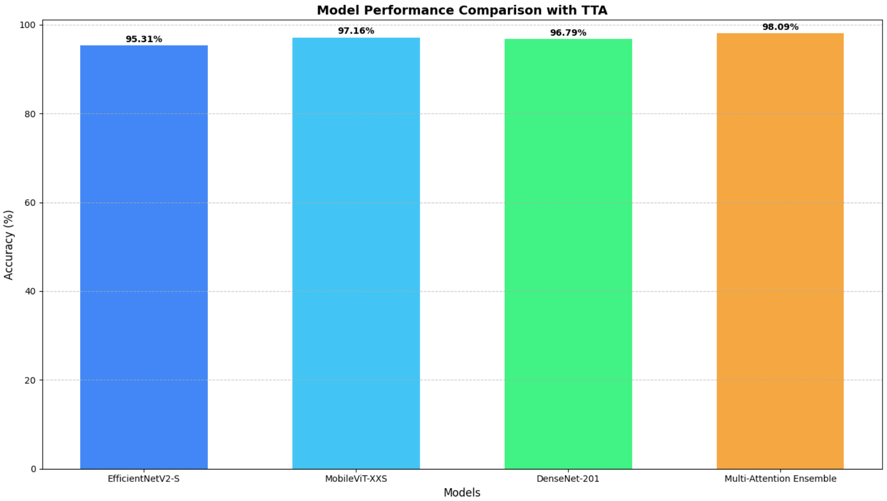


Fig. 13: Accuracy comparison of individual models and the proposed Multi-Attention Stacked Ensemble (MASE) with Test-Time Augmentation (TTA). The MASE architecture outperforms all base models, achieving the highest accuracy of 98.09%.

Table 3: Comparison of our proposed MASE with state-of-the-art approaches on LIDC-IDRI.

Method	Architecture	Year	Accuracy (%)	AUC	Sensitivity (%)	Specificity (%)
D. Zhao et al. [31]	GAN + VGG16	2018	95.24	0.9800	98.67	92.47
Xie et al. [32]	Semi-sup. Adv.	2021	95.68	0.9512	93.60	96.20
A. Halder et al. [33]	2-Path CNN	2022	95.17	0.9936	96.85	96.10
Bushra et al. [34]	LCD CapsNet	2023	94.00	0.9890	94.50	99.07
Gautam et al. [35]	Weighted Ens.	2024	97.23	0.9468	98.60	94.20
Proposed MASE	**Multi-Att. Ens.**	2025	**98.09**	**0.9961**	**98.73**	**98.96**

achieved strong results through architectural innovations, but our multi-attention integration strategy proved more effective. The dynamic attention mechanism adapts to each input, unlike traditional ensembles using fixed weights, while our dual-attention approach explicitly addresses both model-level and class-level weighting.

6.1 Statistical Significance Test

To validate statistical significance, we performed Wilcoxon signed-rank tests comparing our ensemble against each base model across 20 independent runs with different random initializations. The Wilcoxon test is a non-parametric statistical procedure that evaluates whether the median difference in paired observations equals zero [36], making it well-suited for model performance comparisons as it makes no assumptions about normality and is robust to outliers. We applied Bonferroni correction for multiple comparisons, adjusting the significance threshold to $\alpha = 0.0167$.

The ensemble achieved statistically significant improvements over all base models ($p < 0.001$ for all comparisons), demonstrating mean accuracy of 97.53% (±0.41%) versus 93.07% (±4.13%) for EfficientNetV2-S, 95.31% (±4.07%) for MobileViT-XXS, and 90.40% (±8.61%) for DenseNet-201. The most substantial improvement was against DenseNet-201, which showed highest variability ($\sigma = 8.61\%$). Even compared to the strongest individual performer (MobileViT-XXS at 95.31%), our ensemble showed significant improvements ($p = 0.000006$). The Multi-Attention Stacked Ensemble achieved an average 4.60% accuracy improvement with extremely low p-values ($p = 0.000004$ vs. EfficientNetV2-S, $p = 0.000006$ vs. MobileViT-XXS, $p = 0.000250$ vs. DenseNet-201), all well below the adjusted threshold, confirming robust and consistent performance gains.

The clinical implications of these performance gains are substantial. The 0.86% accuracy improvement translates to approximately 8.6 fewer misclassifications per 1,000 nodules examined. Given millions of annual chest CT scans, this could prevent thousands of diagnostic errors [37]. Our enhanced sensitivity of 98.73% for malignant nodules directly addresses the critical concern of missed cancers, where early detection can improve 5-year survival rates by 20-40% [38].

High specificity (98.96%) reduces unnecessary follow-up procedures, each costing \$800-\$1,500 for additional imaging or over \$15,000 for invasive diagnostics [39].

