Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/OJCS.XXXX.XXXXXXX

Distilling Ensemble Knowledge via a Teacher-Assistant for Explainable Cervical Cancer Screening on Edge Devices

Umar Hasan[1], Muhammad Rafsan Kabir[1], Abdullah Al Raiyan[1], Md. Sifat Haque Zidan[1], and Sifat Momen[1]

[1]Department of Electrical and Computer Engineering, North South University, Dhaka, 1229, Bangladesh

Corresponding author: Sifat Momen (Email: sifat.momen@northsouth.edu)

ABSTRACT Cervical cancer remains one of the leading causes of cancer-related mortality among women worldwide, necessitating the urgent need for accurate, efficient, and explainable diagnostic tools. However, existing deep learning approaches often face a trade-off between high predictive accuracy, computational efficiency, and interpretability, which limits their practical deployment, especially in resource-constrained settings. To address these challenges, this study employs an explainable deep learning framework that integrates ensemble learning, Teacher-Assistant Knowledge Distillation (TAKD), and Explainable AI (XAI) techniques. Specifically, we introduce AgileNet, a custom-designed, lightweight student model (0.29 million parameters), distilled from a large, high-performing ensemble through an intermediate teacher assistant. AgileNet achieves state-of-the-art performance on cervical squamous cell classification tasks, reaching 97.88% accuracy, while maintaining low computational complexity. To enhance transparency and clinical trust, we incorporate Grad-CAM++ visualizations that offer interpretable insights into the model's predictions. Extensive experiments demonstrate AgileNet's superior balance between accuracy, model compactness, and inference speed. Overall, the proposed framework provides a promising solution for accessible, reliable, and explainable AI-driven cervical cancer screening, paving the way for future clinical integration. Please write all the new texts like this.

INDEX TERMS Cervical Cancer, Ensemble, Explainable AI, Teacher-Assistant Knowledge Distillation

I. INTRODUCTION

CERVICAL cancer [1], which originates in the cells of the cervix, poses a significant threat to women's health worldwide. It remains a major cause of morbidity and mortality among women, with a disproportionately high impact in developing nations [2], [3]. In developing countries such as Bangladesh, cervical cancer ranks as the second leading malignancy among females, accounting for 12% of cancer cases [4]. Statistics from 2018 reported 8,068 new diagnoses and 5,214 fatalities [4]. Projections suggest that without effective intervention, cumulative deaths from cervical cancer in Bangladesh could reach 505,703 by 2070 and potentially rise to 1,042,859 by 2120 [5]. Microscopic examination of Pap smears remains the gold standard for detecting precancerous and cancerous changes in the cervix [6]. However, the manual interpretation of Pap smears is inherently time-consuming and subjective, requiring highly trained cytopathologists [7], [8]. This poses significant challenges for large-scale screening programs, particularly in regions with limited expert resources, potentially resulting in delays in diagnosis and treatment [2]. These challenges highlight the need for efficient Computer-Aided Diagnosis (CAD) systems. Early detection through screening programs, primarily using the Papanicolaou (Pap) smear test, is essential for enabling timely treatment and improving patient outcomes [2].

Recent research in cervical cancer screening has increasingly leveraged deep learning (DL) models, particularly Convolutional Neural Networks (CNNs), often combined with transfer learning from models pre-trained on large-scale datasets, to automatically extract features and classify cervical cell images [9], [10]. While these methods have demonstrated significant potential in medical image analysis, delivering highly accurate classifications, they face persis-

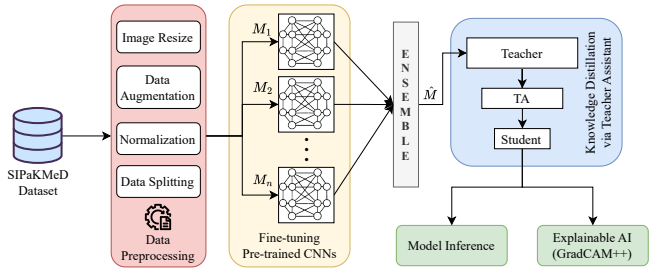

FIGURE 1: Methodology of the proposed framework for cervical cancer screening, utilizing a Teacher-Assistant Knowledge Distillation (TAKD) pipeline.

tent challenges. Most existing approaches produce computationally intensive models with large parameters, resulting in long inference times and substantial memory requirements, which can hinder deployment in resource-constrained clinical environments. Furthermore, publicly available Pap smear datasets, such as SIPaKMeD [11], are relatively small compared to general computer vision datasets [12]–[14], increasing the risk of overfitting. Another critical issue is the lack of transparency in many DL-based solutions, which often function as "black boxes," limiting their clinical acceptability and trust. Although ensemble methods, such as the one proposed by Gangrade et al. [2], have achieved high classification accuracy (94%), a significant research gap remains, highlighting the need for frameworks that not only improve performance but also enhance model efficiency (in terms of parameters and inference time) and strengthen interpretability to support clinical decision-making.

To address these identified gaps, this paper proposes an advanced, explainable DL framework to overcome current limitations in cervical squamous cell classification. Our approach leverages transfer learning to fine-tune multiple CNN models and combines their strengths through ensemble learning to create a robust teacher model. To achieve high model efficiency, we employ a Teacher-Assistant Knowledge Distillation (TAKD) pipeline [15]. Unlike traditional knowledge distillation [16], which transfers knowledge directly from teacher to student, TAKD introduces an intermediate Teacher Assistant (TA) model to bridge the knowledge gap more effectively. In our framework, the ensemble model serves as the teacher, transferring its knowledge to a lightweight student model, AgileNet, which contains only 0.29 million parameters. This strategy aims to deliver a model that is not only highly accurate but also computationally efficient, making it well-suited for deployment in clinical settings. Additionally, we integrate Explainable AI (XAI) techniques using Grad-CAM++ [17] to enhance model transparency and provide interpretability. Figure 1 summarizes the overall methodology of our proposed framework. Experimental results demonstrate that the proposed system performs strongly on the SIPaKMeD dataset while maintaining a lightweight architecture and offering meaningful insights into AgileNet's decision-making process. This work is essential to developing more accurate, efficient, and trustworthy Computer-Aided Diagnosis (CAD) systems for cervical cancer screening.

The primary contributions of this work are as follows:

- We propose an advanced, explainable DL framework designed explicitly for cervical squamous cell classification, addressing key limitations of existing models in terms of efficiency and interpretability.
- We leverage ensemble learning by fine-tuning multiple CNN models through transfer learning, combining their strengths to build a high-performing teacher model.
- We introduce a Teacher-Assistant Knowledge Distillation (TAKD) pipeline, which employs an intermediate Teacher Assistant (TA) model to enable more effective knowledge transfer to a lightweight student model, AgileNet, comprising only 0.29 million parameters.
- The integration of Explainable AI (XAI) techniques, specifically Grad-CAM++, provides visual interpretation and enhances transparency in the model's decision-making process.
- Extensive experiments on the SIPaKMeD dataset demonstrate that our proposed framework achieves high classification performance with a lightweight architecture, making it well-suited for deployment in resource-constrained clinical settings.

II. RELATED WORKS

A. Cervical Squamous Cell Classification

Early efforts in computational pathology explored various approaches for cervical cancer detection and squamous cell classification. Mango et al. [18] combined traditional Pap tests with artificial neural networks (ANN), while Sukumar and Gnanamurthy [19] utilized MRI scans with hybrid SVM and neuro-fuzzy classifiers. The advent of deep learning brought significant advancements. Bora et al. [20] applied deep CNNs for image identification, improving accuracy via feature selection, and Hyeon et al. [21] demonstrated the effectiveness of CNNs like VGG16 for feature extraction from cervical MRIs. Sanyal et al. [22] also developed a CNN for finding abnormal areas in conventional smears with high diagnostic accuracy. Complementary techniques like surface-enhanced Raman scattering (SERS) combined with machine learning also showed promise for predicting sample pathology [23]. These foundational works established the potential of computational methods, particularly deep learning, in analyzing complex medical data for cervical squamous cell classification, inspiring our adoption of deep learning techniques. However, while promising, these early methods often required significant feature engineering or faced limitations in handling the complexity and variability of cytological images. This indicated a clear research gap: the need for more robust and automated feature learning approaches to overcome these limitations, such as those offered by transfer learning.

