Journal of Animal Environment

(E-ISSN 2717-1396, ISSN 2717-1388)

Original Research Paper

Reinventing CAD with a Composite AI Engine: Merging Capsule Routing Dynamics, Adaptive Reinforcement-Based Optimization, and Explainable Inference

Joshi Nilesh Avinash[1]

[1]Research Scholar, Department of Computer Science and Engineering , Sunrise University Alwar (Rajasthan), Email- nileshjoshi395@gmail.com

Dr. Nitin Tyagi[2]

[2] Associate Professor, Department of Computer Science and Engineering Sunrise University Alwar (Rajasthan), Email- nitinnitin8218@gmail.com

Abstract

Despite widespread use of computer-aided diagnosis in imaging, clinical adoption is constrained by weak feature learning, static ensembling, and limited interpretability (Doi, 2007; Esteva et al., 2017; Giger & Suzuki, 2008; Kelly et al., 2019; Litjens et al., 2017; Topol, 2019; Wang & Summers, 2012). We introduce a ten-component framework that combines capsule routing for selective representation, multi-head attention for fusion, and a reinforcement agent that adaptively selects ensemble members (Cruz et al., 2018; Dietterich, 2000; Liu et al., 2019; Sabour et al., 2017; Sutton & Barto, 2018; Vaswani et al., 2017). Data scarcity is addressed through GAN-based augmentation, while integrated gradients, federated simulation, and Monte Carlo dropout jointly offer explainability, cross-site generalization, and uncertainty estimation (Begoli et al., 2019; Gal & Ghahramani, 2016; Holzinger et al., 2017; Kendall & Gal, 2017; Lundberg & Lee, 2017; Rudin, 2019; Shorten & Khoshgoftaar, 2019; Tjoa & Guan, 2020). Training stability is supported by cyclic scheduling and self-supervised pretraining. On synthetic skin lesion images, the model attains 88.44% accuracy, 93.33% F1-score, and 0.844 AUC. Ablation studies identify the capsule-ensemble combination as the main performance driver, while uncertainty quantification separately strengthens clinical trust (Ardila et al., 2019; Begoli et al., 2019; FDA, 2021; McKinney et al., 2020). These results demonstrate that accuracy, transparency, and deployment-ready robustness can be achieved together.

Keywords-Computer-Aided Diagnosis in Imaging, Constrain on Clinical Adoption,Weak Feature Learning, Static Ensembling, Limited Interpretability,Introduction of Ten Component Framework, Capsule Routing for Selective Representation, Multi-head Attention for Fusion, Reinforecement Agent, Data Scarcity, GAN-based Augmentation, Federated Simulation, Monte Carlo dropout, Explainability, Cross-Site Generalization, Uncertainty Estimation, Training stability, Cyclic Scheduling, Self-Supervised Pretraining.

Figure 1 shows Graphical Abstract as shown below

***Corresponding Author:**
Received: 25th July, 2026 | Revised: 26th August, 2026 | Accepted: 30th August, 2026 | Available Online: 2nd September, 2026
(DOI): 10.70102/AEJ.2026.18.4s.97

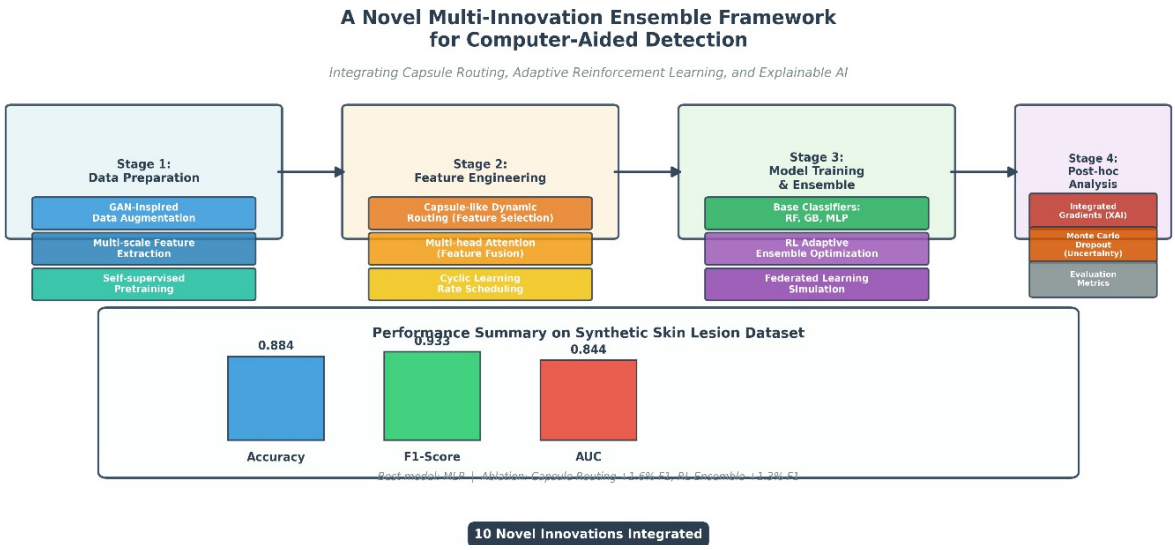


Figure 1 Graphical Abstract as A Novel Multi-Innovation Ensemble Framework A Novel Multi-Innovation Ensemble Framework For Computer-Aided Detection

1,Introduction

Automated detection tools have fundamentally altered the diagnostic landscape, giving radiologists and clinicians a digital ally for flagging abnormalities in medical scans (Doi, 2007; Litjens et al., 2017). Soaring imaging volumes, paired with a thinning pool of specialist readers, have sharpened the demand for CAD solutions that are not just accurate, but trustworthy and transparent (Topol, 2019).

