RESEARCH ARTICLE

Quantile Deviation Ensemble Based on Multi-Layer Perceptrons for Cervical Cancer Classification with Uncertainty Perception

Hui Hu[1], **Jiewu Huang**[1,*], **Shoumei Cao**[1] **and Anna Dai**[1]

[1] College of Data Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, China

Abstract

The accurate classification of cervical cytology in Pap smear images remains a critical challenge in computer-aided diagnosis, largely due to the inherent uncertainty and subtle morphological variations among different pathological categories. To address this, we propose a novel uncertainty-aware ensemble framework that integrates statistical quantile analysis with deep learning for robust and interpretable classification. Our framework first leverages three deep convolutional neural networks DenseNet121, MobileNetV2, and ResNet-50 as base feature extractors. Instead of employing naive ensemble strategies, we introduce a quantile deviation based weighting mechanism to dynamically assess and integrate the prediction confidence of each model, explicitly quantifying performance bias across different probability quantiles. This approach not only enhances ensemble stability but also provides a statistical measure of model uncertainty. Subsequently, the weighted probabilistic outputs are fed into a multi-layer perceptron (MLP) for further non-linear optimization and decision refinement, forming a hybrid statistical deep learning pipeline. Evaluated on the publicly available SIPaKMeD dataset, our framework achieves an average accuracy of 98.10%, outperforming both individual base models and existing ensemble methods. Visualization via Grad-CAM further confirms that the framework focuses on clinically relevant cellular structures, validating its diagnostic relevance. By bridging statistical uncertainty quantification with deep ensemble learning, this work offers a principled and transparent methodology for medical image classification, with potential extensibility to other domains requiring reliable and interpretable predictions under uncertainty.

Keywords: cervical cytology classification, uncertainty quantification, ensemble learning, quantile deviation, deep learning.

1 Introduction

Cervical cancer poses a major threat to women's health worldwide, ranking among the leading causes of morbidity and mortality among female

Submitted: 14 January 2026
Accepted: 29 January 2026
Published: 07 February 2026

Vol. 2, **No.** 1, 2026.
10.62762/JMIA.2026.403659

***Corresponding author:**
Jiewu Huang
ylgw217@163.com

Citation
Hu, H., Huang, J., Cao, S., & Dai, A. (2026). Quantile Deviation Ensemble Based on Multi-Layer Perceptrons for Cervical Cancer Classification with Uncertainty Perception. *Journal of Mathematics and Interdisciplinary Applications*, 2(1), 12–27.

malignancies. Statistics indicate that over 500,000 new cervical cancer cases and more than 300,000 deaths occur annually worldwide [1]. This alarming situation underscores the critical importance of early screening and diagnosis. Liquid-based cytology (LBC) remains the most widely used screening method, analyzing cervical exfoliated cells morphologically to identify precancerous lesions and early signs of malignancy. However, traditional manual slide review processes have inherent limitations: each slide typically contains thousands of cells [2]. Pathologists must subjectively assess microscopic features such as cell morphology, nuclear-to-cytoplasmic ratio, and staining characteristics under high intensity workloads. This process is not only time consuming and labor-intensive but also prone to diagnostic errors due to physician experience, fatigue, and cognitive biases, potentially leading to missed or misdiagnosed cases [3]. In regions with uneven distribution of medical resources, the shortage of experienced pathologists further exacerbates screening bottlenecks. Consequently, developing efficient, objective, and reproducible computer aided diagnostic systems has become an urgent necessity to enhance cervical cancer screening coverage and diagnostic accuracy.

With the rapid advancement of artificial intelligence technology, particularly breakthroughs in deep learning within computer vision, medical image analysis methods based on convolutional neural networks have opened new avenues for automated cervical cytology diagnosis. Unlike traditional machine learning approaches reliant on manually engineered features, deep learning models can autonomously learn hierarchical feature representations within images, demonstrating exceptional performance across diverse medical image classification tasks [4]. However, cervical cytology classification presents unique complexities: First, significant morphological similarities exist between different pathological categories, with features like nuclear size, shape, and chromatin distribution often exhibiting continuous spectrum variations, blurring category boundaries. Second, factors inherent in smear preparation such as staining variations, cell overlap, and impurity interference further complicate feature extraction and interpretation. Furthermore, deep learning models inherently exhibit an "overconfidence" bias, potentially outputting high confidence probabilities even for incorrect predictions a trait that could have severe consequences in high stakes medical decision-making scenarios. Collectively, these factors generate substantial uncertainty in model predictions, yet traditional classification frameworks often fail to explicitly model and leverage this uncertainty.

To enhance model robustness and generalization capabilities, ensemble learning has become a widely adopted technical strategy in medical image analysis by combining predictions from multiple base learners. Typical methods such as majority voting and weighted averaging effectively reduce variance and bias inherent in individual models. However, most existing ensemble methods implicitly assume that all base learners produce predictions of equal reliability. This overlooks the potential for systematic differences in prediction confidence across models when encountering diverse samples. For instance, one model may excel at recognizing a specific type of morphologically variant cell while performing poorly on another; or when image quality is poor, different models may exhibit varying degrees of uncertainty. Fixed weights or simplistic rules fail to adaptively balance contributions across models. In cases of prediction conflicts, a "majority rules" approach may even amplify errors. Thus, quantifying and modeling uncertainty within each base learner's prediction to enable dynamic, adaptive model fusion becomes crucial for enhancing ensemble system performance and reliability.

To address the above challenges, we propose a novel uncertainty aware ensemble framework that systematically integrates statistical quantile analysis with deep learning for cervical cytology classification. Our main contributions are threefold:

- Uncertainty-Aware Weighting Based on Quantile Deviation: We introduce a quantile deviation metric to measure the prediction error of base models across different probability quantile intervals. This method provides a statistical foundation for dynamically assigning ensemble weights, granting higher weights to models with lower uncertainty.

- Hybrid Statistical Deep Learning : We construct a two stage architecture that further optimizes the uncertainty-weighted probabilities generated by a multi-layer deep convolutional neural network through a multi-layer perceptron, achieving effective integration of statistical inference and nonlinear feature mapping.

- Interpretability: Through experiments on two public datasets and Grad-CAM visualization validation, we demonstrate that this framework not only achieves outstanding classification performance but also focuses model attention on pathology relevant regions, significantly enhancing interpretability for clinical users.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 details the proposed methodology. Section 4 presents experimental results and analysis. Section 5 discusses the implications and limitations, and Section 6 concludes the paper.

2 Related Work

In this section, we will discuss some previously published studies on cervical cancer detection from Pap smear images using deep learning methods. Plissiti et al. [5] constructed the SIPaKMeD dataset and achieved 95.35% accuracy using handcrafted features with an SVM classifier. Subsequent research shifted towards deep feature extraction, such as the CNN-SVM pipeline proposed by Nanni et al. [6].

Feature fusion strategies emerged as a key direction for performance improvement. Rahaman et al. [7] proposed a Hybrid Deep Feature Fusion network, while Basak et al. [8] combined deep learning with evolutionary algorithms for feature optimization. Liu et al. [9] extracted and fused local and global features using Xception and DeiT models, respectively.

To address class imbalance, Newaz et al. [10] developed an intelligent decision support system integrating hybrid resampling with genetic algorithms. For model integration, Ghoneim et al. [11] combined CNN with Extreme Learning Machines, and Tawalbeh et al. [12] fused features from ten pretrained models for SVM classification.

Transfer learning has been widely adopted. Kalbhor et al. [13] compared various pretrained networks with machine learning algorithms, while Attallah [4] focused on lightweight architectures (MobileNet, ResNet-18) with feature fusion and dimensionality reduction for efficient deployment. These studies establish a foundation for applying uncertainty quantification methods in this domain.

Manna et al. [1] introduced a CNN ensemble using fuzzy ranking to combine confidence scores from three CNN architectures. Pramanik et al. [22] developed a fuzzy distance-based ensemble that minimizes error through Euclidean, Manhattan, and cosine distances, with defuzzification via the product rule. Sahoo et al. [2] integrated three ensemble techniques Sugeno fuzzy, ranking-based, and gamma function based ensembles sing weighted filtering that considers each intermediate model's validation performance to fuse final classifications.