B. Transfer Learning and Ensemble

To improve upon earlier methods and harness the power of large-scale datasets, transfer learning and ensembling have become common and effective approaches in medical analysis, including cervical cancer screening. Promworn et al. [24] compared several deep learning models on cytopathology images, identifying DenseNet161 [25] as a top performer. The utility of transfer learning, using pre-trained CNNs like AlexNet [26] and VGG-16 [27], became evident, with studies like Taha et al. [28] showing notable accuracy improvements by leveraging pre-trained features. Ensemble methods further boosted performance; Xue et al. [29] used Ensemble Transfer Learning (ETL) for histopathology images, while Chen et al. [30] also achieved high accuracy with CNNs and transfer learning. Ghoneim et al. [31] combined CNNs with ELM classifiers, and Arifianto et al. [32] applied CNNs to diverse datasets, yielding significant results. Hussain et al. [33] proposed several DCNN models with high accuracy. Kang et al. [34] combined Raman spectroscopy with a novel hierarchical CNN (H-CNN), surpassing traditional methods for classifying cancer stages. Pacal et al. [35] achieved state-of-the-art results using Vision Transformers (ViT) and CNNs with data augmentation and ensembles. In contrast, Pramanik et al. [36] introduced a fuzzy distance-based ensemble method. Gangrade et al. [2] specifically used an ensemble of CNN, AlexNet [26], and SqueezeNet [37] on the SIPaKMeD dataset [11], achieving 94% accuracy. Integrating multiple models through ensemble learning, as noted by Solanki et al. [38], consistently shows potential for higher accuracy and robustness compared to single models. The success of transfer learning with diverse architectures (DenseNet, AlexNet, VGG, ViT) and the proven benefit of ensemble methods, as demonstrated by studies like Gangrade et al. [2], highlight their utility in achieving high accuracy, which inspired us to adopt these techniques by fine-tuning multiple state-of-the-art models and combining them using a weighted ensemble. However, a significant research gap remains: while ensemble methods improve accuracy, they often result in large, computationally expensive models, posing challenges for deployment, especially in resource-limited settings. Furthermore, achieving near-perfect accuracy while ensuring robust generalization from limited data continues to be an open challenge, as highlighted by reviews like Youneszade et al. [39].

C. Model Compression for Memory Constrained Devices

To address the challenge of large model sizes resulting from techniques like ensembling, Knowledge Distillation (KD) [16] offers a compelling compression technique. It trains a smaller 'student' network to mimic the outputs (soft targets) of a larger 'teacher' network, transferring knowledge beyond ground truth labels. KD has been successfully applied in various fields, including medical imaging [40], enabling the development of efficient yet powerful models suitable for deployment on memory-constrained devices.

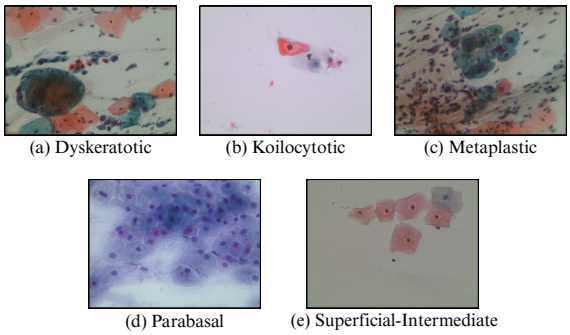
(a) Dyskeratotic (b) Koilocytotic (c) Metaplastic
(d) Parabasal (e) Superficial-Intermediate

FIGURE 2: Representative samples from each class in the SIPaKMeD dataset.

The principles of knowledge distillation [16], which provide a clear pathway to model compression, inspire our use of the Teacher-Assistant Knowledge Distillation (TAKD) framework [15]. Despite the successes of KD, a research gap exists in its specific application within a Teacher-Assistant structure for cervical cell classification. There is a need to investigate how this hierarchical distillation approach can be optimized to create models that are highly accurate, exceptionally efficient, and suitable for deployment on memory-constrained devices in clinical settings.

III. METHODOLOGY

Problem Overview: The primary problem addressed in this study is the automated classification of individual cervical squamous cells from Pap smear images into multiple predefined cytological categories, each representing a distinct stage or type of cellular morphology relevant to cervical cancer screening. Given an input image I of a cervical cell, the objective is to develop a model f that accurately assigns I to one of k classes, $C = \{c_1, c_2, \cdots c_k\}$. The goal is to maximize classification accuracy while considering model efficiency (e.g., parameter count, inference time) and interpretability. This ensures the solution is suitable for clinical application, particularly in resource-constrained environments. This task addresses challenges such as inter-class similarity, intra-class variability, and the need for transparent and explainable decision-making in a critical diagnostic context.

A. Dataset

We utilized the publicly available SIPaKMeD dataset [11] to classify cervical squamous cells from Pap smear images. The dataset comprises 4,049 digitized conventional Pap smear images, which expert pathologists manually cropped. These images are categorized into five classes: Dyskeratotic, Koilocytotic, Metaplastic, Parabasal, and Superficial-Intermediate. Each class represents a specific squamous cell type, corresponding to different stages of cervical cancer severity. The dataset is approximately balanced, with no significant class imbalance, ensuring fair representation across all cell types.

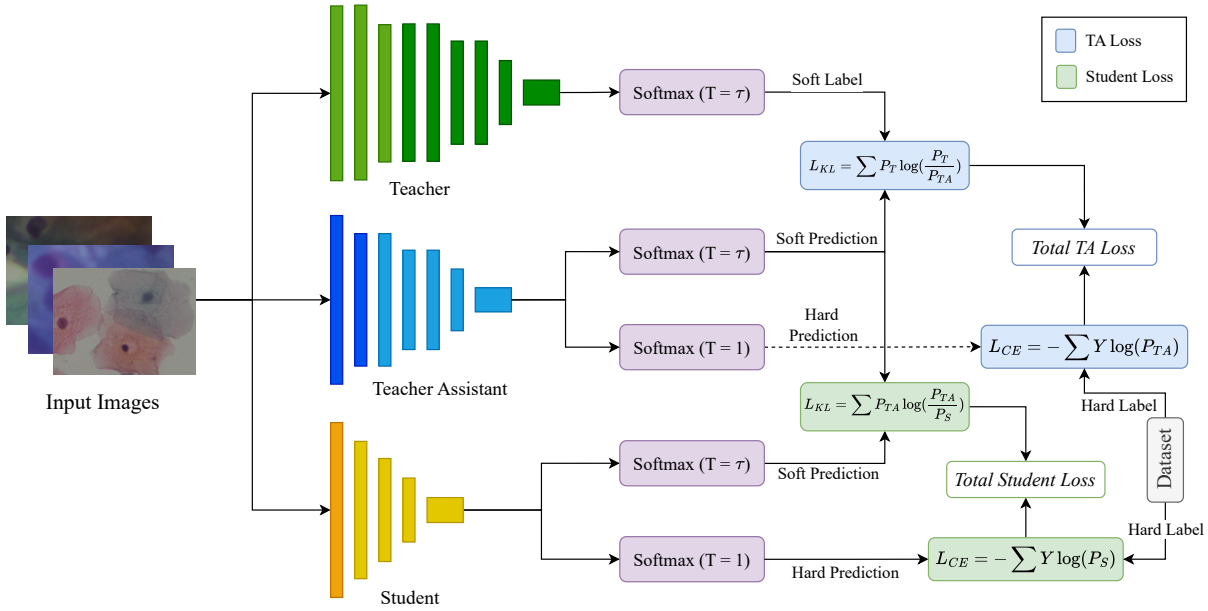

FIGURE 3: Overview of the Teacher-Assistant Knowledge Distillation (TAKD) methodology employed in this study. Using standard knowledge distillation, the large teacher model first transfers knowledge to the intermediate Teacher Assistant (TA) model. Subsequently, the lightweight student model learns from the TA model through a second distillation stage, combining cross-entropy and KL-divergence losses.

