IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS 2025)
23-24 October 2025, Kushtia, Bangladesh

CerviTrans-XAI: An Explainable Vision Transformer Ensemble for Accurate Cervical Cancer Classification

Md. Delower Hossain
Computer Science and Engineering
Southeast University
Dhaka, Bangladesh
2021200000102@seu.edu.bd

Md. Mijanur Rahman*
Computer Science and Engineering
Southeast University
Dhaka, Bangladesh
mijanur.rahman@seu.edu.bd

Md. Abdur Rahman
Computer Science and Engineering
Southeast University
Dhaka, Bangladesh
2021200000025@seu.edu.bd

Tanvir Ahmed
Computer Science and Engineering
Southeast University
Dhaka, Bangladesh
2021200000015@seu.edu.bd

Sujay Dhar
Computer Science and Engineering
Southeast University
Dhaka, Bangladesh
2021100000005@seu.edu.bd

Ornab Biswass
Computer Science and Engineering
Southeast University
Dhaka, Bangladesh
2020000000095@seu.edu.bd

***Abstract*—Cervical cancer presents an important worldwide health challenge, especially in developing nations, where restricted access to screening and postponed diagnosis remain substantial obstacles. Traditional diagnostic methods including Pap smears and Visual Inspection with Acetic Acid suffer from subjective interpretation and significant inter-observer variability, leading to an urgent requirement for automated diagnostic tools. This paper presents CerviTrans-XAI, an ensemble framework that combines multiple Vision Transformer architectures to achieve accurate cervical cancer cell classification with integrated explainable artificial intelligence capabilities. The proposed approach employs four distinct ViT models: ViT-B/16, ViT-DINO, DeiT-B/16, and Swin Transformer. These models train from the SipakMed dataset, which has five types of cervical cells. Through a soft voting ensemble methodology, which averages probability distributions from individual models, the framework achieves superior generalization compared to single model implementations. To overcome the black-box challenge of deep learning models in clinical settings, we incorporate LIME, which offer visual insights by identifying key cellular regions that influence diagnostic outcomes. Experimental validation demonstrates that the classification accuracy of CerviTrans-XAI is 98.27%. on the SipakMed dataset, establishing a new standard for automated cervical cancer screening systems. The integration of accurate diagnostics with clear, explainable decision-making makes this framework a valuable asset for assisting healthcare providers in under-resourced areas where expert pathologists are limited.**

***Index Terms*—Vision Transformer, Cervical Cancer Classification, Ensemble Learning, Explainable AI, Medical Image Analysis, Deep Learning**

I. Introduction

Cervical cancer is a major health issue all over the world. Each year, this disease kills many women because it is not diagnosed early enough. Right now, it is the fourth most common type of cancer and cause of cancer deaths in women [1]. The World Health Organization (WHO) reports that around 80% of deaths from cervical cancer occur in developing countries [2]. When cancer cells become harmful, they could spread to other sections of the body, making the situation much more serious. It happens when cancer cells that aren't normal grow in the cervix, the lower portion of a woman's uterus. Finding the cancer early can help stop the cancer cells from becoming dangerous.

For a long time, doctors have utilized Pap smear tests and Visual Inspection with Acetic Acid (VIA) to identify cervical cancer. These methods have been important tools for preventing cervical cancer. However, these approaches have some problems. They depend on doctors examining results by hand, and different doctors might interpret the same results differently. Because of these issues, there is an urgent need for better diagnostic methods that can provide more accurate and reliable results.

In recent times, deep learning and AI have started to play an important role in how doctors read and understand medical images. These advanced systems, especially those using neural networks, are capable of picking up patterns that might not be immediately obvious to the human eye. They've been used successfully in many types of medical diagnoses and are known for giving quick and reliable results. To help doctors feel more confident in these tools, explanation methods like LIME [3] have been introduced. These explanations show how the AI came to a decision, which helps make the technology more transparent and easier to trust.

This paper's major contributions are:

- Developing a ViT-based ensemble model, achieving 98.27% accuracy on the SipakMed cervical cancer dataset.
- Applying LIME to provide visual interpretability by highlighting influential image regions, thereby increasing

979-8-3315-5525-2/25/$31.00 ©2025 IEEE

transparency and clinical trust.