6.2 GradCAM++

To better understand model decision-making, we employed GradCAM++ visualization, which produces high-resolution, class-discriminative explanations by analyzing gradient flow to the final convolutional layers [40]. GradCAM++ was applied to each base model (DenseNet-201, EfficientNetV2-S, and MobileViT-XXS) to generate heatmaps highlighting the regions most influential in their predictions. Because, compared to the original GradCAM, GradCAM++ offers improved localization of multiple instances of the same class and generates more reliable explanations in complex medical imaging scenarios [41]. These visualizations reveal how each architecture focuses on distinct features of lung nodule images, offering valuable clinical insights into model decision-making and enhancing understanding of the Multi-Attention Stacked Ensemble's superior performance. Each heatmap is also accompanied by a confidence score based on the maximum softmax probability, reflecting the model's certainty in its prediction.

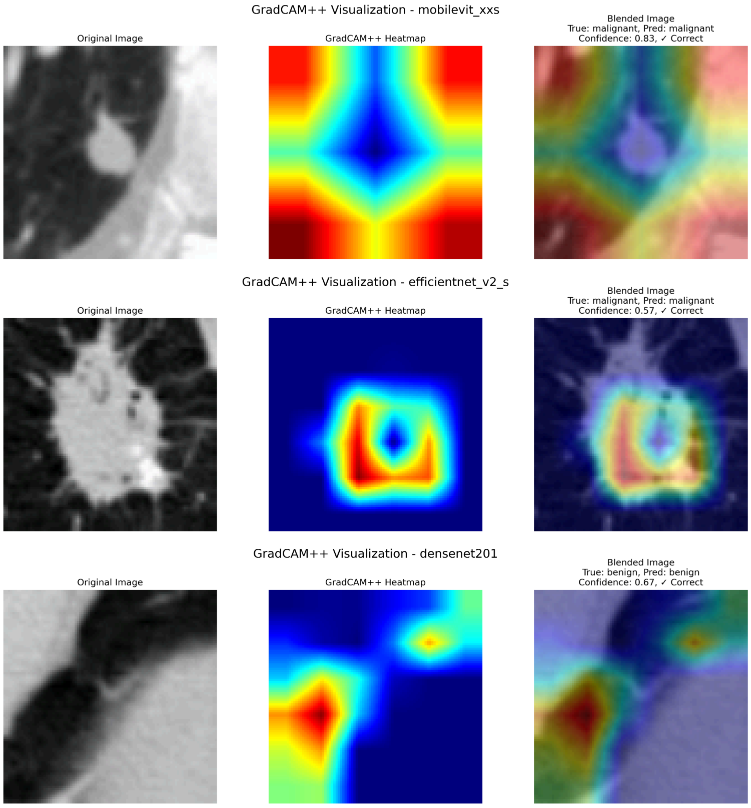

GradCAM++ Visualization - mobilevit_xxs
Original Image
GradCAM++ Heatmap
Blended Image
True: malignant, Pred: malignant
Confidence: 0.83, ✓ Correct
GradCAM++ Visualization - efficientnet_v2_s
Original Image
GradCAM++ Heatmap
Blended Image
True: malignant, Pred: malignant
Confidence: 0.57, ✓ Correct
GradCAM++ Visualization - densenet201
Original Image
GradCAM++ Heatmap
Blended Image
True: benign, Pred: benign
Confidence: 0.67, ✓ Correct

Fig. 14: Comparison of GradCAM++ visualizations across base models for representative lung nodule samples.

Figure 14 illustrates these activation maps, showing how the ensemble effectively combines the complementary feature extraction strengths of individual base models.

7 Conclusion

In this thesis, we presented the Multi-Attention Stacked Ensemble (MASE) for lung nodule classification, integrating model-level and class-level attention with a unified model adapter and Dynamic Focal Loss. On the LIDC-IDRI dataset, MASE achieved 98.09% accuracy and an AUC of 0.9961, translating to a 35% reduction in error rate over previous state-of-the-art methods while preserving high sensitivity (98.73%) and specificity (98.96%).

The key technical contributions include the dual-attention fusion mechanism that adaptively weights both model predictions and class logits, the custom 256-dimensional adapter head that ensures feature consistency and robust regularization. Together, these innovations enable the ensemble to dynamically emphasize the most reliable model or class output for each input, resulting in more nuanced decision boundaries and improved overall performance.

For future work, we envisage three main directions. First, extending MASE to multi-class Lung-RADS categorization would provide clinically actionable risk stratification aligned with reporting standards [42]. Second, we plan to experiment with alternative activation functions, such as the Gompertz function, in our attention and meta-learner modules, aiming to refine convergence behavior and calibration under class imbalance. Third, unifying nodule detection, segmentation, and classification into a single end-to-end architecture could leverage multi-task learning and enable extraction of quantitative biomarkers (e.g., volume, surface-to-volume ratio, texture heterogeneity) for richer clinical insights [43]. Additionally, incorporating explainability techniques like feature attribution, concept-based explanations, counterfactuals, and uncertainty quantification will be essential for transparent decision support and regulatory acceptance in clinical workflows [44].