Akpinar et al.[14] proposed an uncertainty aware liquid neural network (UA-LNN) that quantifies prediction uncertainty by integrating Monte Carlo Dropout into the network. Han et al. [15] propose DM-CNN, a dynamic multi-scale convolutional neural network for medical image classification. This model integrates a dynamic multi scale feature fusion module (DMFF), hierarchical dynamic uncertainty attention, and a multi-scale fusion pooling method (MF Pooling), employing Monte Carlo Dropout for uncertainty quantification. Ullah et al.[16] proposed SVIS-RULEX, an interpretable artificial intelligence framework for medical image classification that integrates statistics, vision, and rules. This framework combines deep feature based statistical feature engineering, a two-stage feature selection method (ZFMIS), rule extraction using decision trees and RuleFit, and a visualization method based on statistical feature map overlay (SFMOV). Wang et al. [17] proposed CreINNs, a confidence set interval neural network for uncertainty estimation in classification tasks. Building upon the traditional interval neural network architecture, this model captures parameter uncertainty through deterministic intervals for weights and biases. In contrast to computationally intensive uncertainty quantification methods like Bayesian Neural Networks (which require learning weight distributions) and Monte Carlo Dropout (which needs multiple stochastic forward passes), our quantile deviation approach offers a computationally efficient alternative. It operates on deterministic pre-trained models through a lightweight post-hoc analysis, providing granular uncertainty assessment across different confidence levels without modifying network architecture or increasing inference time, while maintaining competitive performance. Techniques such as Bayesian neural networks and Monte Carlo dropout provide uncertainty estimates but are computationally expensive. Our work draws inspiration from quantile regression, a robust statistical method for modeling conditional quantiles, and adapts it to measure prediction uncertainty in an ensemble setting. This offers a computationally efficient and statistically principled alternative for uncertainty aware model fusion.

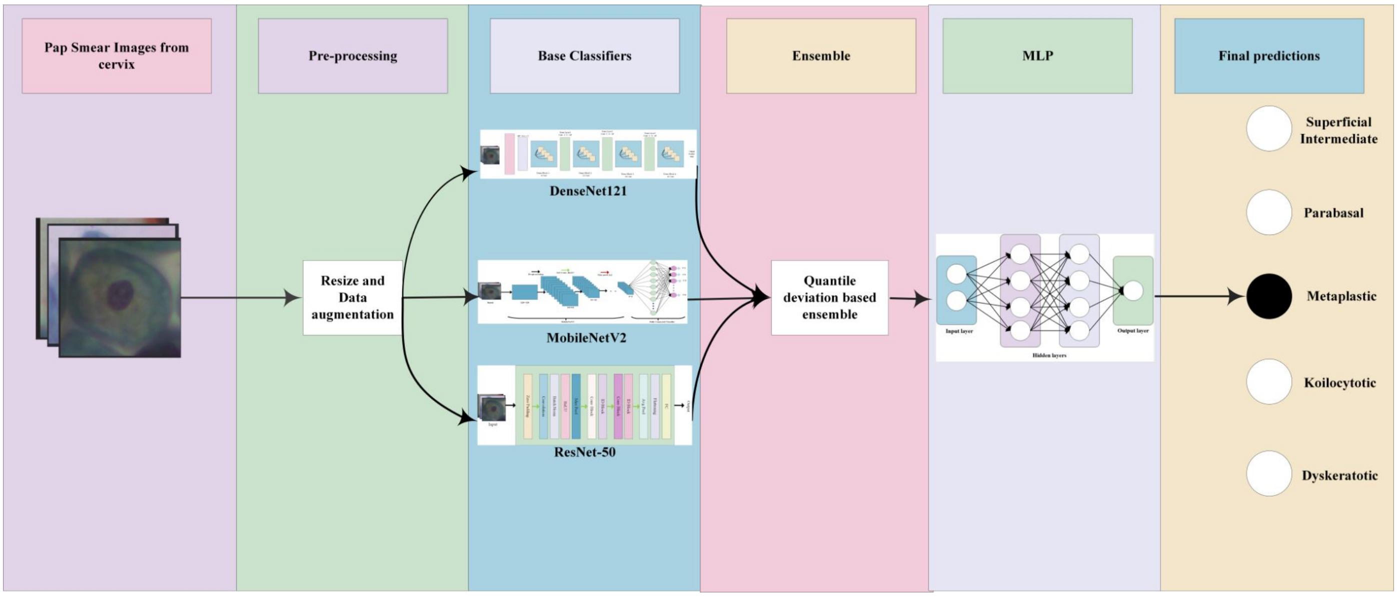


Figure 1. Cervical cancer detection framework based on quantile deviation integration and multi-layer perceptron networks.

3 Methodology

Our proposed framework, illustrated in Figure 1, consists of three main stages: (1) feature extraction using multiple pre-trained CNNs, (2) uncertainty aware ensemble via quantile deviation weighting, and (3) decision refinement using a multi-layer perceptron.

Table 1. Description of dataset categories.

Type	Category	Number
Dyskeratotic	Abnormal	813
Koilocytotic	Abnormal	825
Metaplastic	Benign	793
Parabasal	Normal	787
Superficial-Intermediate	Normal	831

3.1 Dataset description

To evaluate the proposed cervical cancer detection framework based on quantile deviation ensemble and multi-layer perceptron networks, experiments were conducted using the SIPaKMeD dataset. This dataset covers five categories: Dyskeratotic, Koilocytotic, Metaplastic, Parabasal, and Superficial-Intermediate, as shown in Table 1. It consists of 4,049 images cropped from 966 clusters of cell slices, with examples of Pap smear images shown in Figure 2.

3.2 Data preprocessing

Given the varying original image sizes in the Pap smear dataset, before the images were input into the proposed model, the images were first uniformly resized to the standard dimensions of 224×224 pixels. Second, to stabilize the training process and accelerate model convergence, the image pixel values were rescaled to the (0,1) range. Finally, data augmentation was applied using the following strategies: random rotation within 0–40°, horizontal and vertical shifts

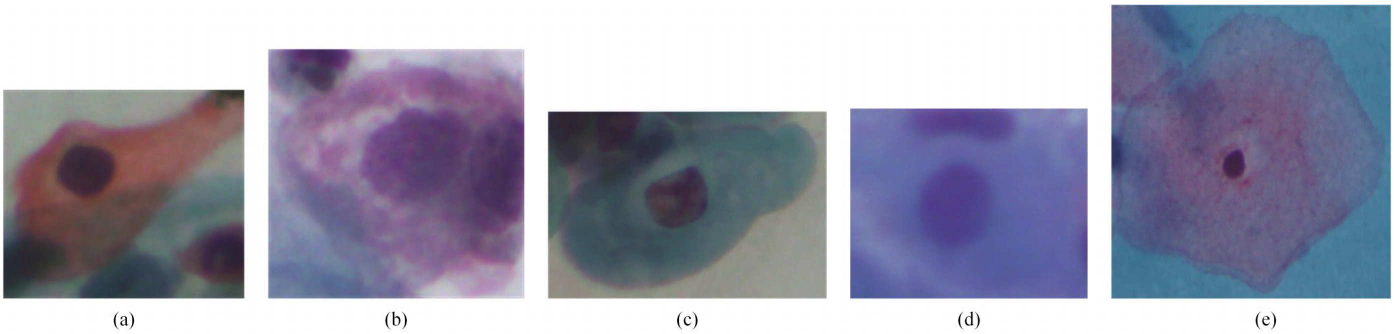


Figure 2. shows examples of Pap smear images from the SIPaKMeD dataset: (a) Dyskeratotic, (b) Koilocytotic, (c) Metaplastic, (d) Parabasal, and (e) Superficial-Intermediate.

of up to 20 in both the width and height directions, shear transformations with an intensity of 0.2, scaling operations within the range of 0.8-1.2 times, and horizontal flipping. After pre-processing, the dataset size was expanded to alleviate the issue of insufficient sample quantity.

3.3 Base CNN Feature Extractors

We select three well-established CNN architectures, pre-trained on ImageNet, for their complementary characteristics.

