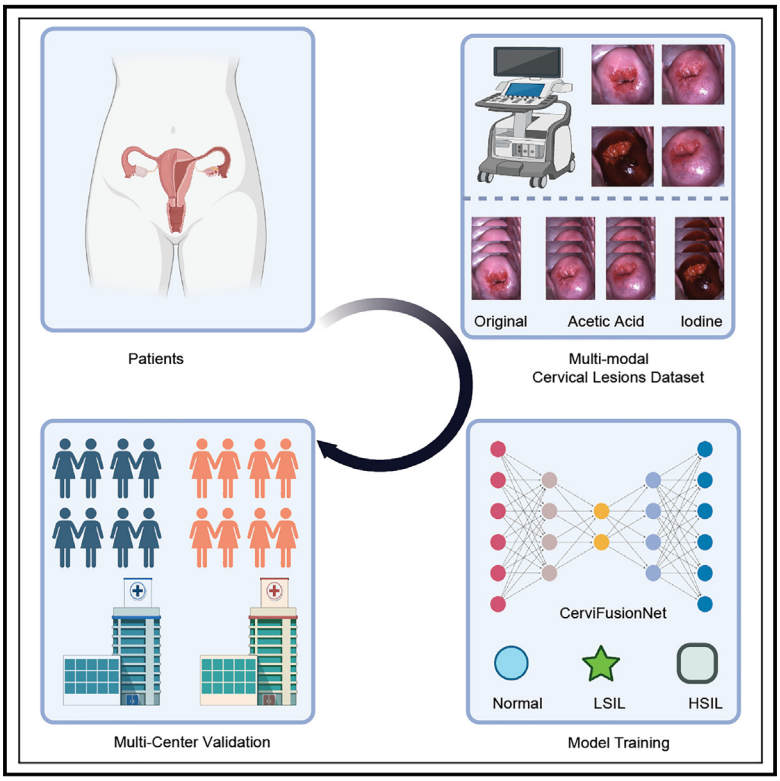
iScience

Article

CerviFusionNet: A multi-modal, hybrid CNN-transformer-GRU model for enhanced cervical lesion multi-classification

Graphical abstract



Highlights

- Establishing a multi-modal dataset for cervical lesions classification
- We developed CerviFusionNet, a novel AI model to process multi-modal dataset
- CerviFusionNet achieved high-precision prediction for cervical lesions diagnosis
- We evaluated the model's robustness and generalization on the external dataset

Authors

Yuyang Sha, Qingyue Zhang, Xiaobing Zhai, ..., Yuefei Wang, Kefeng Li, Jing Ma

Correspondence

kefengl@mpu.edu.mo (K.L.), majing2609@163.com (J.M.)

In brief

Medical imaging; Cervical smear; Classification of bioinformatical subject; Artificial intelligence

Sha et al., 2024, iScience 27, 111313
December 20, 2024
https://doi.org/10.1016/j.isci.2024.111313

CellPress

iScience

CellPress
OPEN ACCESS

Article

CerviFusionNet: A multi-modal, hybrid CNN-transformer-GRU model for enhanced cervical lesion multi-classification

Yuyang Sha,[1,7] Qingyue Zhang,[2,3,7] Xiaobing Zhai,[1] Menghui Hou,[2,3] Jingtao Lu,[4] Weiyu Meng,[1] Yuefei Wang,[5,6] Kefeng Li,[1,*] and Jing Ma[2,3,8,*]

[1]Center for Artificial Intelligence Driven Drug Discovery, Faculty of Applied Sciences, Macao Polytechnic University, Macau SAR 999078, China
[2]First Teaching Hospital of Tianjin University of Traditional Chinese Medicine, Tianjin 300381, China
[3]National Clinical Research Center for Chinese Medicine Acupuncture and Moxibustion, Tianjin 300381, China
[4]Beijing University of Technology, School of Mathematical Statistics and Mechanics, Beijing 100124, China
[5]National Key Laboratory of Chinese Medicine Modernization, State Key Laboratory of Component-based Chinese Medicine, Tianjin University of Traditional Chinese Medicine, Tianjin 301617, China
[6]Haihe Laboratory of Modern Chinese Medicine, Tianjin 301617, China
[7]These authors contributed equally
[8]Lead contact
*Correspondence: kefengl@mpu.edu.mo (K.L.), majing2609@163.com (J.M.)
https://doi.org/10.1016/j.isci.2024.111313