By combining advanced attention-based fusion, rigorous loss design, and explainability, MASE lays the groundwork for AI systems that not only match but complement radiologists expertise, moving closer to real-world deployment in lung cancer screening and beyond.

Our dynamic attention approach demonstrates a pathway toward adaptive AI systems that acknowledge medical data complexity. Despite implementation challenges including validation and regulatory requirements, this work provides a foundation for AI systems that complement human expertise in diagnostic processes, advancing toward collaborative healthcare AI that enhances radiologist capabilities.

These extensions will help translate MASE into robust, interpretable tools that augment radiologists expertise and advance AIassisted lung cancer screening.

References

[1] Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A.: Cancer statistics, 2023. CA: a cancer journal for clinicians **73**(1), 17–48 (2023)

[2] Armato, S.G., McLennan, G., Bidaut, L., al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical Physics **38**(2), 915–931 (2011) https://doi.org/10.1118/1.3528204

[3] El-Baz, A., Beache, G.M., Gimel farb, G., Suzuki, K., Okada, K., Elnakib, A., Soliman, A., Abdollahi, B.: Computer-aided diagnosis systems for lung cancer: challenges and methodologies. International journal of biomedical imaging **2013**(1), 942353 (2013)

[4] Kostis, W.J., Reeves, A.P., Yankelevitz, D.F., Henschke, C.I.: Relationships between nodule location, density, and observer detection performance on low-dose ct images. IEEE Transactions on Medical Imaging **32**(12), 2489–2499 (2013) https://doi.org/10.1109/TMI.2013.2275072

[5] Way, T., Fritscher, K., Mller, H., Noblesse, P.: Texture analysis of pulmonary nodules on ct for categorization using machine learning. Computerized Medical Imaging and Graphics **36**(4), 227–239 (2012) https://doi.org/10.1016/j.compmedimag.2011.12.009

[6] Zhao, B., Kligerman, S., Du, H., GalperinAizenberg, M., Zhang, P., Schwartz, L.H.: Lung nodule classification on ct images using waveletbased multiscale texture analysis and support vector machines. Medical Physics **37**(1), 570–579 (2010) https://doi.org/10.1118/1.3259774

[7] Murphy, K., Ginneken, B., Reinhardt, J.M., Kabus, S., Ding, K., Du, H., Tempany, C.M., Summers, R.M.: A largescale evaluation of automatic pulmonary nodule detection in chest ct using local image features and knearestneighbor classification. Medical Image Analysis **13**(5), 757–770 (2009) https://doi.org/10.1016/j.media.2009.05.004

[8] Froz, B.R., Carvalho Filho, A.O., Silva, A.C., Paiva, A.C., Nunes, R.A., Gattass, M.: Lung nodule classification using artificial crawlers, directional texture and support vector machine. Expert Systems with Applications **69**, 176–188 (2017)

[9] Shen, W., Zhou, M., Yang, F., Yang, C., Tian, J.: Multiscale convolutional neural networks for lung nodule classification. In: Information Processing in Medical Imaging IPMI 2015. Lecture Notes in Computer Science, vol. 9123, pp. 588–599. Springer, ??? (2015). https://doi.org/10.1007/978-3-319-19992-4_47

[10] Nam, J.., Park, J., Hwang, E.J., Lee, J.H., Jin, K.., Lim, K.., Kim, H.K.: Deep learningbased classification of pulmonary nodules on computed tomography images. Journal of Digital Imaging **32**(4), 688–693 (2019) https://doi.org/10.1007/s10278-019-00211-0

[11] Liao, F., Liang, M., Li, Z., Hu, X., Song, S.: Evaluate the malignancy of pulmonary nodules using the 3d deep leaky noisyor network. IEEE Transactions on Neural Networks and Learning Systems **31**(8), 2484–2495 (2020) https://doi.org/10.1109/TNNLS.2019.2917606

[12] Li, X., Cheng, J.., Chen, W., Ni, D., Lei, B., Wang, T.: Translung: Transformer-based model for lung nodule classification in ct. Medical Image Analysis **75**, 102251 (2022) https://doi.org/10.1016/j.media.2021.102251

[13] Sagi, O., Rokach, L.: Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(4), 1249 (2018) https://doi.org/10.1002/widm.1249

[14] Antolovi, M., MajkiSingh, N., Vasiljevi, A.: Attentionbased feature fusion for lung nodule classification using multiple cnns. In: Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), pp. 543–551 (2019)