- DenseNet-121 is a representative dense connection network proposed by Huang et al. [18], whose core lies in its inter-layer dense connection mechanism: each layer establishes forward connections with all preceding layers. This effectively mitigates the vanishing gradient problem through feature reuse while achieving efficient feature extraction with reduced parameters. The network comprises multiple dense blocks, wherein each layer within a block receives features from all preceding layers and feeds them to subsequent layers, facilitating cross-layer feature fusion. Inter-layer transition layers incorporate 1×1 convolutions, mean pooling, and batch normalisation for dimensionality reduction and computational optimisation. DenseNet-121, pre-trained on ImageNet, exhibits robust feature generalisation capabilities and can be transferred to medical image analysis tasks.

- MobileNet-V2, proposed by Sandler et al. [19], is a lightweight convolutional neural network specifically optimised for mobile devices. The network comprises 53 layers, employing an inverted residual architecture with linear bottleneck layers. Through deep separable convolutions, it significantly reduces computational complexity, achieving efficient feature extraction with only approximately 3.5 million parameters. It processes 224×224 pixel images and is pre-trained on the ImageNet dataset, demonstrating strong transfer learning capabilities. MobileNet-V2 supports millisecond-level real-time inference while maintaining high classification accuracy, making it suitable for resource-constrained mobile devices and embedded medical image analysis scenarios.

- He et al. [20] proposed the Residual Network (ResNet), an innovative deep CNN architecture, which effectively solved the core problem of gradient vanishing in deep neural network training through the design of shortcut links with cross-layer jump connections. The core mechanism of this architecture lies in establishing direct connections between the layers and layers of the network, allowing lower level features to bypass intermediate layers and directly participate in higher-level computations. The basic building block of ResNet is the residual block, each of which contains two parallel paths. The residual path learns the residual mapping between input and output through convolution operations, while the shortcut path directly transmits cross-layer signals. The outputs of the two paths are added element-wise and used as input for the next module. This design maintains the integrity of the gradient flow, making the effective training of ultra deep networks possible.

3.4 Quantile deviation ensemble

During the prediction process of deep learning models, the model often generates extremely high probability scores for the input samples. This phenomenon can lead to overconfidence issues in both true positive and false positive classifications, as noted by Hechtlinger et al. [21]. When using such models to construct ensemble frameworks, the highly uneven distribution of raw probability scores hinders the effective learning of complementary information from the different classifiers. To achieve integrated perception of uncertainty, this paper proposes a dynamic weight fusion framework based on quantile deviation. We introduced a temperature soft maximum distribution to transform the probabilities. This distribution is defined by the temperature parameter $T$and the number of categories n, with its mathematical expression given in Formula 1.

$$\text{temp\_softmax}(x_i, T) = \frac{\exp(x_i/T)}{\sum_{j=1}^{j=n} \exp(x_j/T)} \tag{1}$$

where x_i represents the model output for i-th class. The temperature parameter T adjusts the scale of the exponential term, This smoothens the sharpness of the probability distribution. When $T > 1$, the probability distribution becomes flatter, reducing the model's overconfidence. When $T < 1$, the distribution becomes sharper, highlighting the dominant position of high probability categories. Through this standardized transformation, the probability outputs of different classifiers are mapped to a unified scale space, providing a balanced numerical foundation for subsequent ensemble weight calculations based on quantile deviation, This effectively promotes the mining and integration of complementary information.

To quantify the predictive reliability of models at different confidence levels, we introduce quantile deviation as the core metric for uncertainty perception Quantile deviation QD_qMathematical expression See Formula 2.

$$\mathrm{QD}_q = \frac{1}{NC}\sum_{i=1}^{N}\sum_{c=1}^{C}\begin{cases} q\,|\hat{p}_{i,c} - y_{i,c}|, & \hat{p}_{i,c} \geq y_{i,c} \\ (1-q)\,|\hat{p}_{i,c} - y_{i,c}|, & \text{else} \end{cases} \tag{2}$$

where N is the sample size, C is the number of categories, $\hat{p}_{i,c}$ is the model's predicted probability for the i sample in the c category, $y_{i,c}$ is the true label, and q is the quantile parameter. It quantifies the model's prediction bias at a specific quantile by assigning different loss weights based on the magnitude relationship between the predicted and true values, thereby reflecting the model prediction accuracy at the quantile of interest.

Based on quantile deviation, establish a dynamic weight allocation mechanism.The mathematical expression for the model weight $\omega_m(q)$ based on the quantile deviation is shown in Formula 3.

$$\omega_m(q) = \frac{\frac{1}{QD_q(m)}}{\sum_{k=1}^{M}\frac{1}{QD_q(k)}} \tag{3}$$

where M is the number of models, and QD_q is the quantile deviation of the m model under the q quantile. The principle is that models with smaller deviations have higher weights, and the performance is predicted based on the quantile to assign weights to multiple models in preparation for model fusion. Models with lower prediction deviation will be assigned higher weights in the ensemble.

The mathematical expression for the prediction output $\hat{y}(x)$ based on dynamic weighting is shown in Formula 4.

$$\hat{y}(x) = \sum_{m=1}^{M}\omega_m(q)\cdot f_m(x) \tag{4}$$

where $f_m(x)$ is the predicted value of the input x by the m model. By weighting $\omega_m(q)$ and merging the outputs of multiple models, models with small quantile deviations and accurate predictions will have a higher proportion in the merged results.

Mathematical expression for classification probability prediction fusion, see Formula 5.

$$\hat{P}(x) = \text{softmax}\left(\frac{1}{Q}\sum_{q=1}^{Q}\sum_{m=1}^{M}\omega_m(q)\cdot f_m(x), T\right) \tag{5}$$

where Q is the number of quantiles. By first combining the probabilities of different quantiles q and different models m, and then using temperature Softmax to obtain the final classification probability, the prediction is optimized by integrating multiple models and multiple quantile information.This fusion process explicitly incorporates the performance of each model across different confidence quantiles, enabling multi-model collaborative prediction with uncertainty awareness.

Quantile deviation serves as a key tool for quantifying prediction errors, utilizing a unique weighting method to illustrate a model's predictive performance at specific quantiles. Combining temperature Softmax based weight calculations and prediction fusion formulas, a comprehensive system is established from model evaluation to multi-model integration.

The two core hyperparameters of our framework are the temperature T and the set of quantiles Q. Their values are chosen based on calibration principles and the goal of comprehensive uncertainty profiling. **1.Temperature Parameter T:** The temperature parameter in the temp_softmax function (Equation 1) controls the sharpness of the output probability distribution. In line with the temperature scaling approach used in [22], we set $T = 2.0$ (where $T > 1$). This choice produces a flatter probability distribution, effectively reducing the model's overconfidence bias while preserving the ordinal ranking of class predictions, thereby providing a more balanced input for the subsequent quantile deviation analysis.

2. Quantile Set Q: To capture model behavior across the full spectrum of prediction confidence, we select three representative quantiles: $Q = \{0.1, 0.5, 0.9\}$. This choice is clinically and statistically motivated.

- $q = 0.1$ (Low-confidence tail): Penalizes errors where the model is under-confident (predicted probability for the true class is low). This is crucial for identifying ambiguous or borderline cases.
- $q = 0.5$ (Median): Measures the median prediction error, providing a balanced, overall assessment of model accuracy.

- $q = 0.9$ (High-confidence tail): Heavily penalizes errors where the model is over-confident (wrong with high probability).

By integrating these three distinct confidence regimes, our ensemble weighting mechanism becomes sensitive to diverse types of prediction uncertainty, leading to more robust fusion.

3.5 Decision Refinement with Multi-Layer Perceptron

The Multi-Layer Perceptron (MLP) is the most basic feedforward neural network, consisting of an input layer, several hidden layers, and an output layer. Information is transmitted between layers via fully connected connections. Its core functionality lies in utilizing nonlinear activation functions to break the constraints of linear mapping, thereby enabling the learning of complex nonlinear relationships within input data. As a typical representative of feedforward network architectures, MLP optimizes weight parameters through the backpropagation algorithm. The input layer is responsible for receiving raw features, while the hidden layers can be designed as single or multiple layers based on task requirements to perform feature transformations. The output layer employs activation functions tailored to the task type.