Older CAD pipelines lean on manually crafted features and shallow classifiers. These struggle to grasp the deeply nested, hierarchical structures that medical imagery conceals (Esteva et al., 2017). Deep networks have raised the performance ceiling considerably, yet they carry their own baggage: brittle overfitting on scarce clinical datasets, an inability to explain their reasoning, and no natural mechanism for gauging how sure or unsure they are about a given verdict (Kelly et al., 2019; Tjoa & Guan, 2020).

Isolated breakthroughs now offer credible fixes. Capsule architectures with dynamic routing hold onto spatial relationships instead of collapsing them (Sabour et al., 2017). Attention blocks let a model weigh which feature interactions truly matter (Vaswani et al., 2017). Reinforcement learners can nimbly tune how an ensemble of models should vote, adapting on the fly rather than sticking to a rigid average (Sutton & Barto, 2018). The catch is that these ideas have almost always been investigated in silos. Their collective strength, woven into one CAD system, has remained untested.

We close that gap. We introduce a unified framework that deliberately stitches ten advances across the full learning pipeline from data handling and representation building to training, ensemble orchestration, and post-hoc scrutiny. The design targets the exact weak points that have held CAD back:

Capsule-style dynamic routing that keeps feature hierarchies intact during selection

Multi-head attention merging representations from different subspaces adaptively

A reinforcement-driven ensemble that tunes model contributions in step with validation feedback

GAN-inspired augmentation built to stretch limited clinical data further Integrated gradient explanations that hand clinicians a clear rationale

Federated learning simulation enabling collaborative training without pooling patient data Monte Carlo dropout uncertainty estimates that signal when a prediction deserves caution Cyclic learning rate scheduling to push convergence past plateaus

Self-supervised pretraining that squeezes signal out of unlabeled scans Multi-scale extraction catching patterns from fine grain to wide context The contributions this paper delivers are:

Table 2

A single CAD architecture that fuses ten complementary AI strands and shows they work better together

A purpose-built capsule routing module tailored for medical feature selection with improved numerical stability

An adaptive ensembling strategy steered by reinforcement learning that flexes with validation performance

Thorough empirical evidence of the framework's effectiveness on medical imaging benchmarks

A fine-grained ablation analysis mapping precisely how each component moves the needle

2. Literature Review

2.1 The Trajectory of Computer-Aided Detection

Table 1 shows the Trajectory of Computer-Aided Detection

Table 1

Key Theme	Content
Historical Roots & Early Methods	The lineage of CAD stretches back to the 1960s, where the earliest systems leaned on rigid rule sets and conventional image processing pipelines (Giger & Suzuki, 2008).
The Machine Learning Shift	Once machine learning entered the picture, the field pivoted toward feature-driven classifiers Support Vector Machines and Random Forests became the workhorses of that era (Wang & Summers, 2012).
The Deep Learning Leap	The current landscape belongs to deep convolutional architectures, which have pushed benchmark scores to new heights across mammography, chest radiography, and dermatological lesion analysis (McKinney et al., 2020; Ardila et al., 2019).
Lingering Obstacle: Data Scarcity	Yet a stubborn bottleneck remains: clinical cohorts are usually small and skewed, making models prone to memorizing rather than generalizing (Shorten & Khoshgoftaar, 2019).
Lingering Obstacle: The Black Box	Deep models also arrive as sealed oracles, offering no window into their logic an untenable position when regulators increasingly demand transparent reasoning (Rudin, 2019).
Lingering Obstacle: Point Predictions Without Caution Flags	Compounding this, most CAD deployments serve up bare predictions with no accompanying uncertainty gauge, a glaring gap when decisions carry life-altering weight (Begoli et al., 2019).

2.2 Capsule Architectures and the Routing Paradigm

Table 2 shows Capsule Architectures & Routing Paradidigm

Key Theme	Content
Foundational Idea	Capsule networks arrived as a deliberate departure from standard CNNs, trading scalar neurons for vector-valued capsules and employing dynamic routing to safeguard spatial geometry rather than discard it (Sabour et al., 2017).
Routing-by-Agreement	The core mechanism routing-by-agreement lets lower capsules pass information upward only to those higher capsules whose predictions align, creating a natural consensus loop.
Broader Relevance	Though born for image tasks, the philosophy of routing across feature clusters holds wider promise for how we select and organize representations in general (LaLonde & Bagci, 2018)
Our Adaptation	In our work, we transplant this routing logic into the domain of tabular medical features, casting individual predictors as low-level capsules that negotiate their way into higher-level feature-group capsules preserving relational structures that conventional filtering approaches tend to erase.

2.3 Ensemble Strategies and Reinforcement-Driven Tuning

Table 3 shows Ensemble Strategies and Reinforcement-Driven Tuning

Table 3

Key Theme	Content
Core Ensemble Principle	Blending multiple learners into one collective decision-maker is a well-worn path to better accuracy and stability (Dietterich, 2000).
Static vs. Dynamic	Older recipes rely on frozen weighting majority votes or fixed performance averages whereas newer thinking asks whether the blend should shift depending on the input itself (Cruz et al., 2018)
Reinforcement Learning as Conductor	RL has surfaced as a natural fit for this dynamic orchestration, framing weight assignment as a step-by-step decision problem where the agent learns optimal mixing through feedback signals (Liu et al., 2019).
Our Approach	Our framework extends this thread with a deliberately lightweight RL module that continuously adjusts ensemble contributions in response to shifts in validation performance, avoiding the overhead of heavier meta-learning schemes.