Figure 2 presents representative samples from each of the five classes in the SIPaKMeD dataset.

Several image pre-processing steps are applied to the dataset. Following pre-processing steps similar to Gangrade et al. [2], all images are resized to a uniform size of 64 × 64 pixels. Random horizontal flipping with a probability of 0.5 is applied to the training images to improve model generalization and reduce overfitting. This data augmentation increased the dataset size to 5,015 labeled images. All images are then converted to PyTorch tensors and normalized using the ImageNet mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225], as required by the pre-trained CNN models. Finally, the dataset (including the augmented samples) is divided into three subsets: training, validation, and test sets, following a ratio of 7:1.5:1.5, respectively. This results in 3,510 images for training, 752 for validation, and 753 for testing. The validation set is used for hyperparameter tuning and selecting the best model during training, while the test set is reserved for the final model evaluation.

B. Revisiting Knowledge Distillation

Knowledge Distillation (KD), a practical model compression technique introduced by Hinton et al. [16], aims to transfer the knowledge from a large, complex model (referred to as the teacher) to a smaller, simpler model (referred to as the student). This approach enables the student model to approximate the teacher's performance while significantly reducing model size and computational cost. It is particularly valuable for deployment in resource-constrained environments such as mobile devices or clinical settings.

In the KD framework, the student model is not trained solely on the ground truth hard labels provided in the dataset (i.e., one-hot encoded class labels), but also on the soft labels or dark knowledge produced by the teacher model. These soft labels are derived from the teacher's predicted class probabilities, which contain valuable information about the inter-class relationships and uncertainty patterns not captured by the hard labels alone. As a result, the student model can learn richer and more nuanced feature representations, often achieving better generalization and performance compared to training solely on hard labels. The loss function used in knowledge distillation typically combines two components:

1) Hard label loss ($\mathcal{L}_{Hard}$): the cross-entropy loss between the student's predictions P_s and the ground truth labels y.

$$\mathcal{L}_{Hard} = CrossEntropy\ (y, P_s) \tag{1}$$

2) Soft label loss ($\mathcal{L}_{Soft}$): the Kullback–Leibler (KL) divergence between the softened output distributions of the teacher P'_t and student models P'_s. This is calculated after applying a temperature parameter $\mathcal{T}$ (where $\mathcal{T} > 1$) to the logits before the softmax, which smooths the probability distribution and highlights inter-class similarities.

$$\mathcal{L}_{Soft} = KL\ Divergence\ (P'_t, P'_s) \tag{2}$$

The overall distillation loss $\mathcal{L}_{KD}$ is formulated as:

$$\mathcal{L}_{KD} = \alpha \mathcal{L}_{Hard} + (1 - \alpha)\mathcal{L}_{Soft} \tag{3}$$

where α is a weighting factor.

C. Teacher-Assistant Knowledge Distillation

Teacher-Assistant Knowledge Distillation (TAKD), proposed by Mirzadeh et al. [15], extends the standard knowledge distillation (KD) framework [16] by introducing an intermediate model—the Teacher Assistant (TA)—between the large teacher and the small student model. The primary advantage of TAKD over vanilla KD lies in its ability to effectively bridge the substantial knowledge gap that often exists between a very large teacher and a much smaller student network. This staged distillation process leads to improved performance for the student model, as the size gap between the TA and the student is considerably smaller compared to the direct gap between the teacher and the student.

Figure 3 illustrates the overall TAKD framework employed in our study. Initially, the Teacher Assistant (TA) model is trained by distilling knowledge from the large teacher model using the standard knowledge distillation approach, as described in Section B. Subsequently, the lightweight student model is trained similarly, but instead of learning directly from the teacher, it distills knowledge from the TA model. During the student model's training, two distinct loss components are computed: the cross-entropy loss ($\mathcal{L}_{CE}$) and the KL-divergence loss ($\mathcal{L}_{KL}$). The cross-entropy loss is calculated between the ground-truth labels (Y) and the student model's predictions (P_S). In contrast, the KL-divergence loss is computed between the softened predictions of the TA model (P_{TA}) and those of the student model (P_S). To obtain the softened predictions, a temperature $\mathcal{T}$ is applied to the softmax activation function, instead of the standard temperature value of 1.

Finally, the total loss is computed as the weighted sum of the two losses, with a weighting factor α, a key hyperparameter in knowledge distillation:

$$Total\ Loss = \alpha\mathcal{L}_{CE} + (1-\alpha)\mathcal{L}_{KL} \tag{4}$$

This hierarchical distillation pipeline enables the highly compact student model to effectively learn from the distilled knowledge passed down from the large teacher model via the intermediate teacher assistant (TA) model, thereby achieving computational efficiency while maintaining high classification accuracy. Algorithm 1 summarizes the overall pipeline of the employed TAKD approach.

Algorithm 1 Teacher-Assistant Knowledge Distillation

Require: Dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$. Teacher Model M_T. Teacher Assistant Model M_{TA}. Student Model M_S. Distillation Temperature $\mathcal{T}$. Loss weighting coefficient α. Number of training epochs E.

procedure TRAINTA($M_T, M_{TA}, D, \mathcal{T}, \alpha$)
  for epoch = 1 to E **do**
    for each batch (x, y) in D **do**
      $logits_T \leftarrow M_T(x)$
      $logits_{TA} \leftarrow M_{TA}(x)$
      $p'_T \leftarrow \text{softmax}(logits_T/\mathcal{T})$
      $p'_{TA\_log} \leftarrow \text{log\_softmax}(logits_{TA}/\mathcal{T})$
      $p_{TA} \leftarrow \text{softmax}(logits_{TA})$
      $L_{CE\_TA} \leftarrow \text{CrossEntropyLoss}(p_{TA}, y)$
      $L_{KL\_TA} \leftarrow \mathcal{T}^2 \cdot \text{KLDivLoss}(p'_{TA\_log}, p'_T)$
      $L_{TA} \leftarrow \alpha \cdot L_{CE\_TA} + (1-\alpha) \cdot L_{KL\_TA}$
      Update parameters of M_{TA}
    end for
  end for
  return Trained M^*_{TA}
end procedure

procedure TRAINSTUDENT($M^*_{TA}, M_S, D, \mathcal{T}, \alpha$)
  for epoch = 1 to E **do**
    for each batch (x, y) in D **do**
      $logits^*_{TA} \leftarrow M^*_{TA}(x)$
      $logits_S \leftarrow M_S(x)$
      $p'_{TA\_final} \leftarrow \text{softmax}(logits^*_{TA}/\mathcal{T})$
      $p'_{S\_log} \leftarrow \text{log\_softmax}(logits_S/\mathcal{T})$
      $p_S \leftarrow \text{softmax}(logits_S)$
      $L_{CE\_S} \leftarrow \text{CrossEntropyLoss}(p_S, y)$
      $L_{KL\_S} \leftarrow \mathcal{T}^2 \cdot \text{KLDivLoss}(p'_{S\_log}, p'_{TA\_final})$
      $L_S \leftarrow \alpha \cdot L_{CE\_S} + (1-\alpha) \cdot L_{KL\_S}$
      Update parameters of M_S
    end for
  end for
  return Trained M^*_S
end procedure

Initialize M_T, M_{TA}, M_S and their respective optimizers (with pre-configured learning rates)
$M^*_{TA} \leftarrow$ TRAINTA($M_T, M_{TA}, D, \mathcal{T}, \alpha$)
$M^*_S \leftarrow$ TRAINSTUDENT($M^*_{TA}, M_S, D, \mathcal{T}, \alpha$)
output: Trained Student Model M^*_S

D. Model Architectures

This section details the architectures of the teacher, teacher assistant (TA), and student models employed in our framework.