The ensemble probabilities embody both the consensus among models and the underlying uncertainty. To capture potential nonlinear interactions and optimize decision boundaries, we feed these probabilities into a MLP. Acting as a meta-learner, this perceptron generates the final classification prediction through a final nonlinear transformation.

4 Experiments and Results

This section will systematically discuss the evaluation system, hyperparameter tuning, result analysis, horizontal comparison, visualization verification, and cross-dataset generalization. First, we will clarify the logic behind selecting evaluation metrics for deep models and detail the methods for setting basic hyperparameters. We will then analyze the core results of the model outputs and extract key patterns from the numerical performance. Next, we will compare our framework with mainstream ensemble algorithms and state-of-the-art methods, using precise metric data to highlight the technical advantages of our framework. Additionally, we will utilize Class Activation Mapping (CAM) to visually analyze the decision focus of the model, thereby intuitively revealing the feature attention mechanism. Finally, we will conduct supplementary experiments on another independent cervical cancer dataset to validate the framework's generalization capability and stability across datasets.

4.1 Evaluation indicators

To evaluate the effectiveness of the framework, various evaluation metrics were used. These metrics include accuracy, precision, recall, and F1 score. The above metrics are calculated using the following formulas. In addition, confusion matrices and receiver operating characteristic (ROC) curves were used to evaluate the performance of the framework. In a binary classification modelTP, FP, FN and TN represent the number of true positives, false positives, false negatives, and true negatives, respectively.

Accuracy refers to the proportion of samples that are correctly classified in all samples:

$$\text{Accuracy} = \frac{TN + TP}{TP + TN + FP + FN} \tag{6}$$

Precision refers to the proportion of true positives among samples predicted to be positive:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

The recall rate refers to the proportion of true positives correctly predicted among all actual positive samples:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

The F1 score is the harmonic mean of precision and recall, providing a balanced assessment that considers both metrics simultaneously:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

4.2 Hyperparameter settings

For deep CNN models, it is essential to carefully select hyperparameters, particularly the learning rate and batch size. The learning rate controls the extent to which newly learned information overwrites existing knowledge. Batch size refers to the number of samples loaded at once. The hyperparameters used to train the CNN model were determined through extensive experimental settings to identify the optimal combination of base learners for the proposed ensemble technique. The hyperparameters selected for this experiment are detailed in Table 2. The same learning rate and batch size were applied to all three base classifiers. The Adam optimizer was used to optimize the internal weights. The model performed best at a batch size of 16 and a learning rate of 1e-4.

Table 2. Basic hyperparameter settings used for training.

Hyperparameter	Value
Optimizer	Adam
Loss function	Categorical cross entropy
Learning rate	1e-4
Batch size	16
Number of epochs	75

4.3 Analysis of results

Our proposed framework was evaluated on the SIPaKMeD dataset, which contains 5-class cervical cancer cell data, using 5-fold cross-validation to validate the model. Table 3 presents the classification accuracy across all folds. Our uncertainty-aware ensemble achieves an average accuracy of 97.68%, which is significantly higher than any single base model (DenseNet121: 96.52%, MobileNetV2: 94.94%, ResNet-50: 96.94%). After MLP refinement, the accuracy further improves to 98.10%. This demonstrates the effectiveness of both the quantile deviation weighting and the subsequent non-linear optimization.

It is worth noting that the base classifiers did not show obvious overfitting in the experiment, which is a key factor in the success of this method. The visualization results in Figure 3 provide strong evidence for this. Theoretically, the purpose of quantile deviation integration is to accurately estimate the bias at different quantiles. If the model overfits, it will cause the quantile estimation to be biased, such as overestimating or underestimating the bias of extreme quantiles, resulting in the loss of a reliable basis for the decision of quantile deviation integration. Figures 4 and 5 present the receiver operating characteristic (ROC) curves and confusion matrices for the first fold of the experiment, respectively. From the ROC curves, it can be seen that the AUC values for each pathological cell category are close to 1, fully demonstrating the model's strong ability to distinguish between different categories. The elements on the main diagonal of the confusion matrix show a high value distribution, verifying the model's accurate classification of most categories. Although there are a few misclassification cases due to feature overlap, the overall proportion is extremely low. The two sets of visualization results collectively indicate that the model demonstrates outstanding performance in the task of classifying cervical cancer pathological cells. In subsequent work, we can specifically optimize the classification logic for categories with significant feature overlap by combining attention mechanisms, contrastive learning, and other strategies to further reduce the misclassification rate.

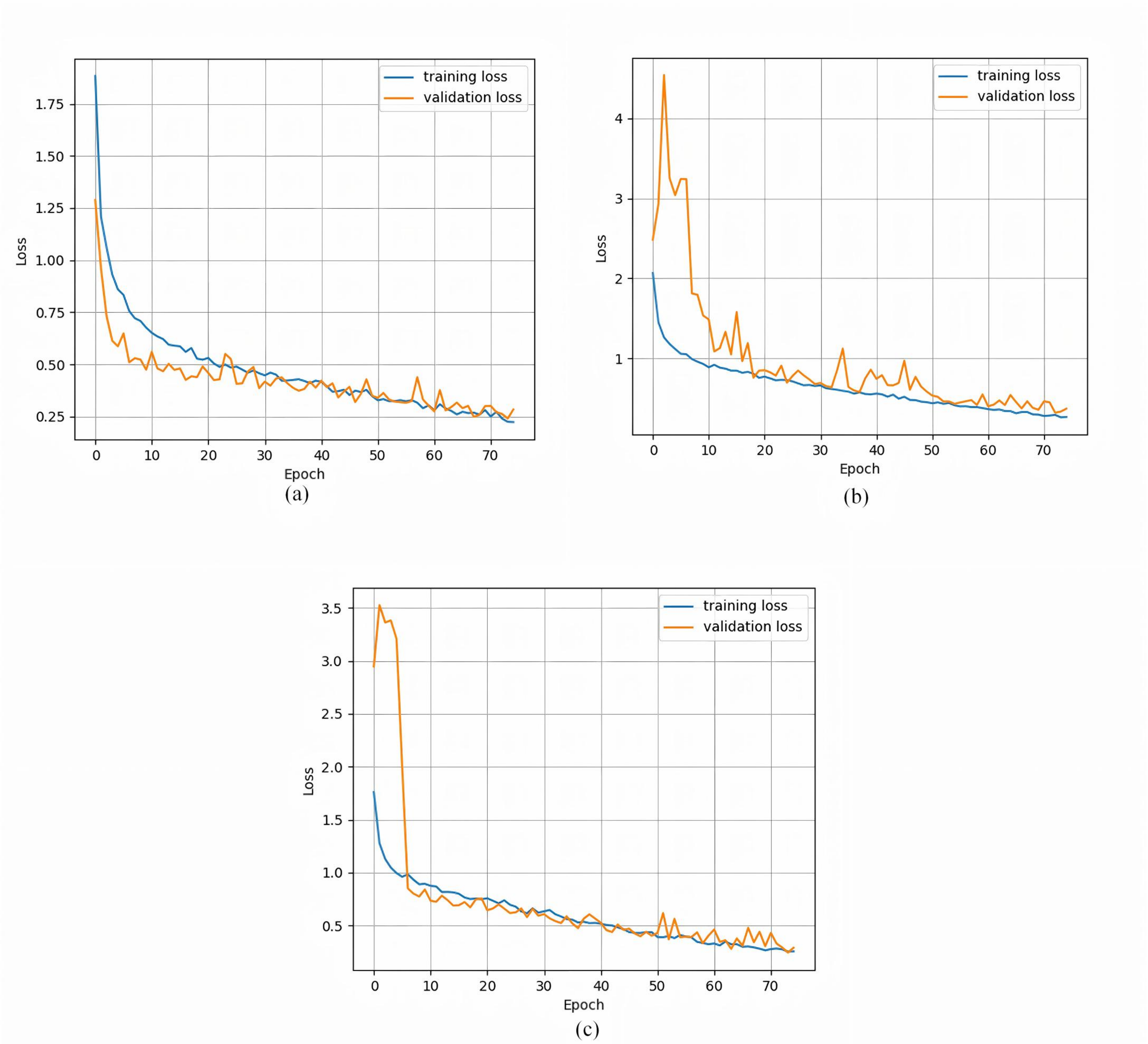
To validate the feasibility of using the base learners in the proposed integrated framework for model integration, a McNemar statistical test [23] was conducted. This test is a non-parametric analysis method for paired data distributions, where the p-value represents the probability that the two models perform similarly. Theoretically, a lower p-value is required to reject the null hypothesis that "the two models are similar." When the p-value is less than 0.05, it can be concluded that the two compared models exhibit significant statistical differences. As shown in Table 4, the null hypotheses were all rejected, clearly indicating that the ensemble model and the base learners exhibit distinct statistical performance, thereby strongly supporting the feasibility of using base learners for ensemble construction.