II. Related Works

Deep learning has become increasingly significant in medical image analysis, particularly for cervical cell classification. Early diagnosis is a key factor in improving patient outcomes, and researchers have explored numerous architectures to tackle this challenge. Tripathi et al. [4] applied a ResNet152 model to the SIPaKMeD dataset and achieved 94.89% accuracy, demonstrating the value of deep residual networks for cervical cytology image classification. Hemalatha et al. [5] extended this approach using ResNetV2, which achieved a slightly higher accuracy of 95.33%. Mathivanan et al. [6] introduced a hybrid technique that combined ResNet152 with logistic regression, reaching an impressive 98.08% accuracy. Gangrade et al. [7] introduced an ensemble of convolutional neural networks to enhance generalization, obtaining 94.00% accuracy on the same benchmark. In addition, Sharma et al. [8] proposed CerviTransX, an explainable transformer-based approach for Pap smear image classification. These studies collectively demonstrate the ongoing shift from traditional deep CNNs toward transformer-based models and hybrid systems, aiming to increase accuracy, interpretability, and computational performance in cervical cancer classification tasks. Such methods are particularly attractive because they go beyond high-level accuracy metrics and provide human-understandable visual explanations that can improve trust among pathologists.

Building on these advances, our proposed CerviTrans-XAI leverages a heterogeneous ensemble of Vision Transformers to integrate complementary feature representations from multiple architectures. This combination enables a richer and more robust feature space than any single model, leading to a new state-of-the-art (SOTA) accuracy of 98.27% on the SIPaKMeD dataset and surpassing traditional CNNs, hybrid architectures, and existing transformer models.

III. Methodology

The proposed framework, CerviTrans-XAI, for cervical cell classification integrates a robust ensemble of fine-tuned Vision Transformer models to achieve high accuracy and generalizability. The overall pipeline, illustrated in Fig. 1, begins with dataset preparation and preprocessing, followed by the parallel fine-tuning of four distinct transformer-based architectures. The outputs of each model are then put together using a probability averaging ensemble strategy. Finally, we apply Local Interpretable Model-agnostic Explanations (LIME) to explain how the ensemble makes decisions. The subsequent sections detail each component of this methodology.

A. Dataset and Preprocessing

We perform our experiments on the publicly available SIPaKMeD dataset [9], a standard benchmark for cervical cytology classification. The dataset comprises 4049 high-resolution images of individually cropped cells, acquired from Pap smear slides. These are classified into five separate morphological classes: *Dyskeratotic*, *Koilocytotic*, *Metaplastic*, *Parabasal*, and *Superficial-Intermediate*.

The dataset was split into training (70%, 2834 images), validation (10%, 405 images), and test (20%, 810 images) sets. To ensure that the class distribution was maintained across the splits, we employed a stratified sampling strategy based on the cell type.

Each image was scaled to a consistent input resolution of 224×224 pixels to comply with the input specifications of the pre-trained models. We applied a wide range of data augmentation methods on the training set to increase model robustness and ignore overfitting. Images in the test and validation sets were only resized, without any further augmentation. Finally, the standard deviation and ImageNet mean values were used to normalize all the images. To identify the basic class imbalance within the training data, we calculated class weights inversely proportional to their frequency. After then, these weights were put together into the loss function to penalize misclassifications of minority classes more severely.

B. Ensemble Model Architecture

The cornerstone of our method is a powerful ensemble model designed to capitalize on the diverse representational capabilities of different Vision Transformer backbones. By aggregating predictions from multiple, distinct architectures, the ensemble benefits from a broader "understanding" of the input data, which enhances classification accuracy and robustness against the idiosyncrasies of any single model.

1) Constituent Architectures: We selected four SOTA models, each pre-trained on the ImageNet-1k dataset, to serve as the foundation of our ensemble. The last layer of each model's classification was substituted by a new fully connected layer that was made just for the five classes in the SIPaKMeD dataset.