Teacher Model: The teacher model is designed as a high-capacity, high-performance system that is the primary knowledge source for the entire distillation process. We first fine-tuned several pre-trained Convolutional Neural Networks (CNNs) of varying sizes and architectures. Specifically, 11 distinct models were employed, including AlexNet, ResNet (ResNet-50, ResNet-101, ResNet-152) [41], VGG (VGG-16, VGG-19), DenseNet (DenseNet-169, DenseNet-201), EfficientNet-B3, ConvNeXt-Tiny, and RegNetY-16GF.

These fine-tuned models were then aggregated using a weighted averaging ensemble approach. The weight w_i assigned to each constituent model M_i was calculated proportionally to its individual test accuracy (acc_i) on the SIPaKMeD dataset:

$$w_i = \frac{acc_i}{\sum_{j=1}^{11} acc_j} \tag{5}$$

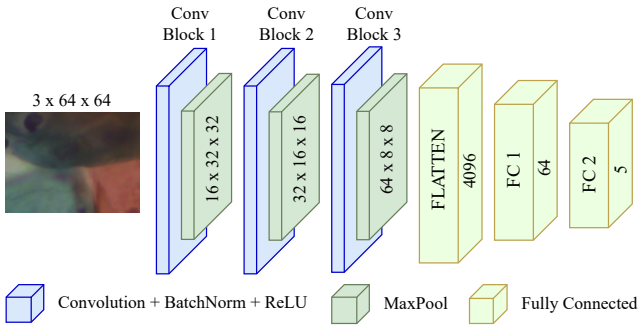

FIGURE 4: Architecture of our proposed CNN model, AgileNet.

The final ensemble prediction $P_T(x)$ was computed as the weighted average of the individual model probabilities P_i:

$$P_T(x) = \sum_{i=1}^{11} w_i \; P_i(x) \tag{6}$$

where x represents an input image.

The resulting ensemble model is a large, high-performing system with approximately 604.78 million parameters, achieving a classification accuracy of 98.01% and an F1 score of 97.96%. Due to its strong classification performance on cervical cancer detection, this ensemble was selected as the teacher model in our employed TAKD framework.

Teacher Assistant Model: In the Teacher-Assistant Knowledge Distillation (TAKD) framework, the Teacher Assistant (TA) model plays a crucial intermediary role, bridging the capacity and complexity gap between the large, high-capacity teacher model and the smaller, lightweight student model. This staged distillation approach facilitates more effective and efficient knowledge transfer, improving the final performance of the student model. For the TA model, we selected SqueezeNet-v1.1 [37], a well-established lightweight architecture known for its favorable balance between model complexity and classification performance. After adapting the final layers to handle five target classes, the modified SqueezeNet architecture consisted of approximately 0.74 million trainable parameters. This adapted SqueezeNet-v1.1 model was then used as the Teacher Assistant, which was trained through knowledge distillation using the outputs from the ensemble teacher model.

Student Model: The student model, named AgileNet, is a custom-designed, lightweight Convolutional Neural Network (CNN) architecture optimized for computational efficiency while maintaining strong performance on the cervical cell classification task. With only 0.29 million parameters, AgileNet is highly suitable for deployment on resource-constrained edge devices. The architecture of AgileNet, illustrated in Figure 4, processes a $3 \times 64 \times 64$ input image through a series of convolutional and fully connected layers.

AgileNet consists of three main convolutional blocks, each followed by batch normalization [42] to stabilize training and a Rectified Linear Unit (ReLU) activation function [43] for non-linearity. Each block utilizes a 3×3 convolutional layer with a padding of 1 and no bias term, followed by a 2×2 max pooling layer for spatial downsampling. After the convolutional blocks, the resulting feature map is flattened into a 4096-dimensional feature vector. This vector is then passed through a fully connected layer, which reduces the dimensionality to 64 before mapping it to the final output layer with five neurons, corresponding to the five target classes in the dataset.

TABLE 1: Key hyperparameters used in the training process.

Hyperparameter	Value
Temperature (for TAKD)	4
Alpha (for TAKD)	0.5
Optimizer	Adam
Learning Rate	1×10^{-3}
Weight Decay	1×10^{-4}
Batch Size	64
Patience (Early Stopping)	30

IV. EXPERIMENTS

A. Setup

Implementation Details: We developed and rigorously evaluated a multi-stage training framework, Teacher-Assistant Knowledge Distillation (TAKD), tailored explicitly for cervical squamous cell classification. Within this framework, all models are optimized using the cross-entropy loss function and incorporate an early stopping mechanism with a patience parameter of 30. Specifically, the training process is automatically halted if the validation loss does not improve over 30 consecutive epochs, preventing overfitting and reducing unnecessary computation. In addition, we employed the *ReduceLROnPlateau* learning rate scheduler, which adaptively reduces the learning rate whenever the validation loss plateaus, enabling more refined convergence during later training stages. Table 1 summarizes the key hyperparameters used across all experiments, including batch size, optimizer [44], learning rate [45], weight decay [46], and the principal knowledge distillation parameters ($\mathcal{T}$: temperature, and α: weighting coefficient). All experiments were implemented using the *PyTorch* framework [47] and executed on a system equipped with two NVIDIA Tesla T4 GPUs.

Evaluation Metrics: To assess the performance of our proposed model, AgileNet, we evaluated it on an unseen test set containing 753 images. We employed standard classification metrics — accuracy, precision, recall, and F1-score — to comprehensively report the model's performance. Additionally, we reported the 95% confidence intervals for accuracy and F1-score to further validate the results' reliability.

Let TP (true positives), TN (true negatives), FP (false positives), and FN (false negatives) represent the counts in the confusion matrix.

TABLE 2: Performance of the baseline CNN models after transfer learning, the ensemble teacher model, the teacher assistant (TA), and the proposed AgileNet. The ensemble refers to the aggregation of 11 pre-trained CNNs using weighted averaging. Inference time is the average per sample.

	Model	Parameter	Inference Time	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
Pre-trained	AlexNet	57.02M	87 µs	96.68	96.66	96.60	96.61
	ResNet-50	23.52M	550 µs	96.41	96.34	96.43	96.38
	ResNet-101	42.51M	918 µs	96.28	96.26	96.27	96.26
	ResNet-152	58.15M	921 µs	96.15	96.11	96.11	96.11
	VGG-16	134.28M	664 µs	96.28	96.33	96.20	96.23
	VGG-19	139.59M	774 µs	96.68	96.65	96.64	96.64
	DenseNet-169	12.49M	733 µs	96.81	96.77	96.79	96.77
	DenseNet-201	18.10M	963 µs	96.02	96.03	95.93	95.97
	EfficientNet-B3	10.70M	380 µs	92.43	92.43	92.30	92.33
	ConvNeXt-Tiny	27.82M	295 µs	97.74	97.79	97.70	97.74
	RegNetY-16GF	80.58M	1055 µs	95.88	95.87	95.83	95.84
Teacher	Ensemble	604.78M	5709 µs	98.01	98.00	97.93	97.96
TA	SqueezeNet	0.74M	75 µs	91.90	91.94	91.81	91.86
TA	SqueezeNet (with KD)	0.74M	66 µs	97.21	97.29	97.16	97.21
Student	AgileNet	0.29M	41 µs	96.95	96.95	96.90	96.92
Student	AgileNet (with TAKD)	0.29M	35 µs	97.88	97.88	97.88	97.88

1) **Accuracy** measures the overall proportion of correctly classified instances out of all predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

2) **Precision** quantifies the proportion of true positive predictions among all positive predictions made.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

3) **Recall** (or sensitivity) captures the proportion of true positives correctly identified out of all actual positive cases.