4.4 Ablation Study and Comparison with Ensemble Strategies

Table 5 compares our method with two baseline integration strategies. The proposed approach

Table 3. Results of 5-fold cross-validation on the SIPaKMeD dataset.

Fold_no	DenseNet121(%)	MobileNetV2(%)	ResNet50(%)	Qua_ensemble(%)	MLP(%)
Fold_1	95.43	95.93	95.56	96.91	97.28
Fold_2	98.02	94.69	96.91	97.53	98.15
Fold_3	97.16	94.20	96.91	98.77	98.89
Fold_4	96.05	94.57	98.02	98.15	98.27
Fold_5	95.92	95.30	97.28	97.03	97.90
Average	96.52	94.94	96.94	97.68	98.10



(a)

(b)

(c)

Figure 3. Training and validation loss curves for the first experimental fold using (a) DenseNet121, (b) MobileNetV2, and (c) ResNet-50.

Table 4. P-values for McNemar's test.

Model	p-value
DenseNet121	0.0414
MobileNetV2	0.0258
ResNet50	0.0436

achieves superior classification performance with an accuracy of 97.68%, compared to 96.30% for soft voting and 97.53% for the summation rule. These results confirm that the dynamic weighting strategy, which adjusts model contributions based on statistically quantified uncertainty, outperforms static and uniform combination rules in medical image classification tasks.

Table 5. Comparison of different ensemble strategies.

Method	Accuracy(%)
Soft voting	96.30
Sum rule	97.53
Our proposed ensemble	97.68

Table 6 illustrates the competitiveness of the proposed integrated method compared to advanced integrated

Table 6. Comparison with other ensemble methods.

Method	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Fuzzy rank [1]	95.43	95.34	95.38	95.36
Fuzzy distance [23]	96.47	96.51	96.53	96.45
MSENet [22]	97.21	97.16	97.15	97.14
Our proposed ensemble	97.68	97.70	97.67	97.69

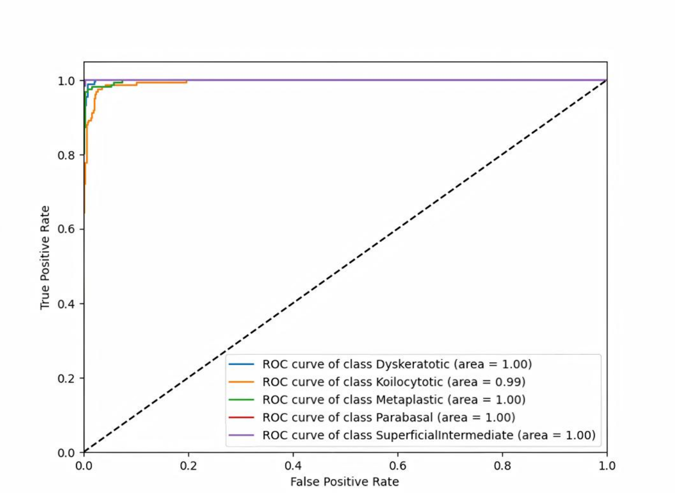


Figure 4. ROC curve for the first fold of the experiment.

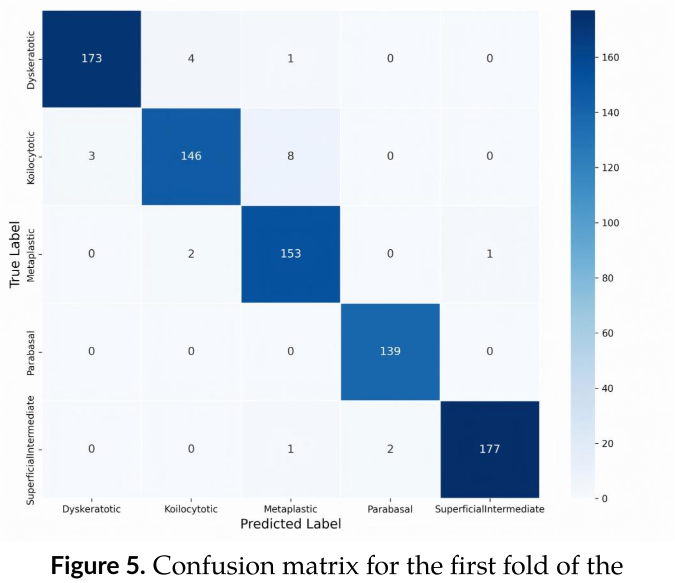


Figure 5. Confusion matrix for the first fold of the experiment.

methods. Manna et al. [1] introduced the concept of fuzzy rank ordering to evaluate the final decision. Pramanik et al. [22] proposed a method that constructs an aggregation mechanism based on fuzzy distance. In another study, Pramanik et al. [23] designed an integration strategy based on mean and standard deviation. As shown in Table 6, the method proposed in this paper outperforms these methods in the comparison.

4.5 Comparison with advanced methods

When introducing any new method, it is extremely important to compare its performance with existing methods. To this end, we have listed the proposed method and some recently proposed methods for comparison in Table 7. Wubineh et al. [24] proposed the RES-DCGAN data augmentation technique (adding residual blocks to the DCGAN generator to enhance data flow and image quality) and incorporated a self-attention mechanism into pre-trained models such as ResNet50V2, Xception, and DenseNet121, using the Pomeranian and SIPaKMeD datasets for cervical cell classification. Sahoo et al. [2] combined three ensemble techniques with advanced data augmentation, using a weighted filtering method to assign weights based on the validation performance of each intermediate ensemble model, and fused the results of the three ensemble methods to obtain the final classification.

4.6 Time analysis

This section will introduce the execution time required for each base classifier. Due to hardware resource constraints, the model was trained separately, with the relevant experimental details shown in Figure 6. The experiment was set up with 75 training iterations and a batch size of 16. Based on calculations, the total training time for a single batch of images is 372ms, which translates to approximately $(134 + 109 + 129) \div 16 \approx 23.25\text{ms/image}$. This result is based on a computing device with 24GB of VRAM, powered by an NVIDIA 3090 GPU, and implemented using the Python 3.9 programming language and the TensorFlow Keras framework. It is important to note that different hardware configurations and software environments may result in variations in execution time.

4.7 Uncertainty Visualization and Model Explainability

The core of deep learning classification tasks lies in feature extraction, and the information content of these features directly determines the upper limit of classification performance. Gradient-Weighted Class

Table 7. Comparison with other advanced methods.

Work Ref.	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Liu et al. [9]	93.48	93.60	93.50	93.50
Wubineh et al. [24]	92	92	92	92
Sahoo et al. [2]	97.62	97.65	97.64	97.64
Our proposed	98.10	98.21	98.05	98.13

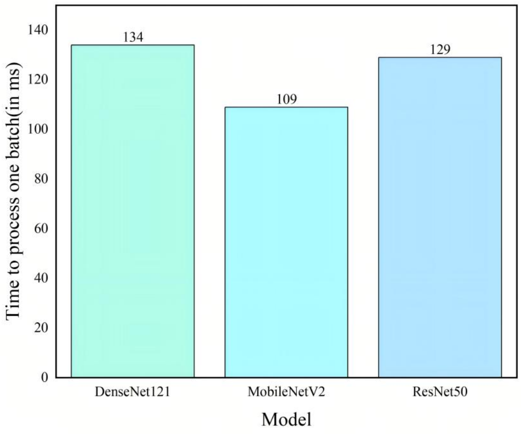


Figure 6. Training time of CNN model using a single batch of images.