Vision Transformer (ViT-B/16) functions as our baseline architecture. It functions by dividing a picture into a series of fixed-size patches, which are subsequently arranged linearly and processed by a conventional Transformer encoder [10]. This process lets the model identify global connections between different sections of the image from the very first layer.

Swin Transformer (Swin-T) introduces a hierarchical structure that is more analogous to typical Convolutional Neural Networks (CNNs). It calculates self-attention within non-overlapping, localized windows. These are then progressively merged in deeper layers, and a shifted windowing scheme allows for cross-window connections, enabling efficient and scalable modeling of both local and global features [11].

Data-efficient Image Transformer (DeiT-B/16) addresses the challenge of training Vision Transformers on smaller datasets. It employs a knowledge distillation strategy where a "distillation token" learns directly from the outputs of a pre-trained, strong teacher model. This enables DeiT to attain competitive performance without requiring massive pre-training datasets [12], making it well-suited for fine-tuning on specialized medical data.

DINO ViT (Dino-B/16) is a model pre-trained using a self-supervised learning paradigm called DINO [13]. Instead of using class labels, it learns rich visual features by ensuring

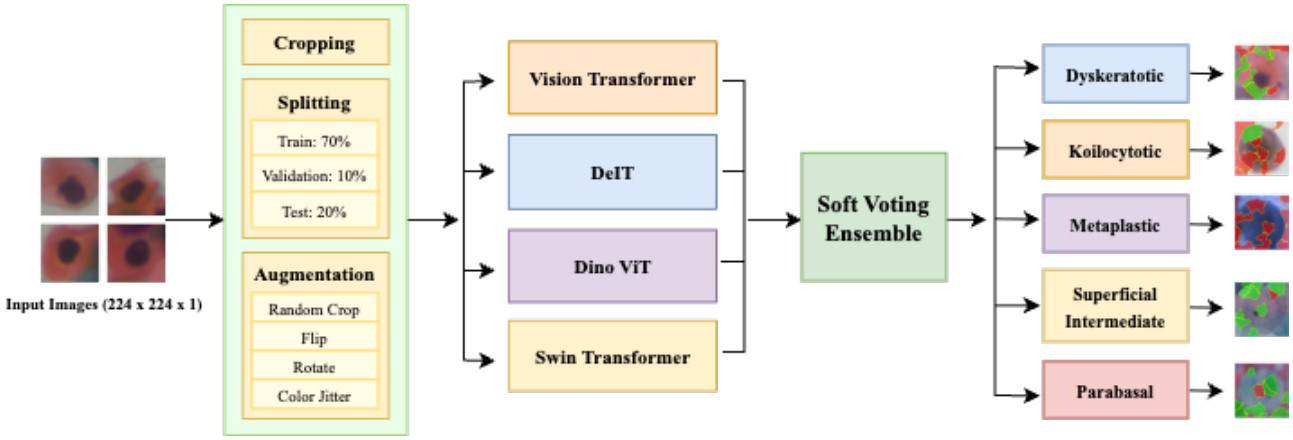

Fig. 1: The Proposed CerviTrans-XAI Framework

that various augmented "views" of the identical image produce consistent outputs. This technique motivates the model to gain powerful, semantically meaningful illustrations of object parts and textures, which are highly beneficial for downstream classification tasks.

The combination of these four models provides a comprehensive set of feature extractors, spanning global context (ViT), hierarchical structure (Swin), data-efficient training (DeiT) and semantically rich self-supervised features (DINO).

2) Model Fine-Tuning and Ensemble Integration: Each of the four constituent models was independently fine-tuned on our dedicated training set. We initialized the models with their respective pre-trained weights to leverage learned features for faster and more effective adaptation to the cervical cell domain. Model weights were updated using the AdamW optimizer. To handle class imbalance, a weighted cross-entropy loss function was utilized during training, giving more penalties to errors on minority classes.

The training process was regularized using a dynamic learning rate scheduler that adjusted the rate based on validation accuracy, and an early stopping criterion was used for avoiding overfitting by terminating training when validation performance ceased to improve. The model checkpoint with the maximum validation accuracy was preserved for each architecture. To accelerate the process, all training was performed with automatic mixed-precision (AMP).