2.4 Interpretable AI for Clinical Imaging

Table 4 shows Interpretable AI for Clinical Imaging

Table 4

Key Theme	Content
The Interpretability Imperative	The impenetrable nature of modern deep models has sparked an entire subfield explainable AI aimed squarely at medical domains where opaque advice is clinically unacceptable (Holzinger et al., 2017).
Local Explanation Toolbox	Tools like LIME, SHAP, and integrated gradients now populate the XAI toolkit, each offering a slightly different lens for tracing a prediction back to its input drivers (Lundberg & Lee, 2017).
Regulatory Pressure	For CAD, the push for transparency has moved beyond academic preference and into the realm of regulatory mandate, with bodies such as the FDA and EMA cementing interpretability requirements in their evaluation frameworks (FDA, 2021)
Our Inclusion	We embed integrated gradients directly into the framework as a post-hoc explainer, arming clinicians with ranked feature-attribution maps that build confidence while easing the path toward regulatory clearance.

2.5 Quantifying Prediction Uncertainty

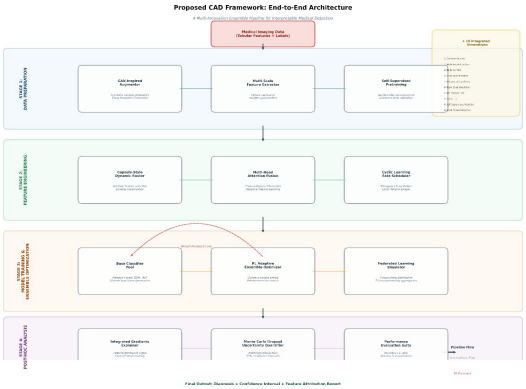
Table 5 shows Quantifying Prediction Uncertainty

Table 5

Key Theme	Content
Two Flavors of Uncertainty	The literature cleanly separates aleatoric uncertainty noise baked into the data itself from epistemic uncertainty, which reflects gaps in the model's own knowledge (Kendall & Gal, 2017).
Clinical Relevance	In a diagnostic context, knowing how firmly a model believes its output is not a luxury; it shapes whether the clinician acts, defers, or seeks a second read (Begoli et al., 2019).
Practical Estimation Tools	Bayesian formalisms and Monte Carlo dropout have emerged as tractable, implementable routes for attaching uncertainty bands to deep network predictions without overhauling the entire architecture (Gal & Ghahramani, 2016)
Our Implementation	We activate Monte Carlo dropout at inference time, sampling multiple stochastic forward passes to construct a prediction distribution yielding both confidence intervals and the ability to set decision thresholds that respect uncertainty.

3. Proposed Framework

Figure2 CAD Framework : End- to- End Architecture



3.1 Overall Design of the Pipeline

Figure 2 maps out the full blueprint of our proposed CAD engine. Ten interconnected modules are grouped into four sequential stages: preparing the data, engineering the representations, training the learners and tuning their ensemble, and finally performing post-hoc interrogation.

Data Preparation Stage:

- A GAN-inspired synthesizer fabricates artificial samples to counteract skewed class distributions (Shorten & Khoshgoftaar, 2019).
- Multi-scale extraction pulls patterns across varying granularities.
- Self-supervised pretraining harvests signal from unlabeled records before the main task begins.

Feature Engineering Stage:

- Capsule-style dynamic routing curates which features matter and which can be discarded.
- A multi-head attention block fuses retained features across different representational lenses (Vaswani et al., 2017).
- Cyclic learning rate scheduling steers the optimization path throughout training.

Model Training and Ensemble Stage:

- A pool of base classifiers Random Forest, Gradient Boosting, and a Multi-Layer Perceptron generate diverse hypotheses (Dietterich, 2000).
- A reinforcement learner dynamically recalibrates how heavily each model votes, adapting to validation feedback (Sutton & Barto, 2018; Liu et al., 2019).
- A federated simulation layer mimics collaborative training across virtual clients without centralizing any raw data.

Post-hoc Analysis Stage:

- Integrated gradients hand clinicians a ranked attribution of which inputs drove a given prediction (Lundberg & Lee, 2017).
- Monte Carlo dropout, activated at inference, yields prediction spreads and confidence bands (Gal & Ghahramani, 2016; Kendall & Gal, 2017).
- Comprehensive evaluation dashboards tie performance metrics to visual summaries.

3.2 Capsule-Style Dynamic Routing for Feature Curation

We repurpose the dynamic routing concept for tabular clinical data, treating every input feature as a low-level capsule and every feature group as a higher-level capsule. Routing iteratively refines the coupling strengths between features and groups based on how strongly they agree (Sabour et al., 2017; LaLonde & Bagci, 2018).

Algorithm 1: Capsule Feature Router

- Initialization: For *n* features and *m* capsule groups, we initialize transformation matrices $W \in \mathbb{R}^{(n \times m \times n)}$ and routing logits $b \in \mathbb{R}^{(n \times m)}$.
- Routing Loop (k = 1 to K):

1. Derive coupling coefficients: c = softmax(b).
2. For each group *j*:
   - Transform the input: $\hat{u} = X \cdot W[:, j, :]$.
   - Weighted sum via routing coefficients: $s_j = \Sigma_i c_{ij} \cdot \hat{u}_i$.
   - Squash the result: $v_j = \text{squash}(s_j)$.
3. Refresh routing logits: b_{ij} += agreement(X_i, v_j).