Activation Maps (Grad-CAM) serve as an effective tool for visualizing the regions of interest in CNN models, providing an intuitive representation of the basis for model decisions[25]. This paper applies this technique to three test samples of cervical cell smear images, with results shown in Figures 7, 8 and 9.

By generating heatmaps from the final convolutional layer outputs, the feature attention mechanisms of different networks are clearly revealed: DenseNet121 achieves feature reuse through its dense connection structure, with its heatmap covering a broad area, demonstrating the ability to capture global contextual information; MobileNetV2, based on separable convolutions and inverted residual structures, precisely focuses its heatmap on core lesion areas, reflecting the optimized design of lightweight architectures for efficient feature extraction; ResNet-50 achieves cross layer feature propagation through residual connections, with its heatmap exhibiting dual characteristics of hierarchical focus and detail supplementation, balancing global semantics with local details.

Visualization results indicate that all three models effectively locate key discriminative regions within images, yet exhibit significant differences in heatmap distribution patterns and focus intensity. These variations intuitively reflect distinct feature extraction preferences and sensitivities across network architectures when processing cervical cell images, providing visual evidence for complementary design in ensemble models. The above analysis validates Grad-CAM's role in explaining model decisions and comparing how different models focus on key features in medical images, providing reference for optimizing models to meet medical diagnostic needs.

Our framework exhibits reliable and robust performance in cervical cancer detection, yet limitations remain. Error case studies and Grad-CAM visualizations reveal high confidence misclassifications when distinguishing morphologically similar cell types, highlighting challenges in capturing subtle pathological features. For a misclassified dyskeratotic cell, DenseNet121's activation is dispersed across nuclear fragments, MobileNetV2's attention is disrupted by background noise, and ResNet50 erroneously focuses on perinuclear bright areas all failing to grasp key overall or cytoplasmic features. For a misclassified koilocytotic cell, none of the three models focus on the critical perinuclear halo: DenseNet121 fixates on intranuclear fragments, MobileNetV2 ignores the halo and misinterprets cytoplasm, and ResNet50 concentrates on irrelevant background. These misclassifications mainly stem from insufficient "borderline samples" in training data, systematic biases in feature extraction, and attention mechanisms being susceptible to interference from complex image backgrounds. Visual evidence for these typical error cases is provided in Figures 10 and 11.

To systematically address these issues and further reduce the misclassification rate, we propose and plan to test the following targeted solutions in future work:

- Attention Mechanism Enhancement: Integrate self-attention or squeeze-and-excitation (SE) blocks into the base CNN architectures. This would explicitly encourage the model to weigh informative cellular regions more heavily

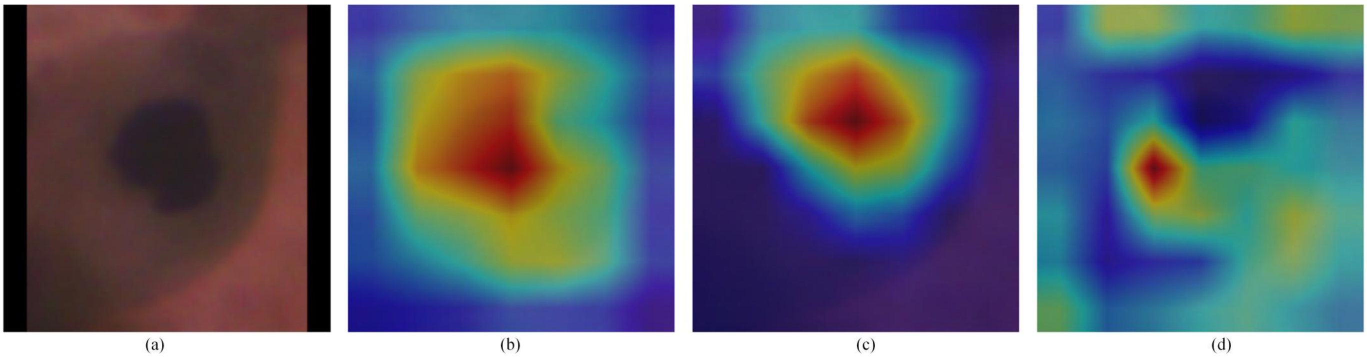


Figure 7. shows the analysis of Dyskeratotic category images using Grad-CAM. (a) Original image, (b) Grad-CAM of DenseNet121, (c) Grad-CAM of MobileNetV2, and (d) Grad-CAM of ResNet50.

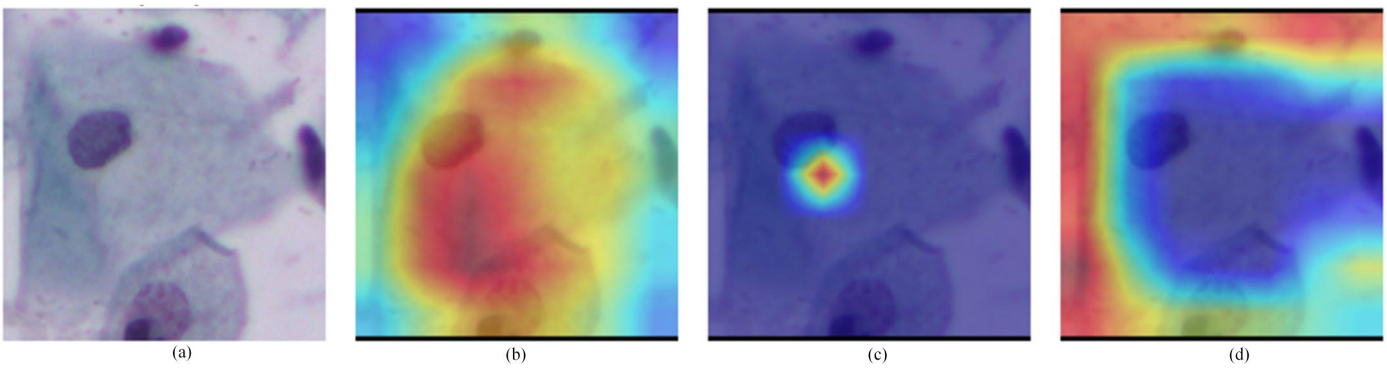


Figure 8. shows the analysis of Superficial-Intermediate category images using Grad-CAM. (a) Original image, (b) Grad-CAM of DenseNet121, (c) Grad-CAM of MobileNetV2, and (d) Grad-CAM of ResNet50.

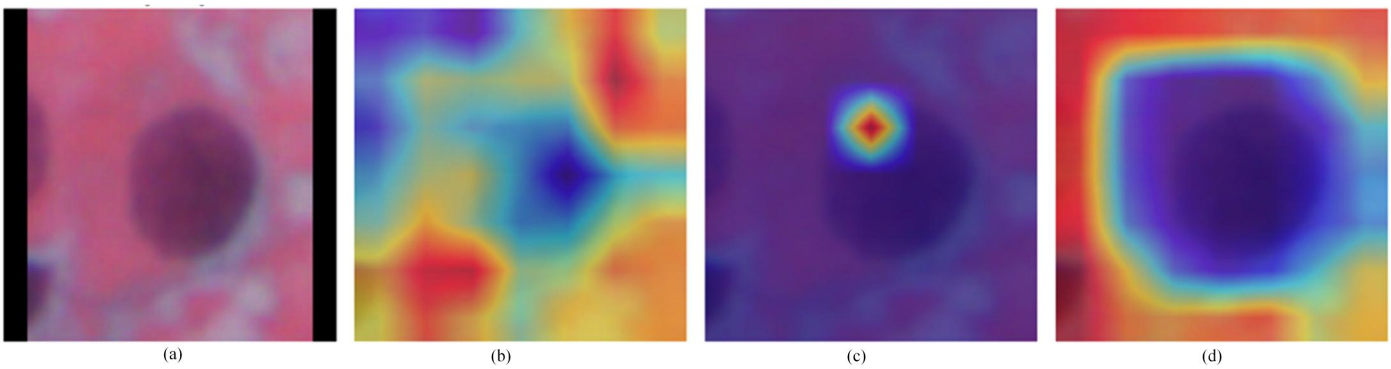


Figure 9. shows the analysis of Parabasal category images using Grad-CAM. (a) Original image, (b) Grad-CAM of DenseNet121, (c) Grad-CAM of MobileNetV2, and (d) Grad-CAM of ResNet50.