[15] Zhang, Y., Jiang, J., Li, Y.: Multilevel attention ensemble for pulmonary nodule classification. IEEE Access **9**, 45678–45687 (2021) https://doi.org/10.1109/ACCESS.2021.3066782

[16] Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 510–519 (2019)

[17] Schlemper, J., Oktay, O., Chen, L., Matthew, J., Knight, C., Kainz, B., Glocker, B., Rueckert, D.: Attention-Gated Networks for Improving Ultrasound Scan Plane Detection (2018). https://arxiv.org/abs/1804.05338

[18] Lin, T.., Goyal, P., Girshick, R., He, K., Dollr, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988 (2017)

[19] Zhang, H., Xia, Y., Sun, J.: Dynamic focal loss for class imbalance in medical image classification. IEEE Transactions on Medical Imaging **40**(12), 3568–3578 (2021) https://doi.org/10.1109/TMI.2021.3095400

[20] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: Mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (ICLR) (2018)

[21] Wang, X., Peng, Y., Lu, L., *et al.*: Testtime augmentation for improved inference in medical imaging. In: IEEE 15th International Symposium on Biomedical Imaging (ISBI), pp. 1415–1418 (2018). https://doi.org/10.1109/ISBI.2018.8363601

[22] Chen, S., Zhang, H., Dong, Z., *et al.*: Enhanced lung nodule classification with testtime augmentation. IEEE Journal of Biomedical and Health Informatics **24**(6), 1621–1631 (2020) https://doi.org/10.1109/JBHI.2019.2930469

[23] Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. In: Advances in Neural Information Processing Systems, vol. 32, pp. 3342–3352 (2019)

[24] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)

[25] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)

[26] Tan, M., Le, Q.: Efficientnetv2: Smaller models and faster training. In: International Conference on Machine Learning, pp. 10096–10106 (2021). PMLR

[27] Mehta, S., Rastegari, M.: Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178 (2021)

[28] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollr, P.: Focal Loss for Dense Object Detection (2018). https://arxiv.org/abs/1708.02002

[29] Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Test-time augmentation with uncertainty estimation for deep learning-based medical image segmentation (2018)

[30] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)

[31] Zhao, D., Zhu, D., Lu, J., Luo, Y., Zhang, G.: Synthetic medical images using f&bgan for improved lung nodules classification by multi-scale vgg16. Symmetry **10**(10) (2018) https://doi.org/10.3390/sym10100519

[32] Xie, Y., Zhang, J., Xia, Y.: Semi-supervised adversarial model for benign–malignant lung nodule classification on chest ct. Medical image analysis **57**, 237–248 (2019)

[33] Halder, A., Chatterjee, S., Dey, D.: Adaptive morphology aided 2-pathway convolutional neural network for lung nodule classification. Biomedical Signal Processing and Control **72**, 103347 (2022) https://doi.org/10.1016/j.bspc.2021.103347

[34] Bushara, A.R., Kumar, V.R.S., Kumar, S.S.: Lcd-capsule network for the detection and classification of lung cancer on computed tomography images. Multimedia Tools and Applications **82**(24), 37573–37592 (2023) https://doi.org/10.1007/s11042-023-14893-1

[35] Gautam, N., Basu, A., Sarkar, R.: Lung cancer detection from thoracic ct scans using an ensemble of deep learning models. Neural Computing and Applications **36**(5), 2459–2477 (2024)

[36] Wilcoxon, F.: In: Kotz, S., Johnson, N.L. (eds.) Individual Comparisons by Ranking Methods, pp. 196–202. Springer, New York, NY (1992). https://doi.org/10.1007/978-1-4612-4380-9_16

[37] Rubin, G.D.: Lung nodule and cancer detection in computed tomography screening. Journal of thoracic imaging **30**(2), 130–138 (2015)

[38] National Lung Screening Trial Research Team: Lung cancer incidence and mortality with extended follow-up in the national lung screening trial. Journal

of Thoracic Oncology **14**(10), 1732–1742 (2019) https://doi.org/10.1016/j.jtho.2019.05.044

[39] Tanner, N.T., Aggarwal, J., Gould, M.K., Kearney, P., Diette, G., Vachani, A., Fang, K.C., Silvestri, G.A.: Management of pulmonary nodules by community pulmonologists: a multicenter observational study. Chest **148**(6), 1405–1414 (2015)

[40] Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847 (2018). IEEE

[41] Kaur, S., Aggarwal, D., Dabas, R., Gupta, D., Gupta, A.: Deep visual explanation for deep learning models in medical image analysis: A survey. IEEE Access **10**, 14033–14056 (2022)

[42] Pinsky, P.F., Gierada, D.S., Black, W., Munden, R., Nath, H., Aberle, D., Kazerooni, E.: Performance of lung-rads in the national lung screening trial: a retrospective assessment. Annals of internal medicine **162**(7), 485–491 (2015)

[43] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)

[44] Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)