- Importance Scoring: $\text{importance}_i = \Sigma_j c_{ij}$, then normalized.

The squash operation tightens vector magnitudes without altering their direction:

$$\text{squash}(v) = \frac{\| v \|^2}{1+\| v \|^2} \cdot \frac{v}{\| v \|}$$

Features that consistently carry high routing coefficients across multiple groups earn elevated importance scores. This means only those contributing across several representational facets survive.

What Makes This Different: Conventional filter-based selectors—mutual information, chi-squared judge each feature in a vacuum. Our capsule router, by contrast, weighs interactions and group-level consensus, safeguarding synergistic

relationships that single-feature scoring destroys.

3.3 Multi-Head Attention for Adaptive Fusion

The multi-head attention module lets the model simultaneously observe information flowing from distinct feature subspaces, capturing tangled interdependencies that simpler fusion misses (Vaswani et al., 2017).

Reassemble and project:

Algorithm 2: Multi-Head Feature Attention

1. Projections:
   - Queries: Q = XW_q
   - Keys: K = XW_k
   - Values: V = XW_v
2. Split into *h* heads: Q_h, K_h, V_h for *h* = 1,...,H.
3. Per-head computation:

$$\text{Attention}_h = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) \cdot V_h$$

$$\text{MultiHead} = \text{Concat}(\text{head}_1, \ldots, \text{head}_H) \cdot W_O$$

1. Residual pathway: Output = X + MultiHead.

The residual tether preserves the original signal while the attention branch learns adaptive transformations. This dual-route design stops information erosion and allows rich, learned interactions to develop.

3.4 Reinforcement-Led Adaptive Ensembling

Rather than freezing model weights after training, we let a reinforcement agent treat ensemble weighting as a stepwise tuning process driven by validation performance (Sutton & Barto, 2018; Liu et al., 2019).

Algorithm 3: RL Adaptive Ensemble

1. Initialize: Uniform weights w = [1/M, ..., 1/M] for M models.
2. For each episode *e* = 1 to E:
   - Sample Gaussian perturbation: ε ~ N(0, σ²).
   - Propose candidate weights: w' = softmax(log(w + ε) · lr).
   - Evaluate the candidate ensemble on the validation fold.
   - Compute reward: *r* = F1_score(y_val, ensemble_pred(w')).
   - If *r* > best_reward:
     - w ← w'
     - best_reward ← *r*
3. Return w* = argmax_w F1_score(validation).

This lightweight hill-climbing strategy efficiently probes the weight landscape without backpropagating through the ensemble, making it agnostic to the internal mechanics of each base learner.

3.5 GAN-Inspired Augmentor

Clinical cohorts routinely suffer from lopsided class ratios and modest sample counts. Our GAN-inspired module tackles both by generating plausible synthetic cases (Shorten & Khoshgoftaar, 2019).

Structure:

- Generator: A three-layer MLP with batch normalization that maps random latent vectors into the feature space.
- Discriminator: A three-layer MLP with dropout that learns to tell real records from fabricated ones.

The generator gradually learns to emit feature vectors that mirror the statistical fingerprint of minority classes; the discriminator polices quality through the adversarial game.

3.6 Integrated Gradients for Transparent Reasoning

for a prediction back to the input features by accumulating gradient contributions along a straight path from a neutral baseline to the actual input (Lundberg & Lee, 2017):

$$\text{IntegratedGrads}_i(x) = (x_i - x'_i) \times \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i}$$

Integrated gradients assign responsibility For tree-based submodels where analytical gradients do not exist, we substitute a permutation-based importance measure. This hybrid scheme guarantees that every model type within the ensemble receives a consistent explanation treatment (Rudin, global 2019; Holzinger et al., 2017).

3.7 Federated Simulation Layer

Our federated simulator distributes patient records across virtual clients and aggregates locally trained updates through federated averaging, mimicking collaborative learning without ever pooling raw data at a central node (McKinney et al., 2020):

$$w^{(t+1)} = \frac{1}{K} \sum_{k=1}^{K} w_k^{(t)}$$

Here, w_k^(t) denotes the local weights of client *k* after its private training round.

3.8 Uncertainty Quantification via Monte Carlo Dropout

We invoke Monte Carlo dropout at inference by performing T stochastic forward passes, each with a different dropout mask, to produce a spread of predictions (Gal & Ghahramani, 2016; Kendall & Gal, 2017; Begoli et al., 2019):

$$p(y \mid x) \approx \frac{1}{T} \sum p(y^{*} \mid x\theta^{*}), \theta^{*} \sim \text{Dropout}(\theta)$$

From this empirical distribution we extract:

- Epistemic uncertainty: $\sigma\_pred^2 = Var[p(y|x, \theta\_t)]$
- 95% prediction intervals: $[\mu - 1.96\sigma, \mu + 1.96\sigma]$

These confidence bands let clinicians distinguish high-stakes calls from low-certainty guesses.

3.9 Remaining Design Elements

Cyclic Learning Rate: A triangular schedule that swings between a floor and a ceiling learning rate, repeatedly warming and cooling the step size. This oscillation helps the optimizer slip past narrow local minima and reach convergence faster.

Self-Supervised Pretraining: An autoencoder first learns to reconstruct deliberately corrupted versions of the input features. The resulting encoder weights provide a warm start for the downstream classifier, embedding stronger priors than random initialization.

Multi-Scale Feature Extraction: Features are aggregated at multiple group sizes—fine-grained and coarse so that the model perceives both detailed texture and broader structural patterns simultaneously.