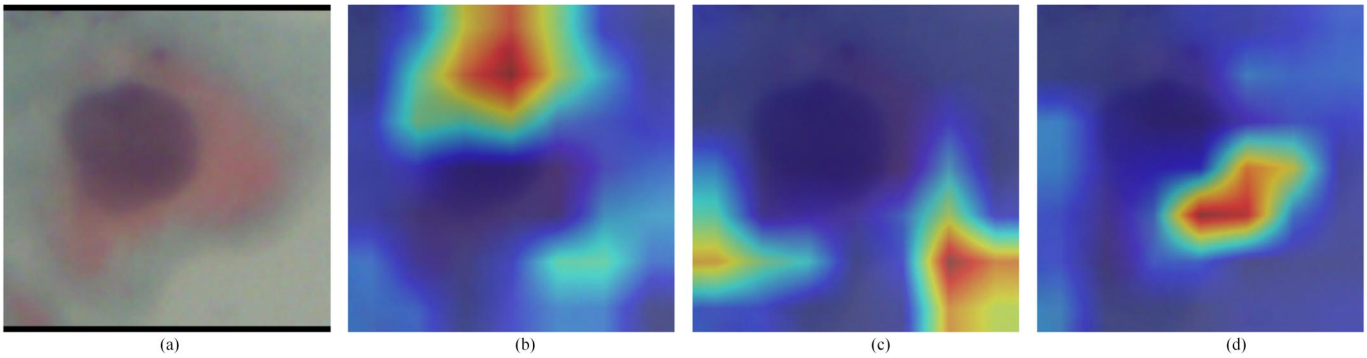


Figure 10. shows the misclassification analysis of Dyskeratotic images using Grad-CAM. (a) Original image, (b) Grad-CAM of DenseNet121, (c) Grad-CAM of MobileNetV2, and (d) Grad-CAM of ResNet50.

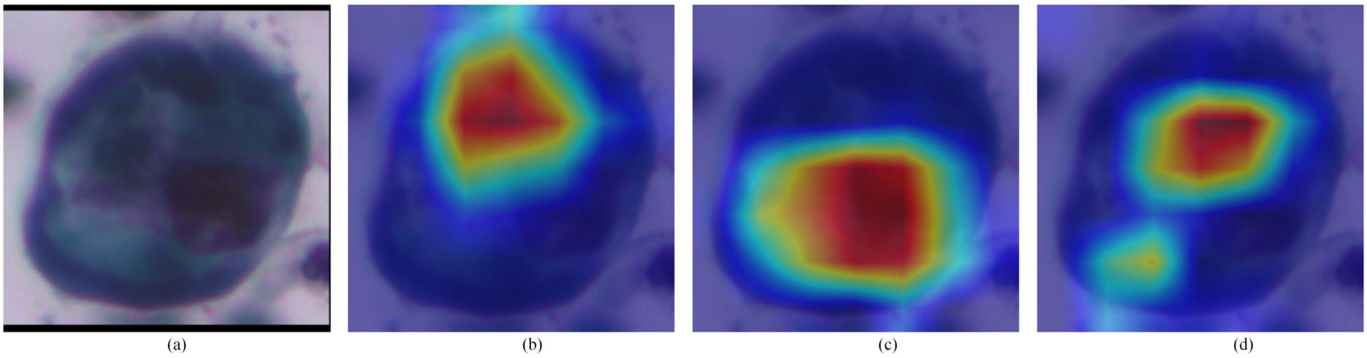


Figure 11. shows the misclassification analysis of Koilocytotic images using Grad-CAM. (a) Original image, (b) Grad-CAM of DenseNet121, (c) Grad-CAM of MobileNetV2, and (d) Grad-CAM of ResNet50.

while suppressing responses from irrelevant background, directly mitigating the noise interference problem observed.

- Contrastive Learning for Fine-Grained Discrimination: Employ supervised contrastive learning during pre-training or as an auxiliary loss. By learning an embedding space where morphologically similar but pathologically distinct cell types (e.g., borderline cases) are pushed apart, the model can be trained to extract more subtle and discriminative features, improving its ability to separate classes with overlapping characteristics.

- Targeted Data Augmentation: Beyond general augmentations, develop pathology-aware augmentation strategies. For instance, simulate varying degrees of cell overlap, stain intensity variations, or add controlled synthetic noise to the background during training. This would explicitly increase the model's robustness to the confounding factors identified in error cases.

We employ gradient-weighted class activation mapping (Grad-CAM) to visualize the focal regions of attention for each base model. The visualizations reveal complementary attention patterns: DenseNet121 typically captures broader contextual information, MobileNetV2 precisely focuses on core cellular structures, while ResNet-50 balances global and local features. This diversity validates the rationale for our ensemble approach. More importantly, when classification errors occur, Grad-CAM traces them back to specific model biases such as attention disrupted by background noise or atypical fragments providing valuable insights for model debugging and uncertainty interpretation.

4.8 Cross-Dataset Generalization

To assess generalization, we tested our framework on an independent liquid-based cytology dataset (Mendeley-LBC). As shown in Figure 12, our method achieved 99.48% accuracy, demonstrating its robustness and transferability to different data distributions and preparation protocols.

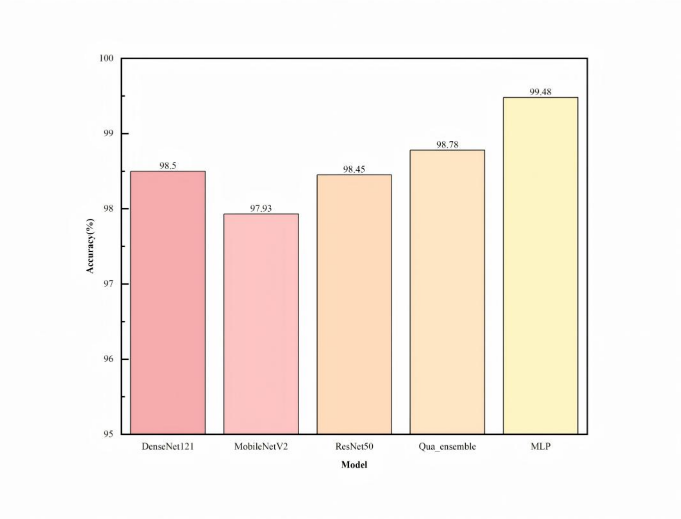


Figure 12. Results obtained from additional experiments conducted on the Mendeley-LBC dataset.

5 Discussion and Limitations

The proposed framework successfully bridges statistical uncertainty quantification with deep ensemble learning, offering a principled, explainable, and high performing solution for cervical cytology classification. The quantile deviation mechanism provides a computationally efficient way to measure and utilize prediction uncertainty, guiding a more intelligent model fusion.

5.1 Limitations and Future Work

Dataset Scope: While the current evaluation is comprehensive, it is primarily based on publicly available datasets with a predefined set of abnormality categories. Real-world cervical cancer screening involves a wider spectrum of cytological presentations, including rare subtypes and diagnostically challenging borderline cases, which are underrepresented in our training data. To address this, future work will focus on expanding the dataset through multi center collaborations to collect diverse clinical samples. Additionally, we plan to explore advanced synthetic data generation techniques, such as generative adversarial networks (GANs), to artificially create realistic images of rare or borderline cytological features, thereby enhancing the model's robustness and generalizability to real clinical scenarios.

Computational Cost: The training phase of our framework, which involves multiple deep CNNs, remains computationally intensive. While the inference is efficient, the initial training overhead could be a consideration for resource constrained settings. To enhance practical deployment, future work will investigate two directions: (1) Lightweight modifications: exploring further architectural optimizations or employing more efficient base networks (e.g., EfficientNet) to reduce the model footprint. (2) Efficient training strategies: implementing knowledge distillation to transfer the ensemble's knowledge into a single, compact student model, thereby drastically reducing inference time and memory requirements without significant performance degradation, making it more suitable for real-time or mobile health applications.