SUMMARY

Cervical lesions pose a significant threat to women's health worldwide. Colposcopy is essential for screening and treating cervical lesions, but its effectiveness depends on the doctor's experience. Artificial intelligence-based solutions via colposcopy images have shown great potential in cervical lesions screening. However, some challenges still need to be addressed, such as low algorithm performance and lack of high-quality multi-modal datasets. Here, we established a multi-modal colposcopy dataset of 2,273 HPV+ patients, comprising original colposcopy images, acetic acid reactions at 60s and 120s, iodine staining, diagnostic reports, and pathological results. Utilizing this dataset, we developed CerviFusionNet, a hybrid architecture that merges convolutional neural networks and vision transformers to learn robust representations. We designed a temporal module to capture dynamic changes in acetic acid sequences, which can boost the model performance without sacrificing inference speed. Compared with several existing methods, CerviFusionNet demonstrated excellent accuracy and efficiency.

INTRODUCTION

Long-term cervical lesions without effective treatment may transform into cervical cancer. Cervical cancer is the fourth most common malignancy tumor affecting women's health, with approximately 604,000 new cases and 342,000 deaths in 2020 reported by the World Health Organization (WHO).[1–3] Unfortunately, over 80% of cervical cancer cases occur in low- and middle-income countries, lacking screening infrastructure.[4,5] In some undeveloped regions, cervical cancer has become the leading malignant tumor affecting local women. Risk factors such as the early age of first sexual intercourse and unsafe sexual behaviors increase the likelihood of developing cervical lesions and cancer.[6–8] Unlike other types of cancers, cervical cancer has a known cause and can be prevented through early screening and timely intervention. Therefore, designing simple, effective, and economical methods for cervical lesions detection is an important research topic, especially for undeveloped regions.[9]

Cervical lesions are primarily caused by persistent cervix infection with high-risk human papillomavirus (HPV), such as HPV 16 and HPV 18.[10,11] Squamous intraepithelial lesion (SIL) is generally used to describe the abnormal growth of squamous cells on the surface of the cervix. Based on the Bethesda system,[12] SIL can be classified into low-grade SIL (LSIL) and high-grade SIL (HSIL). LSIL cells have only minor abnormal characteristics but still resemble normal ones. Observation and monitoring are typically recommended for patients with LSIL, as the immune system may clear abnormal cells over time. However, HSIL cells exhibit abnormality under the microscope. Unlike LSIL, patients with HSIL are recommended to receive a loop electrosurgical excision procedure (LEEP).[13] Otherwise, some of these high-grade diseased cells may cause cervical cancer. Therefore, the main goal of colposcopy is to distinguish pathological normal, LSIL, or HSIL, which is essential for selecting the appropriate treatment approaches.[14,15]

Currently, cervical lesions can be detected through various screening methods, such as HPV testing, ThinPrep cytologic test (TCT), and colposcopy. HPV testing is a DNA-based detection that can identify whether the patients carry high-risk HPV. The WHO recommends using HPV testing for cervical cancer