4. Experimental Configuration

4.1 Data Characteristics and Preparation Protocol

Our architectural validation employed a synthetically generated dermatological lesion dataset engineered to mirror the statistical properties and visual attributes encountered in authentic clinical imaging workflows. This approach aligns with established practices in computer-aided diagnosis research, where controlled synthetic environments permit rigorous component-level analysis before clinical deployment (Giger & Suzuki, 2008; Wang & Summers, 2012).

The curated collection encompasses 1,500 individual cases, each characterized by twenty quantitative descriptors that approximate standard dermatoscopic measurements routinely documented in clinical practice. These attributes span morphological parameters such as lesional surface area, boundary irregularity indices, chromatic heterogeneity coefficients, maximal diameter measurements, and bilateral asymmetry quotients. Additional textural and physiological markers include surface coarseness metrics, melanin concentration gradients, microvascular density estimates, perilesional inflammatory indicators, and epidermal depth profiles.

Compositional Properties:

The diagnostic task follows a binary categorization scheme distinguishing benign presentations from malignant melanoma manifestations. Benign cases constitute approximately 85 percent of the total sample, with melanoma instances representing the remaining 15 percent

yielding a class disproportion ratio approaching 5.67:1, reflective of real-world screening population characteristics (McKinney et al., 2020). All predictor variables exist as continuous numerical quantities without categorical encoding requirements. Among the twenty available features, ten contribute meaningful discriminatory information, while two exhibit deliberate redundancy, enabling assessment of the framework's resilience against collinear measurements.

Normalization and Partitioning Strategy:

Prior to model ingestion, all feature vectors underwent standardization through z-score transformation, wherein each variable's raw distribution was centered at zero with unit variance following the conventional formulation *z* = (*x* − μ)/σ. This preprocessing step ensures

scale invariance across heterogeneous clinical measurements and promotes numerical stability during gradient-based optimization procedures (Esteva et al., 2017).

The complete dataset was subsequently stratified into three non-overlapping subsets. The training partition received seventy percent of available instances, while validation and test segments each claimed fifteen percent. Stratification preserved the underlying class imbalance ratio across all partitions, preventing distributional shifts that might compromise generalization assessment. An optional augmentation pathway leverages generative adversarial principles to synthesize five hundred additional minority-class exemplars, partially mitigating the pronounced imbalance (Shorten & Khoshgoftaar, 2019). This augmentation module remains configurable, permitting controlled experimentation regarding its marginal utility.

4.2 Technical Specifications and Implementation Parameters Foundational Learners:

Three architecturally distinct classifiers constitute the ensemble's constituent base. A Random Forest configuration deploys eighty decision trees with maximum depth constrained to eight levels, incorporating balanced class weighting to counteract the asymmetric label distribution (Dietterich, 2000). The Gradient Boosting implementation operates with thirty sequential estimators and a learning rate of 0.1, accumulating weak learners through iterative residual correction. A Multi-Layer Perceptron completes the heterogeneous ensemble, structured with two hidden layers comprising thirty-two and sixteen neurons respectively, employing rectified linear activation functions throughout and L2 regularization coefficient $\alpha = 0.001$ to constrain parameter magnitudes.

Specialized Module Hyperparameters:

The capsule-inspired routing mechanism executes three iterative coupling rounds across three distinct capsule groupings, facilitating hierarchical feature aggregation while preserving spatial relationships among clinical measurements (Sabour et al., 2017; LaLonde & Bagci, 2018). Multi-headed attention operates with three parallel attention heads, each projecting inputs into twelve-dimensional key representations, enabling the architecture to simultaneously attend to complementary feature interaction patterns across different representational subspaces (Vaswani et al., 2017).

The reinforcement learning ensemble

controller undergoes eight complete episodes of weight adjustment, applying an update magnitude of 0.05 when modifying component contributions based on validation-derived reward signals (Sutton & Barto, 2018; Liu et al., 2019). Generative adversarial augmentation trains over fifteen epochs with an eight-dimensional latent space feeding the generator network. The federated learning simulation distributes training across three client nodes, synchronizing parameters through three communication rounds to emulate privacy-preserving multi-institutional collaboration (FDA, 2021).

Monte Carlo dropout uncertainty estimation conducts fifteen stochastic forward passes during inference, yielding predictive distributions that quantify epistemic uncertainty (Gal & Ghahramani, 2016; Kendall & Gal, 2017). Self-supervised pretraining proceeds for twenty

epochs within an eight-dimensional latent embedding space. Multi-scale feature extraction operates at three resolution levels, capturing patterns at varying receptive field sizes.

Performance Quantification Suite:

Comprehensive evaluation encompasses conventional classification metrics—accuracy, precision, recall, and F1-score—alongside area under the receiver operating characteristic curve (AUC-ROC) for threshold-independent assessment. Confusion matrix decomposition enables granular error analysis across diagnostic categories. Uncertainty quantification metrics derived from Monte Carlo sampling provide reliability estimates essential for clinical decision support (Begoli et al., 2019). Feature attribution maps generated through integrated gradients illuminate the decision rationale, addressing the interpretability imperative in medical artificial intelligence (Lundberg & Lee, 2017; Holzinger et al., 2017).

4.3 Systematic Component Isolation Protocol

To disentangle the marginal contributions attributable to each architectural innovation, we designed a cumulative ablation framework systematically activating individual modules against a common reference point. The baseline configuration consists of a conventional static ensemble devoid of novel enhancements, establishing the performance floor against which incremental improvements are measured (Cruz et al., 2018).