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

4) **F1-score** provides the harmonic mean of precision and recall, offering a balanced metric that accounts for false positives and false negatives.

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

B. Results and Analysis

This section presents the experimental results of our proposed framework and comparative models. Table 2 summarizes the comprehensive performance scores, parameter count, and inference time for all individual pre-trained models, the ensemble teacher model, the teacher assistant (TA), and our proposed student model, AgileNet. This table provides a detailed overview of how each model performed under the specified experimental conditions.

From Table 2, the observations are as follows: ***(1)*** Ensembling the 11 fine-tuned models yielded the highest classification accuracy of 98.01%, which motivated its selection as the high-performance teacher model for the subsequent knowledge distillation process. ***(2)*** Among the pre-trained models, SqueezeNet, with just 0.74 million parameters, is the lightest and achieved 91.90% accuracy after fine-tuning. Notably, when enhanced with knowledge distillation from the ensemble teacher, its performance significantly improved to 97.21%, making it an ideal assistant model (TA) within the teacher-assistant knowledge distillation framework (TAKD). ***(3)*** The proposed student model, AgileNet, trained with TAKD, achieved an impressive accuracy of 97.88% and an F1-score of 97.88%, using only 0.29 million parameters. This ranks as the second-best performance overall, surpassed only by the much larger teacher ensemble, which comprises 604.78 million parameters. ***(4)*** When comparing AgileNet with and without TAKD, the baseline AgileNet (trained conventionally) achieved 96.95% accuracy. The application of TAKD effectively transferred knowledge from the teacher and assistant models, substantially improving classification performance. ***(5)*** In terms of computational efficiency, AgileNet demonstrated the lowest parameter count and the fastest inference time, substantially outperforming all other models in terms of lightweight design and speed. ***(6)*** Overall, while the teacher ensemble achieves the highest accuracy, its massive size and long inference time make it impractical for real-world clinical deployment, particularly in low-resource settings. In contrast, the distilled AgileNet model offers an excellent balance, achieving near-teacher-level performance while remaining extremely lightweight and efficient. This makes AgileNet highly suitable for deployment on low-memory devices, underscoring its potential for practical use.

TABLE 3: Performance variability within the 95% confidence interval for the teacher, teacher assistant (TA), and student models.

	Model	Accuracy(%)	F1(%)
Teacher	Ensemble	96.95 - 98.94	96.88 - 98.91
TA	SqueezeNet (with KD)	96.02 - 98.41	95.95 - 98.40
Student	AgileNet (with TAKD)	96.87 - 98.84	96.86 - 98.83

TABLE 4: Class-wise classification report for the proposed AgileNet model.

Class	Precision	Recall	F1	Support
Dyskeratotic	0.9826	0.9826	0.9826	172
Koilocytotic	0.9640	0.9640	0.9640	139
Metaplastic	0.9615	0.9740	0.9677	154
Parabasal	0.9929	1.0000	0.9964	139
Superficial-Intermediate	0.9932	0.9732	0.9831	149
Macro Average	0.9788	0.9788	0.9788	753
Weighted Average	0.9788	0.9788	0.9788	753

To enhance the reliability and trustworthiness of our findings, we calculated the 95% confidence intervals for the employed teacher, teacher assistant (TA), and student models, as presented in Table 3. These intervals provide a range within which the actual performance of the models is likely to lie. Furthermore, Table 4 presents the detailed class-wise classification report for our proposed distilled AgileNet model. It can be observed that across all classes and metrics, the AgileNet model consistently achieved high and balanced scores, demonstrating its robustness in distinguishing between different cervical cell types. In addition, Figure 5 illustrates the training and validation accuracy and loss curves of the distilled assistant model (TA) and the student model across epochs during the distillation process, providing insight into their learning behavior. Finally, the confusion matrices for the teacher ensemble, the TA model, and the distilled student model (AgileNet) are presented in Figure 6 to assess their classification performance across the various classes visually.

A comparative analysis of our proposed approach against recent notable studies in cervical cancer screening is presented in Table 5. This comparison demonstrates that our distilled AgileNet model achieves the highest accuracy and has the lowest parameter count (0.29 million). This significantly contributes to the cervical cytopathology domain by providing a highly accurate yet computationally efficient solution.

Explainable AI: To enhance the reliability and trustworthiness of the proposed AgileNet model, we employed explainable AI (XAI) techniques, specifically Grad-CAM++ [17]. This approach was applied to gain deeper insights into

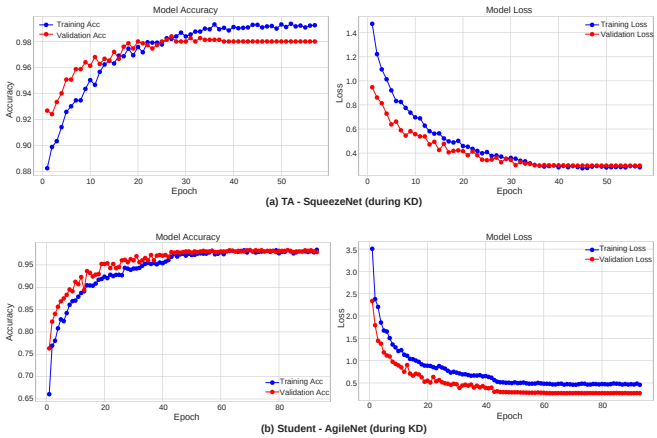

FIGURE 5: Training and validation accuracy and loss curves for both the teacher assistant (TA) model and the student model.

TABLE 5: Performance comparison of AgileNet with existing studies on cervical cancer screening.

Year	Reference	Model	Param	Accuracy
2020	Win et al. [48]	Ensemble	-	94.09%
2021	Manna et al. [49]	Ensemble	59.85M	95.43%
2022	Li et al. [50]	VGG16 w/ SENET	-	96.57%
2023	Alsubai et al. [51]	CNN	14.87M	91.13%
2023	Neerukonda et al. [52]	InceptionResNetV2	55.87M	92.00%
2023	Sahoo et al. [53]	Ensemble	112.15M	97.18%
2024	Wubineh et al. [54]	DenseNet121 w/ SA	10.17M	92.00%
2024	Al-asbaily et al. [55]	CNN w/ PCA	-	94.20%
2024	Joynab et al. [56]	CNN-based FL	0.42M	94.36%
2024	Fang et al. [57]	DIFF w/ CNN	52.04M	96.02%
2024	Alzahrani et al. [58]	ViT	-	97.54%
2025	Gangrade et al. [2]	Ensemble	23.21M	94.00%
2025	This work	AgileNet	**0.29M**	**97.88%**

the student model's decision-making process following the knowledge distillation phase. Visual explanations were generated for correctly classified samples selected from the test set, with one representative example shown for each class in Figure 7.

The resulting heatmaps predominantly highlight key cellular regions, particularly the nuclei and abnormal cytoplasmic areas, which align well with established cytopathological diagnostic features [59]. These visualizations provide strong evidence that AgileNet effectively focuses on clinically relevant regions when making predictions, thereby increasing confidence in the model's outputs. Additionally, we present the probability scores for each class alongside these visualizations to further validate the model's performance and enhance its interpretability. Overall, these XAI results demonstrate the valuable role of explainability in verifying and building trust in distilled deep learning models.