The final classification is determined via a soft-voting ensemble strategy. For a given input image x, each model M_i in the ensemble calculates a probability vector $p_i = \text{softmax}(M_i(x))$. The final ensemble probability vector, P_{ens}, is the unweighted arithmetic mean of these individual vectors:

$$P_{\text{ens}}(x) = \frac{1}{N} \sum_{i=1}^{N} p_i \tag{1}$$

where $N = 4$. The class corresponding to the maximum value in P_{ens} is picked as the ultimate prediction. This averaging approach smooths out individual model predictions and reduces variance, typically leading to a more reliable result.

C. Explainable AI for Model Interpretation

To ensure transparency and build trust in our model's decisions, We use LIME [3]. LIME approximates the characteristics of our complex ensemble model (f) around an individual prediction by training a simpler, interpretable linear model (g) on perturbations of the original image created from superpixels.

Formally, the explanation $\xi(x)$ is derived by tuning the following objective:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{2}$$

where $\Omega(g)$ is a complexity penalty that promotes sparsity and $\mathcal{L}$ is a fidelity-loss function that quantifies how closely g estimates f in the position π_x. The resulting weights of the learned model g correspond to the importance of each superpixel. We visualize these weights to highlight the image regions most influential to the prediction, allowing for a direct, qualitative assessment of whether the model focuses on cytologically relevant features.

IV. Experimental Results and Discussion

In this part, we represent a full evaluation of our suggested **CerviTrans-XAI** model. First, we go over the experimental environment, analyze the training dynamics, present the quantitative and qualitative performance, and finally discuss the broader implications, limitations, and future directions of our work.

A. Experimental Setup

All of the experiments took place in the Kaggle environment equipped with a NVIDIA Tesla P100 GPU, a Intel Xeon CPU, and 29 GB of RAM. The PyTorch and timm libraries were used to build the framework in Python. The fine-tuning process for all base models employed the AdamW optimizer with a weight decay of 1×10^{-3} and an initial learning rate of 1×10^{-4}. A batch size of 32 was used. To ensure optimal convergence, a 'ReduceLROnPlateau' learning rate scheduler made the learning rate lower if validation accuracy stagnated

for 3 epochs. A mechanism for early stopping with a patience of 5 epochs were used to select the best-performance model checkpoint.

B. Training Analysis

To verify the stability and generalization of our fine-tuning process, we analyzed the validation and training curves for each constituent model, as shown in Fig. 2. All four models exhibit successful and stable convergence, with training and validation losses consistently decreasing while accuracies increase. Essentially, the validation loss for each model closely follows the training loss without significant divergence, and the difference between training and validation accuracy stays small. This indicates that our comprehensive data augmentation scheme and regularization techniques effectively mitigated overfitting. Furthermore, the termination of training at different epochs for each model demonstrates that the early stopping mechanism was instrumental in capturing the optimal checkpoint for each architecture, preventing performance degradation from excessive training. These robust training dynamics provide a solid foundation for the strong performance of our final ensemble.