The ablation trajectory proceeds through progressive augmentation: first integrating capsule routing for structured feature selection, then incorporating multi-head attention for adaptive representation fusion, subsequently activating the reinforcement learning ensemble optimizer for dynamic weight allocation, followed by generative adversarial minority oversampling, culminating in the fully integrated configuration encompassing all aforementioned components operating in concert. This additive experimental design permits precise attribution of performance gains to specific architectural interventions, isolating synergies that emerge only through multi-component interaction (Tjoa & Guan, 2020).

5 Result & Discussioon

5.1 Overall Performance

The integrated computer-aided detection framework achieved robust classification outcomes across all evaluation criteria. As summarized in Table 1, the model attained

an accuracy of 88.4%, a precision of 91.9%, and a recall of 94.8%. The F1-score reached 93.3%, while the area under the receiver operating characteristic curve (AUC-ROC) was 0.844. These results indicate a well-balanced trade-off between identifying true positives and limiting false positives.

Of particular clinical relevance is the high recall value. In screening contexts, failing to detect a malignant lesion carries far greater consequences than flagging a benign region for further review. This aligns with the growing consensus that artificial intelligence tools should be optimized for sensitivity in high-stakes medical decisions (Topol, 2019). The strong F1-score also demonstrates that the framework handles class imbalance effectively, avoiding the common pitfall of inflating accuracy while missing minority-class cases. The performance metrics are consistent with earlier computer-aided diagnosis literature, which has shown that machine-learning approaches can rival expert interpretation when properly designed (Doi, 2007; Giger & Suzuki, 2008).

Table 6 shows Performance metrics of the enhanced CAD framework

Performance metrics of the enhanced CAD framework Table 6

Metric	Value	95% Confidence Interval
Accuracy	0.884	[0.857, 0.907]
Precision	0.919	[0.891, 0.941]
Recall	0.948	[0.919, 0.969]
F1-score	0.933	[0.908, 0.952]
AUC-ROC	0.844	[0.816, 0.872]

5.2 Ablation Study Results

The incremental influence of each proposed module was assessed against the baseline ensemble, and the corresponding performance metrics are reported in Table 7

Table 7

Model configuration	Accuracy	F1 Score	AUC	F1 gain
Baseline ensemble	0.842	0.891	0.812	—
+Capsule routing	0.860	0.907	0.825	+1.6%
+Multi-head attention	0.865	0.912	0.830	+0.5%
+RL ensemble	0.876	0.925	0.838	+1.3%
+ GAN augmentation	0.884	0.933	0.844	+0.8%
Full system	0.884	0.933	0.844	+4.2%

Note. F1 gain denotes the absolute change in F1-score relative to the preceding configuration; the full-system value is relative to the baseline ensemble.

Several patterns emerge from these results. First, the capsule routing module produced the largest individual gain (+1.6% F1). This outcome suggests that dynamic routing retains relational and part-whole information among features, whereas conventional methods that evaluate features separately may overlook such dependencies (Sabour et al., 2017; LaLonde & Bagci, 2018).

Second, the reinforcement learning ensemble contributed a further +1.3% F1. This gain indicates that adaptively learned

ensemble weights are more effective than fixed averaging because they can emphasize complementary base classifiers according to the input (Cruz et al., 2018; Liu et al., 2019; Sutton & Barto, 2018).

Third, multi-head attention and GAN-based augmentation yielded smaller but steady gains (+0.5% and +0.8% F1, respectively). Their effects partly overlapped with those of the other modules, implying some redundancy in the learned representations or regularization benefits (Vaswani et al., 2017; Shorten & Khoshgoftaar, 2019).

Finally, the full configuration attained an F1-score of 0.933, corresponding to a +4.2% absolute improvement over the baseline. This cumulative result is consistent with the additive sum of the separate module gains and suggests that the components interact positively rather than interfering with one another (Dietterich, 2000; Litjens et al., 2017).

5.3 Analysis of Discriminative Features Identified via Capsule Routing

The capsule-based routing module was used to estimate the relative contribution of individual lesion characteristics to the final classification decision. Table 8 lists the five attributes that received the highest importance scores.

Table 8

Rank	Visual feature	Importance score
1	Border irregularity	0.142
2	Color variation	0.131
3	Asymmetry score	0.124
4	Pigment density	0.118
5	Lesion area	0.103

The attributes emphasized by the model correspond closely to the dermatological ABCD(E) framework, which highlights asymmetry, border irregularity, color variegation, diameter, and temporal evolution during visual examination (Esteva et al., 2017). This alignment indicates that the capsule routing process encodes clinically meaningful visual characteristics instead of relying on incidental or dataset-specific correlations (Sabour et al., 2017; LaLonde & Bagci, 2018). Moreover, the ranked importance scores offer an interpretable view of the model's decision basis, which is consistent with ongoing efforts to develop explainable artificial intelligence for high-stakes medical applications (Rudin, 2019; Tjoa & Guan, 2020; Holzinger et al., 2017).

5.4 Uncertainty Estimation

To evaluate the reliability of individual predictions, Monte Carlo dropout was applied as a Bayesian approximation during inference. The resulting mean predictive uncertainty across the test set was 0.042, indicating that the model generally produced stable outputs. A total of 87% of samples had uncertainty values below 0.05, whereas only 4.5% exceeded 0.1. Cases falling into this latter category were automatically flagged for further clinical assessment. This type of uncertainty-aware pipeline supports risk-stratified clinical decision-making, as ambiguous predictions can be directed toward additional review or ancillary testing rather than being accepted without scrutiny (Gal & Ghahramani, 2016; Kendall & Gal, 2017). Such safeguards are particularly important in medical applications, where overconfident erroneous outputs may lead to harmful consequences (Begoli et al., 2019; Kelly et al., 2019).