Uncertainty Granularity: The current framework quantifies prediction uncertainty at the ensemble model level. While this informs the overall reliability of a classification decision for a whole image, it does not provide pixel-wise or instance-wise uncertainty estimates, which could offer pathologists more detailed insights into which specific cellular regions contribute most to diagnostic uncertainty. As a key direction for future work, we plan to extend our framework to incorporate fine-grained uncertainty estimation. This could be achieved by integrating our ensemble approach with probabilistic segmentation networks or developing uncertainty back propagation techniques to generate spatial uncertainty heatmaps. Such maps would highlight ambiguous areas within a cell (e.g., overlapping nuclei, staining artifacts), offering a deeper level of interpretability and aiding in the review of challenging cases.

Future Work: integrating more advanced uncertainty estimation techniques, expanding the framework to handle whole slide images, and validating the system in prospective clinical studies to assess its real world impact on diagnostic workflow and patient outcomes.

6 Conclusion

We propose a novel uncertainty-aware ensemble framework for cervical cytology classification that integrates statistical quantile analysis with deep learning. By introducing a weighting scheme based on quantile deviation, we achieve dynamic, uncertainty aware fusion of multi-layer convolutional neural network predictors. Subsequent refinement through multilayer perceptrons further enhances performance. Experimental results demonstrate that our framework outperforms several existing state-of-the-art methods on benchmark datasets while providing interpretable decision insights through visualization techniques. This work establishes a robust methodological paradigm for uncertainty aware medical image analysis, whose scalability holds broad promise for other interdisciplinary applications particularly in scenarios where reliable and explainable predictions are critical.

Data Availability Statement

The data used to support the findings of this study are derived from the SIPaKMeD dataset, which is publicly available on Kaggle at https://www.kaggle.com/datasets/?search=sipakmed

Funding

This work was supported by the National Natural Science Foundation of China under Grant 62566012.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that AI-assisted tools were used solely for language translation and proofreading purposes. DeepL Translator (v3) was employed to translate portions of the manuscript from Chinese into English, and DeepSeek-R1 was used for language editing and proofreading. No generative AI was used for content creation, data analysis, or scientific interpretation.

Ethical Approval and Consent to Participate

Not applicable. This study utilized publicly available datasets that do not involve direct interaction with human participants or collection of new identifiable data.

References

[1] Manna, A., Kundu, R., Kaplun, D., Sinitca, A., & Sarkar, R. (2021). A fuzzy rank-based ensemble of CNN models for classification of cervical cytology. *Scientific Reports*, *11*(1), 14538. [CrossRef]

[2] Sahoo, P., Saha, S., Sharma, S. K., Kanungo, A., & Verma, N. (2025). Boosting cervical cancer detection with a multi-stage architecture and complementary information fusion. *Soft Computing*, *29*(2), 1191–1206. [CrossRef]

[3] Wasswa, W., Ware, A., Basaza-Ejiri, A. H., & Obungoloch, J. (2018). A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. *Computer Methods and Programs in Biomedicine*, *164*, 15–22. [CrossRef]

[4] Attallah, O. (2022). Deep learning-based CAD system for COVID-19 diagnosis via spectral-temporal images. In *Proceedings of the 12th International Conference on Information Communication and Management* (pp. 25–33). [CrossRef]

[5] Plissiti, M. E., Dimitrakopoulos, P., Sfikas, G., Nikou, C., Krikoni, O., & Charchanti, A. (2018). Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. In *2018 25th IEEE International Conference on Image Processing (ICIP)* (pp. 3144–3148). IEEE. [CrossRef]

[6] Nanni, L., Ghidoni, S., & Brahnam, S. (2017). Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognition*, *71*, 158–172. [CrossRef]

[7] Rahaman, M. M., Li, C., Yao, Y., Kulwa, F., Wu, X., Li, X., & Wang, Q. (2021). DeepCervix: A deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques. *Computers in Biology and Medicine*, *136*, 104649. [CrossRef]

[8] Basak, H., Kundu, R., Chakraborty, S., & Das, N. (2021). Cervical cytology classification using PCA and GWO enhanced deep features selection. *SN Computer Science*, *2*(5), 369. [CrossRef]

[9] Liu, W., Li, C., Xu, N., Jiang, P., & Jiang, Y. (2022). CVM-Cervix: A hybrid cervical Pap-smear image classification framework using CNN, visual transformer and multilayer perceptron. *Pattern Recognition*, *130*, 108829. [CrossRef]

[10] Newaz, A., Muhtadi, S., & Haq, F. S. (2022). An intelligent decision support system for the accurate diagnosis of cervical cancer. *Knowledge-Based Systems*, *245*, 108634. [CrossRef]

[11] Ghoneim, A., Muhammad, G., & Hossain, M. S. (2020). Cervical cancer classification using convolutional neural networks and extreme learning machines. *Future Generation Computer Systems*, *102*, 643–649. [CrossRef]

[12] Tawalbeh, S., Alquran, H., & Alsalatie, M. (2023). Deep feature engineering in colposcopy image recognition: A comparative study. *Bioengineering*, *10*(1), 105. [CrossRef]

[13] Kalbhor, M. M., & Shinde, S. V. (2023). Cervical cancer diagnosis using convolution neural network: feature learning and transfer learning approaches. *Soft Computing*, 1-11. [CrossRef]

[14] Akpinar, M. H., Atila, O., Sengur, A., Salvi, M., & Acharya, U. R. (2025). A novel uncertainty-aware liquid neural network for noise-resilient time series forecasting and classification. *Chaos, Solitons & Fractals*, *193*, 116130. [CrossRef]

[15] Han, Q., Qian, X., Xu, H., Wang, Y., He, F., & Shi, Y. (2024). DM-CNN: Dynamic multi-scale convolutional neural network with uncertainty quantification for medical image classification. *Computers in Biology and Medicine*, *168*, 107758. [CrossRef]

[16] Ullah, N., Guzmán-Aroca, F., Martínez-Álvarez, F., & Troncoso, A. (2025). A novel explainable AI framework for medical image classification integrating statistical, visual, and rule-based methods. *Medical Image Analysis*, *97*, 103665. [CrossRef]

[17] Wang, K., Shariatmadar, K., Manchingal, S. K., Mirzazadeh, M., Ghorbani, M., Kuhn, S., & Cuzzolin, F. (2025). Creinns: Credal-set interval neural networks for uncertainty estimation in classification tasks. *Neural Networks*, *185*, 107198. [CrossRef]

[18] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017, July). Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2261-2269). IEEE Computer Society. [CrossRef]

[19] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018, June). MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4510-4520). IEEE. [CrossRef]

[20] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). [CrossRef]

[21] Hechtlinger, Y., Póczos, B., & Wasserman, L. (2018). Cautious deep learning. *arXiv preprint arXiv:1805.09460*.

[22] Pramanik, R., Banerjee, B., & Sarkar, R. (2023). MSENet: Mean and standard deviation based ensemble network for cervical cancer detection.

Engineering Applications of Artificial Intelligence, *123*, 106336. [CrossRef]

[23] Pramanik, R., Biswas, M., Sen, S., de Souza Júnior, L. A., Papa, J. P., & Sarkar, R. (2022). A fuzzy distance-based ensemble of deep models for cervical cancer detection. *Computer Methods and Programs in Biomedicine*, *219*, 106776. [CrossRef]

[24] Wubineh, B. Z., Rusiecki, A., & Halawa, K. (2024). Classification of cervical cells from the Pap smear image using the RES_DCGAN data augmentation and ResNet50V2 with self-attention architecture. *Neural Computing and Applications*, *36*(34), 21801–21815. [CrossRef]

[25] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). [CrossRef]

Hui Hu is currently pursuing a master's degree in statistics at Guizhou Minzu University. (Email: 1686836244@qq.com)

Jiewu Huang received a Master's degree in Probability Theory and Mathematical Statistics from Central South University (Changsha, China) in 2006, and a Ph.D. in Computational Mathematics from Chongqing University (Chongqing, China) in 2014. (Email: ylgw217@163.com)

Shoumei Cao is currently pursuing a master's degree in statistics at Guizhou Minzu University. (Email: 2459344243@qq.com)

Anna Dai is currently pursuing a master's degree in statistics at Guizhou Minzu University. (Email: 2087998360@qq.com)