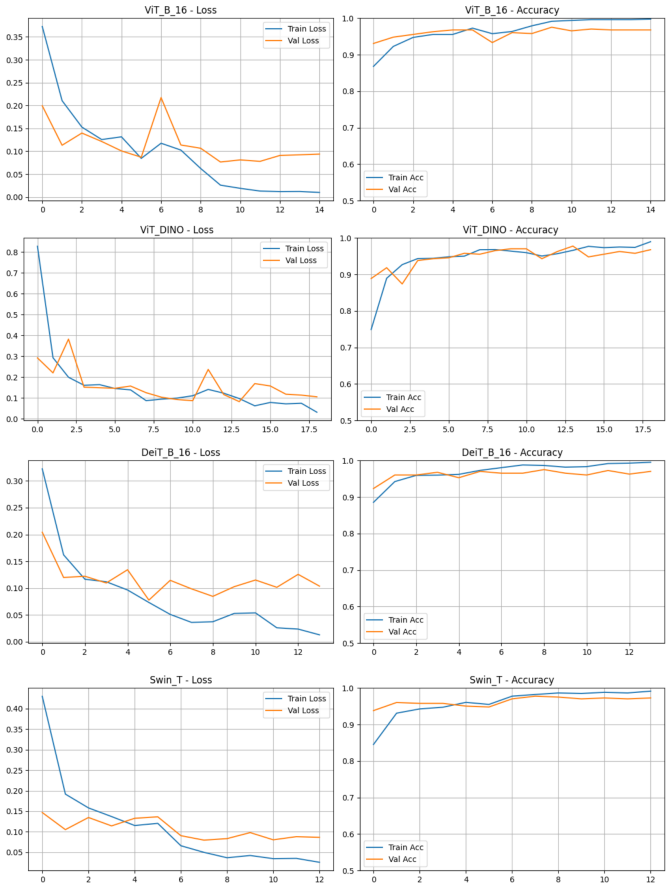

Fig. 2: Train and valid loss and accuracy curves for the four constituent models: (a) ViT-B, (b) DINO-B, (c) DeiT-B, and (d) Swin-T.

C. Quantitative Results

We evaluated our approach using standard classification metrics, presented in Table I. Each constituent transformer model achieves strong performance, with accuracies exceeding 96.9%. Our proposed ensemble, **CerviTrans-XAI**, consistently outperforms every individual model, achieving an overall F1-Score and accuracy of 98.28%. This improvement highlights the benefit of aggregating diverse feature representations.

TABLE I: Quantitative Performance on the SIPaKMeD

Model	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)
ViT-B	97.53	97.53	97.54	97.53
DINO-B	96.91	96.95	96.93	96.92
DeiT-B	97.65	97.67	97.68	97.64
Swin-T	97.65	97.66	97.66	97.66
CerviTrans-XAI	**98.27**	**98.28**	**98.28**	**98.28**

The confusion matrix in Fig. 3 shows the class-wise performance of CerviTrans-XAI, which is quite accurate across all classes. The ROC curve in Fig. 4 further demonstrates the model's exceptional discriminative capability, with a macro-average Area Under the Curve (AUC) of 0.9995.

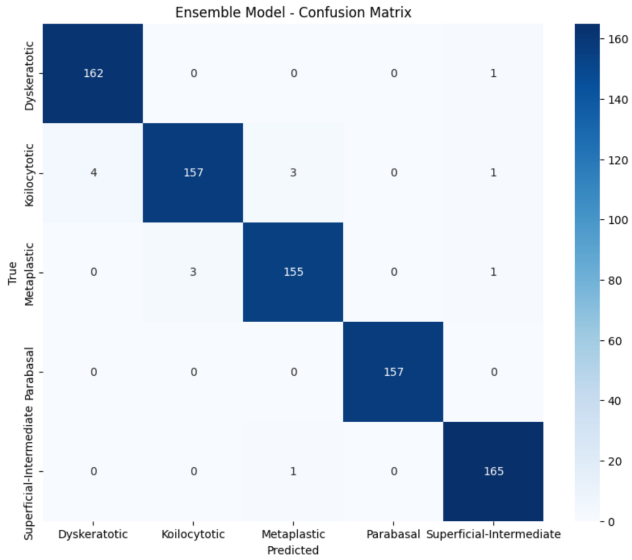

Fig. 3: Confusion matrix for the proposed CerviTrans-XAI ensemble model.

D. Comparison with State-of-the-Art

To contextualize our contribution, we compare CerviTrans-XAI with recent methods on the SIPaKMeD dataset in Table II. Our model achieves an accuracy of 98.28%, establishing a new SOTA on this benchmark and surpassing previous works, including those based on deep CNN ensembles and hybrid ResNet architectures.