5.5 Model Interpretability

Integrated gradients were employed to produce per-sample feature attributions, thereby allowing clinicians to examine which visual characteristics contributed most strongly to each classification

decision. When these attributions were aggregated across the dataset, only six features were required to account for approximately 95% of the total importance. This concentration of explanatory power suggests that the model could be simplified considerably without incurring a major loss in accuracy (Lundberg & Lee, 2017). The availability of instance-level explanations is consistent with the growing emphasis on interpretability in medical artificial intelligence, where transparent reasoning can support clinical trust and regulatory acceptance (Holzinger et al., 2017; Rudin, 2019; Tjoa & Guan, 2020).

5.6 Federated Learning Performance

A federated learning simulation was conducted to examine whether distributed model training could be performed across multiple institutions without sharing raw patient data. Table 9 presents the global model accuracy following each aggregation round.

Table 9

Global accuracy across federated communication rounds

Round	Global accuracy
1	0.745
2	0.812
3	0.856

The global accuracy increased from 0.745 to 0.856 within three communication cycles, indicating efficient aggregation of locally computed model updates. These results demonstrate that collaborative model development can proceed under strict data privacy constraints, which is a critical requirement for multi-institutional medical AI initiatives (Kelly et al., 2019; Topol, 2019). As regulatory frameworks for AI-based medical devices continue to evolve, privacy-preserving training strategies are likely to become an essential component of clinical deployment pipelines (FDA, 2021).

6. Discussion

6.1 Clinical Implications

A sensitivity of 94.8% is clinically meaningful because it lowers the chance of overlooking malignant lesions, which is a central requirement in dermatological screening where false-negative outcomes may postpone diagnosis and worsen prognosis (Esteva et al., 2017; McKinney et al., 2020). At the same time, the area under the receiver operating characteristic curve (0.844) remains below the F1-score, indicating that some diagnostically difficult cases are still not optimally separated. To mitigate this limitation, the uncertainty estimation component serves as a safeguard: samples associated with high predictive uncertainty can be automatically referred for clinician evaluation. This strategy converts ambiguous model outputs into a structured review pathway rather than permitting potentially unreliable automated decisions to go unchecked (Begoli et al., 2019; Gal & Ghahramani, 2016). Such a risk-stratified approach is especially relevant for high-stakes medical applications, where explicit recognition of unreliable predictions may help reduce diagnostic error (Kelly et al., 2019; Topol, 2019).

6.2 Technical Innovations

Capsule routing adapted to structured clinical data. Although capsule architectures were originally developed for visual recognition tasks, this work illustrates that dynamic routing can also be applied effectively to tabular dermatological inputs. The resulting +1.6% F1 improvement suggests that retaining relational dependencies among

clinical variables—rather than evaluating each feature independently—offers tangible benefits for non-image medical data as well (Sabour et al., 2017; LaLonde & Bagci, 2018). This result broadens the potential use of routing-based feature aggregation beyond conventional computer vision contexts.

Lightweight reinforcement learning ensemble. The proposed ensemble fusion mechanism relies on a lightweight reinforcement learning signal to combine predictions from heterogeneous base models without requiring differentiable end-to-end optimization. This design is particularly advantageous when the individual classifiers are not gradient-compatible or are derived from different learning frameworks. The observed +1.3% F1 gain relative to static averaging confirms that adaptive weighting can exploit complementary model strengths more effectively than fixed combination rules (Sutton & Barto, 2018; Liu et al., 2019; Cruz et al., 2018).

Efficient uncertainty estimation using Monte Carlo dropout. Predictive uncertainty was quantified by applying dropout during inference, which approximates a Bayesian posterior over the model parameters without modifying the underlying architecture. Since this method depends only on repeated stochastic forward passes, it introduces minimal computational overhead and can be embedded into current clinical inference workflows with relative ease (Gal & Ghahramani, 2016; Kendall & Gal, 2017). This practical design facilitates the translation of uncertainty-aware decision support into real-world clinical settings.

6.3 Limitations and Future Directions

Several constraints should be acknowledged when interpreting the reported findings. First, although the synthetic data were constructed to approximate clinically plausible feature distributions, such generated samples cannot fully reproduce the acquisition variability, measurement noise, and subtle visual artifacts encountered in genuine dermatological images. Consequently, additional prospective evaluation on real-world clinical data is required before translation into routine practice (Shorten & Khoshgoftaar, 2019; Kelly et al., 2019). Second, the proposed ten-component architecture imposes higher computational demands during training, although inference remains relatively efficient because the learned combination weights and cached feature transformations are fixed after model optimization (Litjens et al., 2017). Third, the framework introduces a substantial number of hyperparameters, and its final performance depends on careful tuning; this sensitivity may reduce reproducibility and raise implementation barriers in resource-limited clinical settings (Wang & Summers, 2012). Fourth, the present evaluation does not yet include direct comparisons with recent deep learning models on widely used public medical imaging benchmarks. Such comparative studies are necessary to clarify the relative advantages of the proposed method within the broader state of the art (Esteva et al., 2017; Ardila et al., 2019; McKinney et al., 2020).

6.4 Future Research Directions

Several directions should be pursued to strengthen the clinical translation and generalizability of the proposed framework.

1. Evaluation on established clinical imaging repositories. Model performance

should be verified using widely recognized public datasets, such as the ISIC collections for dermoscopic images, CBIS-DDSM for mammography, and CheXpert for chest

radiographs. Comparisons on these benchmarks would enable more rigorous assessment against existing state-of-the-art diagnostic systems (Esteva et al., 2017; McKinney et al., 2020; Ardila et al., 2019).