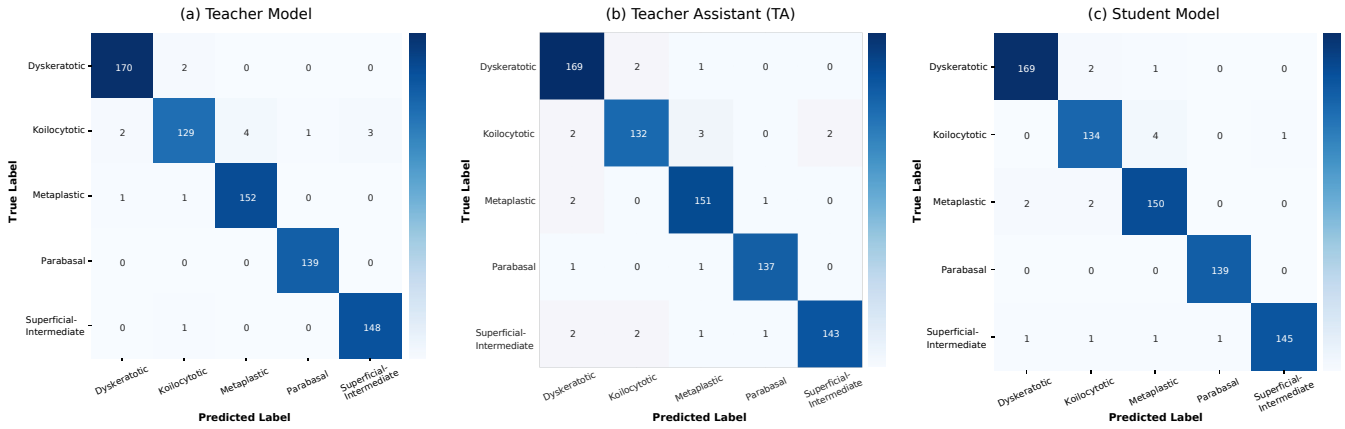

FIGURE 6: Confusion matrices showing the test set performance of the **(a)** teacher model, **(b)** teacher assistant (TA) model, and **(c)** student model.

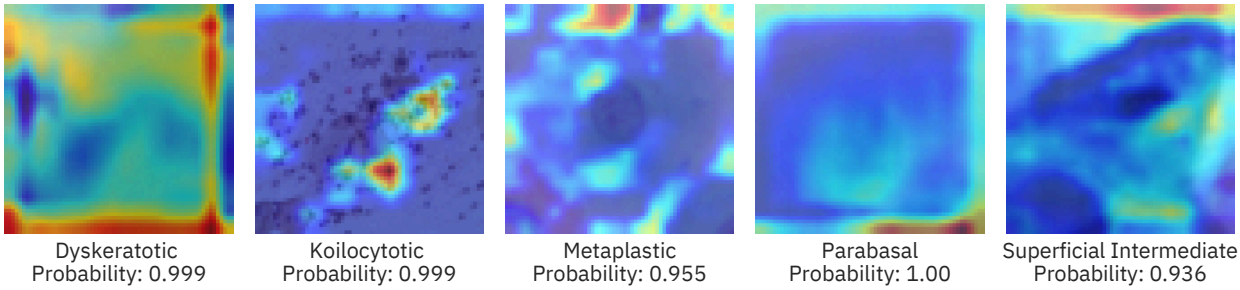

FIGURE 7: Explainable AI (XAI) visualizations using Grad-CAM++ for AgileNet predictions on correctly classified test images, following Teacher-Assistant Knowledge Distillation (TAKD).

C. Ablation Study

We conducted a series of ablation experiments to assess the contribution of key components within our proposed framework. These experiments specifically focused on three critical aspects: **(a)** the impact of conventional knowledge distillation versus Teacher-Assistant Knowledge Distillation (TAKD) on the performance of AgileNet; **(b)** the influence of employing different teacher models; and **(c)** the sensitivity of AgileNet to key distillation hyperparameters, namely the temperature ($\mathcal{T}$) and the weighting coefficient (α).

Impact of KD and TAKD: We first evaluated AgileNet's performance under two distinct training regimes: conventional Knowledge Distillation (KD) and the Teacher-Assistant Knowledge Distillation (TAKD) framework, incorporating an intermediate assistant model. As shown in Figure 8(a), AgileNet achieves a notably higher accuracy and F1 score when trained using the TAKD pipeline than it does under conventional KD. These results underscore the effectiveness of integrating an intermediate assistant model, facilitating more efficient and robust knowledge transfer from the teacher to the student.

Impact of different teacher models on AgileNet: To examine how the choice of teacher model influences the performance of the final student model, we explored the effect of varying the teacher architecture. Specifically, we compared two settings: one where the teacher was a comprehensive 11-model ensemble and another where the teacher was a single high-performing pre-trained model (ConvNeXt-Tiny). As illustrated in Figure 8(b), we observe that employing a stronger, higher-performing teacher consistently leads to developing a superior student model. Notably, the ensemble outperforms the single ConvNeXt-Tiny model's performance and ability to guide the student. ConvNeXt-Tiny was selected as the single-teacher baseline because it achieved the highest performance among all individual models evaluated.

Impact of varying alpha and temperature values: The effectiveness of knowledge distillation is highly sensitive to its key hyperparameters, particularly the distillation temperature ($\mathcal{T}$) and the loss weighting factor (α). We conducted experiments to systematically evaluate their influence using three distinct α values (0.3, 0.5, 0.7) and four temperature settings (1, 2, 4, 8). The student model, AgileNet, was trained under the TAKD framework for 10 epochs across these combinations to identify the optimal configuration. As illustrated in Figure 8(c), the highest 10-epoch accuracy was achieved when the temperature was set to 4 and α was set to 0.5. Notably, these results reflect early-stage training (after

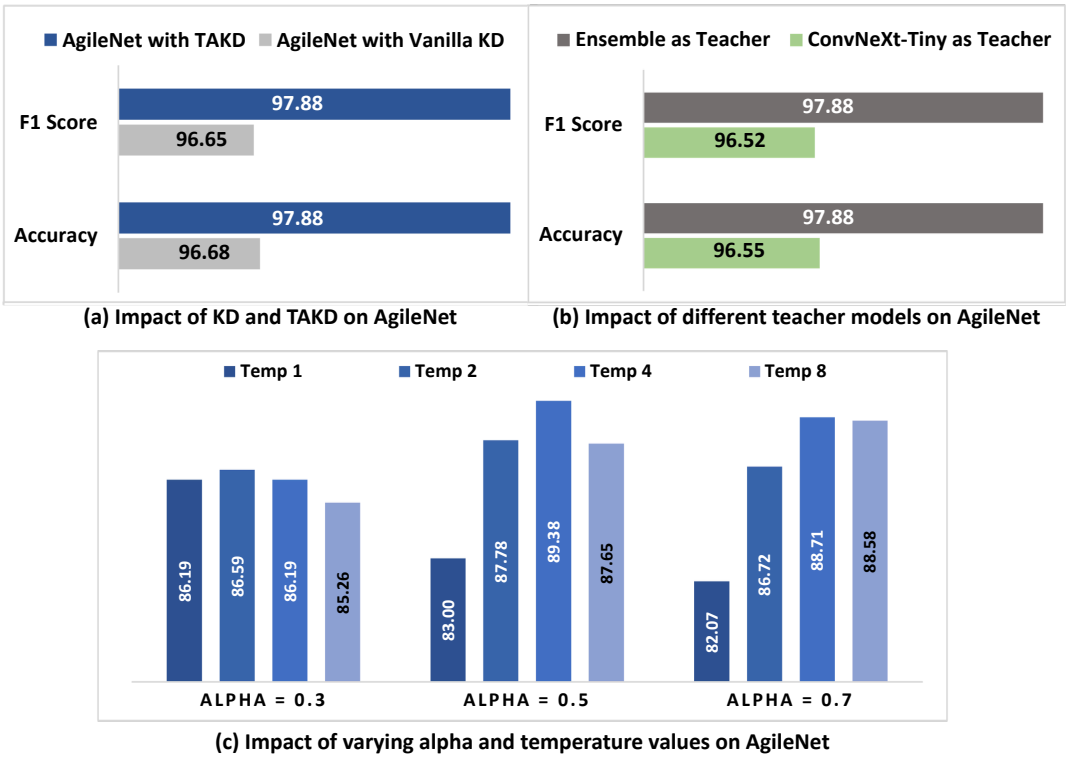

FIGURE 8: The Impact of KD and TAKD on AgileNet's Accuracy.