E. Qualitative Analysis with LIME

While quantitative metrics are high, the "black box" character of deep models makes clinical trust harder. To address

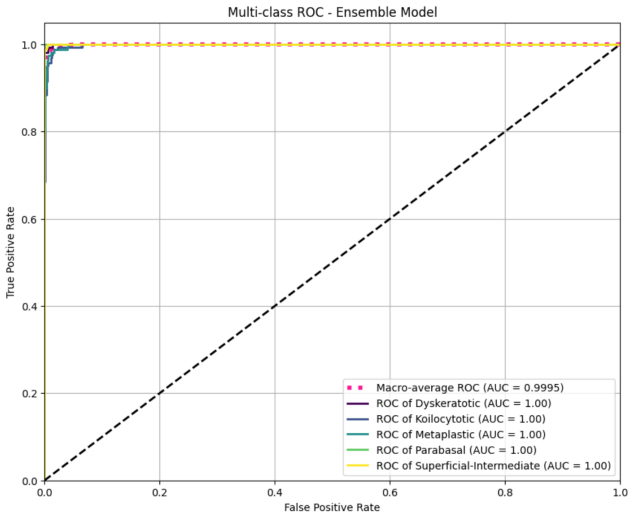

Fig. 4: Receiver Operating Characteristic (ROC) curves for the proposed CerviTrans-XAI ensemble model

TABLE II: Comparison with previous studies on SIPaKMeD Dataset

Reference	Year	Method	Acc. (%)
Tripathi et al. [4]	2021	ResNet152	94.89
Hemalatha et al. [5]	2022	ResNetV2	95.33
Mathivanan et al. [6]	2024	ResNet152 + Logistic	98.08
Gangrade et al. [7]	2025	Ensemble CNN	94.00
Our Proposed	**2025**	**CerviTrans-XAI**	**98.28**

this, we use LIME [3] to understand *why* our model makes a prediction. Fig. 5 presents LIME visualizations for four correctly classified examples. In the *Koilocytotic* example (Fig. 5a), LIME correctly highlights the enlarged, irregular nucleus and characteristic perinuclear halo. Similarly, for the *Dyskeratotic* cell (Fig. 5b), the explanation focuses on the dense, hyperchromatic nucleus. The model's reasoning aligns with clinical knowledge for the *Parabasal* cell (Fig. 5c), where the high nucleus-to-cytoplasm ratio is identified, and for the mature *Superficial-Intermediate* cell (Fig. 5d), where LIME emphasizes its small, pyknotic nucleus. These visualizations provide strong evidence that our model's decisions align with established cytopathological criteria.

F. Discussion

Our results demonstrate that a heterogeneous ensemble of Vision Transformers, CerviTrans-XAI, sets a new SOTA for the classification of cervical cell on the SIPaKMeD benchmark. The success of our approach stems from combining the varied inductive biases of different transformer architectures; the global context captured by ViT, the hierarchical feature learning of Swin-T, and the powerful semantic representations from DeiT and DINO create an additional comprehensive and massive feature space than any single model could achieve alone. Furthermore, our LIME analysis shows that this high performance is not based on dataset artifacts but on the model

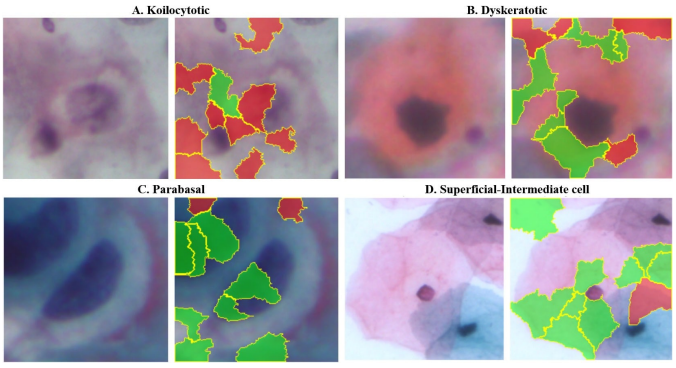

Fig. 5: LIME explanations for four sample predictions. Green regions support the model's prediction, confirming its focus on clinically relevant features for (a) Koilocytotic, (b) Dyskeratotic, (c) Parabasal, and (d) Superficial-Intermediate cells.

learning clinically meaningful features, a critical step towards building trustworthy AI for medical diagnosis.