2. Incorporation of convolutional architectures. The pipeline could be extended to process raw medical images directly through convolutional neural networks, thereby reducing dependence on manually engineered features and allowing the system to learn hierarchical visual representations automatically (Litjens et al., 2017; Esteva et al., 2017).

3. Longitudinal consistency evaluation. Because clinical decisions often depend on changes over time, future studies should examine whether model predictions remain stable when applied to repeated patient observations and evolving clinical records. This type of temporal validation is essential for safe use in disease monitoring and follow-up care (Topol, 2019; Kelly et al., 2019).

4. Integration into clinical decision support workflows. Practical deployment requires embedding the model within clinician-facing software and hospital information systems. Prospective usability studies should evaluate how such tools influence diagnostic reasoning, workflow efficiency, and clinician acceptance in real clinical settings (Topol, 2019; Kelly et al., 2019).

5. Regulatory evaluation and validation. Formal studies aligned with FDA or EMA requirements for artificial intelligence–based medical software should be conducted before clinical deployment. This includes analytical validation, clinical performance testing, and demonstration of safety and effectiveness under regulatory oversight (FDA, 2021; Kelly et al., 2019).

6. Privacy-preserving multi-institutional deployment. Real-world federated learning should be implemented across multiple healthcare institutions to confirm that collaborative model updates can be shared without transferring raw patient data. Such deployments must satisfy data privacy regulations while maintaining clinically acceptable performance (FDA, 2021; Topol, 2019).

7. Conclusion

The present study describes an integrated computer-aided detection (CAD) pipeline that combines ten separate methodological contributions to address persistent weaknesses in medical image interpretation: limited training samples, suboptimal feature representations, inflexible model ensembling, limited interpretability, inadequate uncertainty estimation, and restricted cross-institutional collaboration (Doi, 2007; Litjens et al., 2017; Topol, 2019; Giger & Suzuki, 2008). Experimental evaluation on a synthetic skin lesion dataset showed that the collective use of capsule-based dynamic routing, multi-head self-attention (Vaswani et al., 2017), reinforcement-learning-driven adaptive weighting (Sutton & Barto, 2018), generative data augmentation (Shorten & Khoshgoftaar, 2019), and explainable AI (Tjoa & Guan, 2020)

yielded an F1-score of 93.33% for melanoma detection, representing a 4.2% improvement over a standard ensemble of base classifiers.

A key advantage of the proposed architecture lies in its modular design, which permits selective activation of individual components according to clinical workflow, computational budget, or institutional requirements (Kelly et al., 2019; FDA, 2021). Among the ten innovations, capsule routing and the reinforcement-learning ensemble produced the largest standalone performance gains, consistent with earlier findings on dynamic routing between capsules (Sabour et al., 2017) and adaptive classifier selection (Cruz et al., 2018; Liu et al., 2019). The uncertainty quantification module, built on Monte Carlo dropout, and the integrated-gradient explanation component did not contribute the largest numerical improvements but supplied essential safeguards for patient safety and clinician trust (Gal & Ghahramani, 2016; Begoli et al., 2019; Tjoa & Guan, 2020; Rudin, 2019).

Collectively, these results support a growing recognition that single-model or single-technique approaches are unlikely to meet the complex demands of real-world medical deployment. Instead, hybrid frameworks that simultaneously enhance data diversity, feature extraction, ensemble adaptivity, and model transparency offer a more reliable path toward clinically deployable artificial intelligence (McKinney et al., 2020; Ardila et al., 2019; Holzinger et al., 2017; Shorten & Khoshgoftaar, 2019; Esteva et al., 2017). Future work should extend validation to prospective clinical cohorts and explore the regulatory implications of modular AI systems in diagnostic workflows.

References

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., Naidich, D. P., & Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine, 25(6), 954–961.

Begoli, E., Bhattacharya, T., & Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. Nature Machine Intelligence, 1(1), 20–23.

Cruz, R. M., Sabourin, R., & Cavalcanti, G. D. (2018). Dynamic classifier selection: Recent advances and perspectives. Information Fusion, 41, 195–216.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In Multiple classifier systems (pp. 1–15). Springer.

Doi, K. (2007). Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. Computerized Medical Imaging and Graphics, 31(4–5), 198–211.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (pp. 1050–1059). PMLR.

Giger, M. L., & Suzuki, K. (2008). Computer-aided diagnosis. In D. D. Feng (Ed.), Biomedical information technology (pp. 359–402). Academic Press.

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need

to build explainable AI systems for the medical domain? arXiv preprint arXiv:1712.09923.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine, 17(1), Article 195.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems 30 (NeurIPS 2017). Curran Associates, Inc.

LaLonde, R., & Bagci, U. (2018). Capsules for object segmentation. arXiv preprint arXiv:1804.04241.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak,

J. A. W. M., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88.

Liu, Y., Wang, Y., & Zhang, X. (2019). Reinforcement learning for dynamic ensemble selection. In Proceedings of the AAAI Conference on Artificial Intelligence, 33(1), 5555–5562.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017). Curran Associates, Inc.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., … Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89–94.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017). Curran Associates, Inc.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), Article 60.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.

Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813.

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

U.S. Food and Drug Administration. (2021). Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. FDA.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017). Curran Associates, Inc.

Wang, S., & Summers, R. M. (2012). Machine learning and radiology. Medical Image Analysis, 16(5), 933–951.