10 epochs) and should not be interpreted as the model's fully converged performance.

D. Discussion

This study addresses the pressing need for automated diagnostic tools in cervical cytopathology that are highly accurate, computationally efficient, and transparent—key requirements for real-world clinical adoption. Conventional deep learning models often face a trade-off between performance and practicality, achieving high accuracy at the cost of large, resource-intensive architectures that limit their deployment, particularly in resource-constrained settings. Our work presents a comprehensive framework designed to overcome these challenges by synergistically integrating multiple well-established and novel techniques. Central to this contribution is the strategic implementation of a Teacher-Assistant Knowledge Distillation (TAKD) pipeline, which enables the rich diagnostic knowledge embedded in a large, high-performing ensemble to be effectively distilled into AgileNet—a custom-designed, lightweight student model with significantly fewer parameters—through the use of an intermediate assistant model.

Moreover, integrating Grad-CAM++ for explainability directly addresses the crucial demand for transparency in clinical AI, allowing medical experts to visualize and interpret the model's decision-making process, thereby enhancing trust and supporting the validation of diagnostic outputs. Experimental results underscore the strength of our approach, demonstrating that AgileNet achieves an impressive balance between performance and efficiency while providing interpretable visual rationales for its predictions. Collectively, these outcomes pave the way for more accessible, reliable, and trustworthy AI-driven cervical cancer screening solutions, particularly in settings with limited computational resources.

V. CONCLUSION

This study presents an explainable deep learning framework for classifying cervical squamous cells from Pap smear images, leveraging ensemble learning, teacher-assistant knowledge distillation (TAKD), and explainable AI (XAI) techniques—specifically, Grad-CAM++. By incorporating TAKD, we developed a highly lightweight model, AgileNet, containing only 0.29 million parameters. Despite its compact size, AgileNet achieves remarkable classification performance, attaining 97.88% accuracy, reduced parameter count, and faster inference time. Notably, the proposed model outperforms existing state-of-the-art approaches, including those fine-tuned explicitly for cervical cancer screening. Overall, the framework delivers high precision and accuracy while maintaining suitability for deployment on memory-constrained edge devices. Furthermore, integrating XAI enhances the interpretability and trustworthiness of the model's predictions, offering meaningful support to clinical experts in diagnostic settings.

Future Work: The promising results of this study open several avenues for future investigation. Future efforts should prioritize the rigorous validation of AgileNet on larger, more diverse, multi-center datasets, addressing real-world challenges such as class imbalance, variations in staining protocols, and differences in image acquisition equipment. Additionally, more sophisticated knowledge distillation strategies could be explored to enhance efficiency further, while reducing reliance on large teacher ensembles would make the training pipeline more accessible in low-resource settings. Beyond Grad-CAM++, applying complementary explainable AI (XAI) techniques could further strengthen the interpretability and trustworthiness of the diagnostic outputs. Finally, incorporating additional data modalities—such as patient clinical history, HPV status, or other relevant biomarkers—alongside image-based classification holds promise for developing a more comprehensive, multimodal diagnostic support system.

DATA AVAILABILITY

The dataset employed for this study is a publicly available dataset. It can be accessed through https://www.cs.uoi.gr/~marina/sipakmed.html.

REFERENCES

[1] National Cancer Institute, "Cervical cancer." https://www.cancer.gov/types/cervical, 2025. Last Accessed: April 19, 2025.

[2] J. Gangrade, R. Kuthiala, S. Gangrade, Y. P. Singh, M. R, and S. Solanki, "A deep ensemble learning approach for squamous cell classification in cervical cancer," *Scientific Reports*, vol. 15, p. 7266, Mar 2025.

[3] R. Hull, M. Mbele, T. Makhafola, C. Hicks, S. Wang, R. Reis, R. Mehrotra, Z. Mkhize-Kwitshana, G. Kibiki, D. Bates, and Z. Dlamini, "Cervical cancer in low and middle-income countries (Review)," *Oncology Letters*, vol. 20, 06 2020.

[4] A. K. Uddin, M. A. Sumon, S. Pervin, and F. Sharmin, "Cervical cancer in Bangladesh," *South Asian Journal of Cancer*, vol. 12, no. 01, pp. 036–038, 2023.

[5] K. Canfell, J. J. Kim, M. Brisson, A. Keane, K. T. Simms, M. Caruana, E. A. Burger, D. Martin, D. T. Nguyen, É. Bénard, *et al.*, "Mortality impact of achieving WHO cervical cancer elimination targets: a comparative modelling analysis in 78 low-income and lower-middle-income countries," *The Lancet*, vol. 395, no. 10224, pp. 591–603, 2020.

[6] O. E. Aina, S. A. Adeshina, and A. Aibinu, "Classification of cervix types using convolutional neural network (CNN)," in *2019 15th IEEE International Conference on Electronics, Computer and Computation (ICECCO)*, pp. 1–4, 2019.

[7] M. M. Patel, A. N. Pandya, and J. Modi, "Cervical Pap smear study and its utility in cancer screening, to specify the strategy for cervical cancer control," *National Journal of Community Medicine*, vol. 2, p. 49–51, Jun. 2011.

[8] P. L. Sachan, M. Singh, M. L. Patel, and R. Sachan, "A study on cervical cancer screening using Pap smear test and clinical correlation," *Asia-Pacific Journal of Oncology Nursing*, vol. 5, no. 3, pp. 337–341, 2018.

[9] A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, and M. A. Azim, "Transfer learning: a friendly introduction," *Journal of Big Data*, vol. 9, no. 1, p. 102, 2022.

[10] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *CoRR*, vol. abs/1511.08458, 2015.

[11] M. E. Plissiti, P. Dimitrakopoulos, G. Sfikas, C. Nikou, O. Krikoni, and A. Charchanti, "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, 2018.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[13] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Tech. Rep. 0, University of Toronto, Toronto, Ontario, 2009.

[14] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari, "The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale," *International Journal of Computer Vision*, vol. 128, p. 1956–1981, Mar. 2020.

[15] S. Mirzadeh, M. Farajtabar, A. Li, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher," *CoRR*, vol. abs/1902.03393, 2019.

[16] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

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

[18] L. J. Mango, "Computer-assisted cervical cancer screening using neural networks," *Cancer Letters*, vol. 77, no. 2, pp. 155–162, 1994. Computer applications for early detection and staging of cancer.

[19] S. Ponnusamy and R. Gnanamurthy, "Computer aided detection of cervical cancer using Pap smear images based on adaptive neuro fuzzy inference system classifier," *Journal of Medical Imaging and Health Informatics*, vol. 6, pp. 312–319, 04 2016.

[20] K. Bora, M. Chowdhury, L. B. Mahanta, M. K. Kundu, and A. K. Das, "Pap smear image classification using convolutional neural network," in *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing*, ICVGIP '16, (New York, NY, USA), Association for Computing Machinery, 2016.

[21] J. Hyeon, H. Choi, K. Lee, and B. Lee, "Automating Papanicolaou test using deep convolutional activation feature," in *Proceedings - 18th IEEE International Conference on Mobile Data Management, MDM 2017*, Proceedings - 18th IEEE International Conference on Mobile Data Management, MDM 2017, pp. 382–385, Institute of Electrical and Electronics Engineers Inc., June 2017. Publisher Copyright: © 2017 IEEE.; 18th IEEE International Conference on Mobile Data Management, MDM 2017 ; Conference date: 29-05-2017 Through 01-06-2017.

[22] P. Sanyal, P. Ganguli, and S. Barui, "Performance characteristics of an artificial intelligence based on convolutional neural network for screening conventional Papanicolaou-stained cervical smears," *Medical Journal Armed Forces India*, vol. 76, no. 4, pp. 418–424, 2020.