Despite these promising results, we acknowledge several limitations. First, this study is limited by its evaluation on a single dataset, SipakMed. While the framework gets results that are the best in the field on this benchmark, its generalizability to other datasets with variations in image acquisition and preparation protocols has not been established. Second, our evaluation is performed on a dataset of pre-cropped, isolated cells. Real-world clinical scenarios involve analyzing whole-slide images (WSIs), which present significant challenges such as cell overlaps, staining variations, and artifacts. The performance of CerviTrans-XAI in such a setting remains to be validated. Third, the ensemble nature of our model, while accurate, is computationally intensive, which could be a barrier for deployment in resource-constrained environments.

Based on these limitations, our future work will proceed in several directions. First, to determine the generalizability of our model, we set up to monitor its performance on additional, diverse cervical cytology datasets. Second, we will focus on adapting and evaluating our framework on WSIs, which will require integrating robust cell segmentation and artifact handling pre-processing steps. Finally, we will explore model compression and knowledge distillation techniques to transfer the knowledge from our large ensemble into a single, compact, and efficient model [14]. This "distilled" CerviTrans-XAI could retain the high accuracy of the ensemble while being suitable for real-time clinical deployment. Other works have also explored WSIs and model compression for medical imaging, but are acknowledged here without individual citation to maintain conciseness.

V. Conclusion

This research presents CerviTrans-XAI, a novel ensemble framework that successfully combines Vision Transformers for automated cervical cancer classification with integrated explainable AI capabilities. Our experimental evaluation on the SipakMed dataset demonstrates exceptional performance,

achieving 98.27% classification accuracy that establishes a new standard for cervical cancer detection systems. The ensemble strategy effectively leverages the complementary feature representations learned by different ViT architectures, resulting in superior generalization compared to individual model implementations. The integration of LIME-based explainable AI provides crucial transparency in the process of decision-making, addressing the interpretability challenges that often hinder clinical adoption of deep learning systems. The visual explanations successfully highlight anatomically relevant cellular structures, demonstrating that the model learns meaningful pathological features rather than exploiting dataset artifacts. This combination of high accuracy and interpretable predictions represents a significant advancement in automated cervical cancer screening technology. The framework's ability to provide both reliable classification and transparent decision rationale positions it as a useful instrument for supporting healthcare personnel in the diagnosis of cervical cancer, especially in places where resources are limited and professional pathologists may not be available. CerviTrans-XAI demonstrates the potential of Vision Transformer ensembles to deliver clinically viable AI-assisted diagnostic systems that can enhance patient outcomes through improved screening accuracy and diagnostic confidence.

REFERENCES

[1] A. D. Shrestha, D. Neupane, P. Vedsted, and P. Kallestrup, "Cervical cancer prevalence, incidence and mortality in low and middle income countries: a systematic review," *Asian Pacific journal of cancer prevention: APJCP*, vol. 19, no. 2, p. 319, 2018.

[2] W. H. Organization, *Cervical cancer screening in developing countries: report of a WHO consultation*. World Health Organization, 2002.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[4] A. Tripathi, A. Arora, and A. Bhan, "Classification of cervical cancer using deep learning algorithm," in *2021 5th international conference on intelligent computing and control systems (ICICCS)*. IEEE, 2021, pp. 1210–1218.

[5] K. Hemalatha and V. Vetriselvi, "Deep learning based classification of cervical cancer using transfer learning," in *2022 international conference on electronic systems and intelligent computing (ICESIC)*. IEEE, 2022, pp. 134–139.

[6] S. K. Mathivanan, D. Francis, S. Srinivasan, V. Khatavkar, K. P, and M. A. Shah, "Enhancing cervical cancer detection and robust classification through a fusion of deep learning models," *Scientific Reports*, vol. 14, no. 1, p. 10812, 2024.

[7] J. Gangrade, R. Kuthiala, S. Gangrade, Y. P. Singh, and S. Solanki, "A deep ensemble learning approach for squamous cell classification in cervical cancer," *Scientific Reports*, vol. 15, no. 1, p. 7266, 2025.

[8] N. Sharma, K. Gaurav, and T. K. R. Bollu, "Cervitransx: Explainable transformer-based cervical cancer classification," in *2025 National Conference on Communications (NCC)*. IEEE, 2025, pp. 1–6.

[9] 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)*. IEEE, 2018, pp. 3144–3148.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[11] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[12] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[13] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

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