[23] V. Karunakaran, V. N. Saritha, M. M. Joseph, J. B. Nair, G. Saranya, K. G. Raghu, K. Sujathan, K. S. Kumar, and K. K. Maiti, "Diagnostic spectro-cytology revealing differential recognition of cervical cancer lesions by label-free surface enhanced raman fingerprints and chemometrics," *Nanomedicine: Nanotechnology, Biology and Medicine*, vol. 29, p. 102276, 2020.

[24] Y. Promworn, S. Pattanasak, C. Pintavirooj, and W. Piyawattanametha, "Comparisons of Pap smear classification with deep learning models," in *2019 IEEE 14th International Conference on Nano/Micro Engineered and Molecular Systems (NEMS)*, pp. 282–285, 2019.

[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[28] B. Taha, J. Dias, and N. Werghi, "Classification of cervical-cancer using pap-smear images: A convolutional neural network approach," in *Medical Image Understanding and Analysis* (M. Valdés Hernández and V. González-Castro, eds.), (Cham), pp. 261–272, Springer International Publishing, 2017.

[29] D. Xue, X. Zhou, C. Li, Y. Yao, M. M. Rahaman, J. Zhang, H. Chen, J. Zhang, S. Qi, and H. Sun, "An application of transfer learning

and ensemble learning techniques for cervical histopathology image classification," *IEEE Access*, vol. 8, pp. 104603–104618, 2020.
[30] W. Chen, X. Li, L. Gao, and W. Shen, "Improving computer-aided cervical cells classification using transfer learning based snapshot ensemble," *Applied Sciences*, vol. 10, no. 20, 2020.
[31] A. Ghoneim, G. Muhammad, and M. S. Hossain, "Cervical cancer classification using convolutional neural networks and extreme learning machines," *Future Gener. Comput. Syst.*, vol. 102, p. 643–649, Jan. 2020.
[32] D. Arifianto and A. S. Agoes, "Cervical cancer image classification using CNN transfer learning," in *Proceedings of the 2nd International Seminar of Science and Applied Technology (ISSAT 2021)*, pp. 145–149, Atlantis Press, 2021.
[33] E. Hussain, L. B. Mahanta, C. R. Das, and R. K. Talukdar, "A comprehensive study on the multi-class cervical cancer diagnostic prediction on Pap smear images using a fusion-based decision from ensemble deep convolutional neural network," *Tissue and Cell*, vol. 65, p. 101347, 2020.
[34] Z. Kang, Y. Li, J. Liu, C. Chen, W. Wu, C. Chen, X. Lv, and F. Liang, "H-CNN combined with tissue Raman spectroscopy for cervical cancer detection," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 291, p. 122339, 2023.
[35] I. Pacal and S. Kılıcarslan, "Deep learning-based approaches for robust classification of cervical cancer," *Neural Comput. Appl.*, vol. 35, p. 18813–18828, July 2023.
[36] R. Pramanik, M. Biswas, S. Sen, L. A. d. Souza Júnior, J. a. P. Papa, and R. Sarkar, "A fuzzy distance-based ensemble of deep models for cervical cancer detection," *Comput. Methods Prog. Biomed.*, vol. 219, June 2022.
[37] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size," *CoRR*, vol. abs/1602.07360, 2016.
[38] S. Solanki, V. Dehalwar, J. Choudhary, M. L. Kolhe, and K. Ogura, "Spectrum sensing in cognitive radio using CNN-RNN and transfer learning," *IEEE Access*, vol. 10, pp. 113482–113492, 2022.
[39] N. Youneszade, M. Marjani, and C. P. Pei, "Deep learning in cervical cancer diagnosis: Architecture, opportunities, and open research challenges," *IEEE Access*, vol. 11, pp. 6133–6149, 2023.
[40] X. Xing, Y. Hou, H. Li, Y. Yuan, H. Li, and M. Q.-H. Meng, "Categorical relation-preserving contrastive knowledge distillation for medical image classification," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pp. 163–173, Springer, 2021.
[41] R. Wightman, H. Touvron, and H. Jégou, "ResNet strikes back: An improved training procedure in timm," *CoRR*, vol. abs/2110.00476, 2021.
[42] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, p. 448–456, JMLR.org, 2015.
[43] A. F. Agarap, "Deep learning using rectified linear units (ReLU)," 2019.
[44] D. Choi, C. J. Shallue, Z. Nado, J. Lee, C. J. Maddison, and G. E. Dahl, "On empirical comparisons of optimizers for deep learning," *CoRR*, vol. abs/1910.05446, 2019.
[45] Y. Wu, L. Liu, J. Bae, K. H. Chow, A. Iyengar, C. Pu, W. Wei, L. Yu, and Q. Zhang, "Demystifying learning rate polices for high accuracy training of deep neural networks," *CoRR*, vol. abs/1908.06477, 2019.
[46] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," *CoRR*, vol. abs/1711.05101, 2017.
[47] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *PyTorch: an imperative style, high-performance deep learning library*. Red Hook, NY, USA: Curran Associates Inc., 2019.
[48] K. P. Win, Y. Kitjaidure, K. Hamamoto, and T. Myo Aung, "Computer-assisted screening for cervical cancer using digital image processing of Pap smear images," *Applied Sciences*, vol. 10, no. 5, 2020.
[49] A. Manna, R. Kundu, D. Kaplun, A. Sinitca, and R. Sarkar, "A fuzzy rank-based ensemble of CNN models for classification of cervical cytology," *Scientific Reports*, vol. 11, p. 14538, Jul 2021.
[50] M. Li, A. Feng, Y. Yan, S. You, and C. Li, "Deep convolutional neural network based cervical cancer exfoliated cell detection," in *Proceedings of International Conference on Image, Vision and Intelligent Systems 2022 (ICIVIS 2022)* (P. You, H. Li, and Z. Chen, eds.), (Singapore), pp. 589–598, Springer Nature Singapore, 2023.
[51] S. Alsubai, A. Alqahtani, M. Sha, A. Almadhor, S. Abbas, H. Mughal, and M. Gregus, "Privacy preserved cervical cancer detection using convolutional neural networks applied to Pap smear images," *Computational and Mathematical Methods in Medicine*, vol. 2023, no. 1, p. 9676206, 2023.
[52] S. P. Neerukonda, "Transfer learning for cervical cancer image classification," Master's thesis, California State University, Northridge, Jun 2023.
[53] P. Sahoo, S. Saha, S. Mondal, M. Seera, S. K. Sharma, and M. Kumar, "Enhancing computer-aided cervical cancer detection using a novel fuzzy rank-based fusion," *IEEE Access*, vol. 11, pp. 145281–145294, 2023.
[54] B. Z. Wubineh, A. Rusiecki, and K. Halawa, "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*, vol. 36, pp. 21801–21815, Dec 2024.
[55] S. A. Al-asbaily, S. Almoshity, S. Younus, and K. Bozed, "Classification of cervical cancer using convolutional neural networks," in *2024 IEEE 4th International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA)*, pp. 735–739, 2024.
[56] N. S. Joynab, M. N. Islam, R. R. Aliya, A. R. Hasan, N. I. Khan, and I. H. Sarker, "A federated learning aided system for classifying cervical cancer using PAP-SMEAR images," *Informatics in Medicine Unlocked*, vol. 47, p. 101496, 2024.
[57] M. Fang, M. Fu, B. Liao, X. Lei, and F.-X. Wu, "Deep integrated fusion of local and global features for cervical cell classification," *Computers in Biology and Medicine*, vol. 171, p. 108153, 2024.
[58] M. Alzahrani, U. A. Khan, and S. Al-Garni, "Ensemble and transformer encoder-based models for the cervical cancer classification using Pap-smear images," *Journal of Electrical Systems*, vol. 20, pp. 1637–1646, 04 2024.
[59] R. Nayar and D. C. Wilbur, *The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria, and Explanatory Notes*. Springer International Publishing, Jan. 2015.