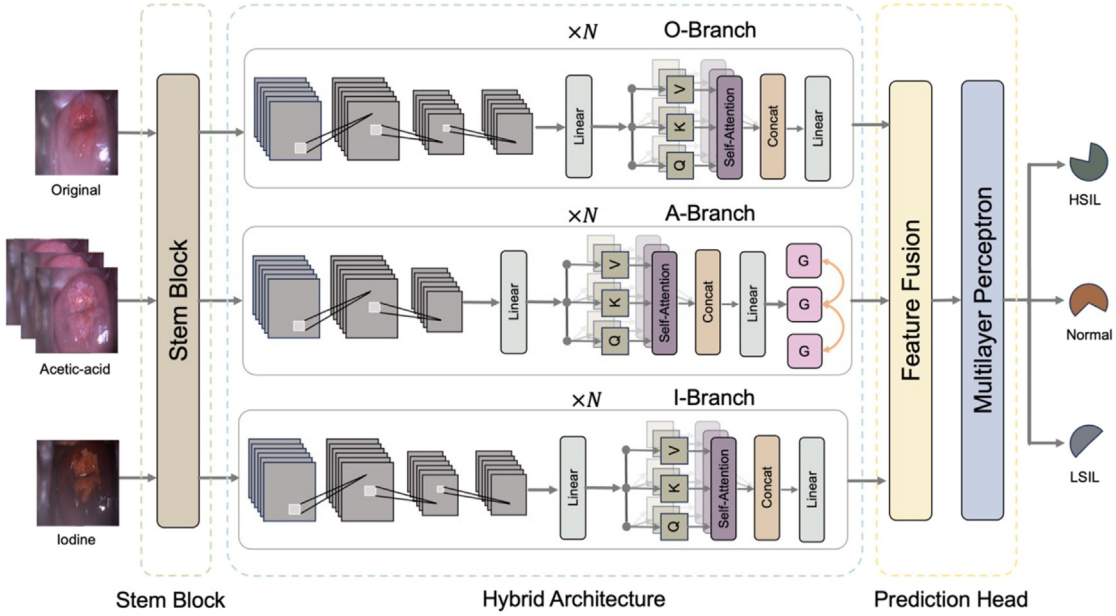


Figure 1. Overview of the proposed CerviFusionNet architecture for cervical lesions classification

The proposed method consists of three parts: Stem Block, Hybrid Architecture, and Prediction Head. First, the Stem Block can extract detailed information from input multi-modal samples and reduce the resolution quickly. Second, the designed Hybrid Architecture, composed of CNN and ViT, is responsible for extracting robustness representations following the Stem Block. Then, these extracted features are fused and refined by the proposed Feature Fusion Module. Finally, we use a multilayer perceptron (MLP) to map these refined features into cervical lesions classification, including Normal, LSIL, and HSIL.

screening, which does not require advanced medical instruments or experienced doctors, making it accessible even in undeveloped areas.[16] However, HPV testing often produces false-positive results, which can lead to misdiagnosis and increase the burden on healthcare systems.[17] Colposcopy is a more effective way of screening cervical cancer. During this examination, doctors focus on observing the vascular morphology and epithelial structure of the cervix. Some studies[18–20] have shown that colposcopy effectively reduces cervical cancer incidence. However, the effectiveness of coloscopy heavily depends on the doctors' experience. While cervical cancer screening has been widely promoted worldwide, its effectiveness is far from satisfactory. For instance, the Chinese government has offered free cervical screening for several years, but more than 59,060 women still died of cervical cancer in 2020.[2] Computer-aided diagnosis (CAD) frameworks have demonstrated strong potential in various medical tasks.[21,22] Therefore, it is worthwhile to research and develop high-precision cervical lesions analysis systems based on artificial intelligence and colposcopy images.

Due to the groundbreaking advancements in artificial intelligence technology and computer hardware, deep learning methods have been widely used in many fields, such as image analysis,[23–25] data mining,[26] and medical diagnosis.[27–29] Currently, several deep learning-based systems for cervical lesions classification have been proposed, significantly improving doctors' diagnostic efficiency. For instance, Zhang et al.[20] introduced a patch-wise solution for cervical cancer image recognition, which can integrate the domain knowledge to boost the model performance. Chen et al.[30] designed a high-performance backbone model with EfficientNet[31] for colposcopy sample feature extraction. Then, it employed the GRU module to encode the global relationship between input sequences. Due to the immature squamous metaplasia and cervical ectropion, these methods may produce some false positive predictions, leading to misdiagnosis. In order to handle this challenge, some studies have combined additional information to boost the model performance. MSCI[32] developed a C-GCNN model for cervical cancer screening. It could consider time series and combined multistate cervical images to enhance model performance. Li et al.[33] developed the CMF model to classify cervical lesions by mining the correlation between colposcopy images and diagnostic information, improving model accuracy and robustness. Additionally, other researches[34–36] efforts have integrated image representations with diagnostic reports, incorporating age information, cytology reports, HPV test results, and cervix transformation zone type to provide more comprehensive information for diagnosis.

Existing CAD systems[20,30,32] for cervical lesions classification via colposcopy images have made considerable progress, but some problems still need to be solved urgently. First, many existing methods can only process single-frame samples, making them ignore important information in acetic acid reaction sequences and iodine staining. Most studies focus solely on the classification of LSIL and HSIL, disregarding the normal state, which may hinder precision treatment for patients. Third, methods based on information fusion introduce a large amount of redundant representations, significantly affecting the model's efficiency. Simultaneously, these models may be prone to overfitting when the training sample is small or incomplete. Additionally, most colposcopy datasets are proprietary and difficult to

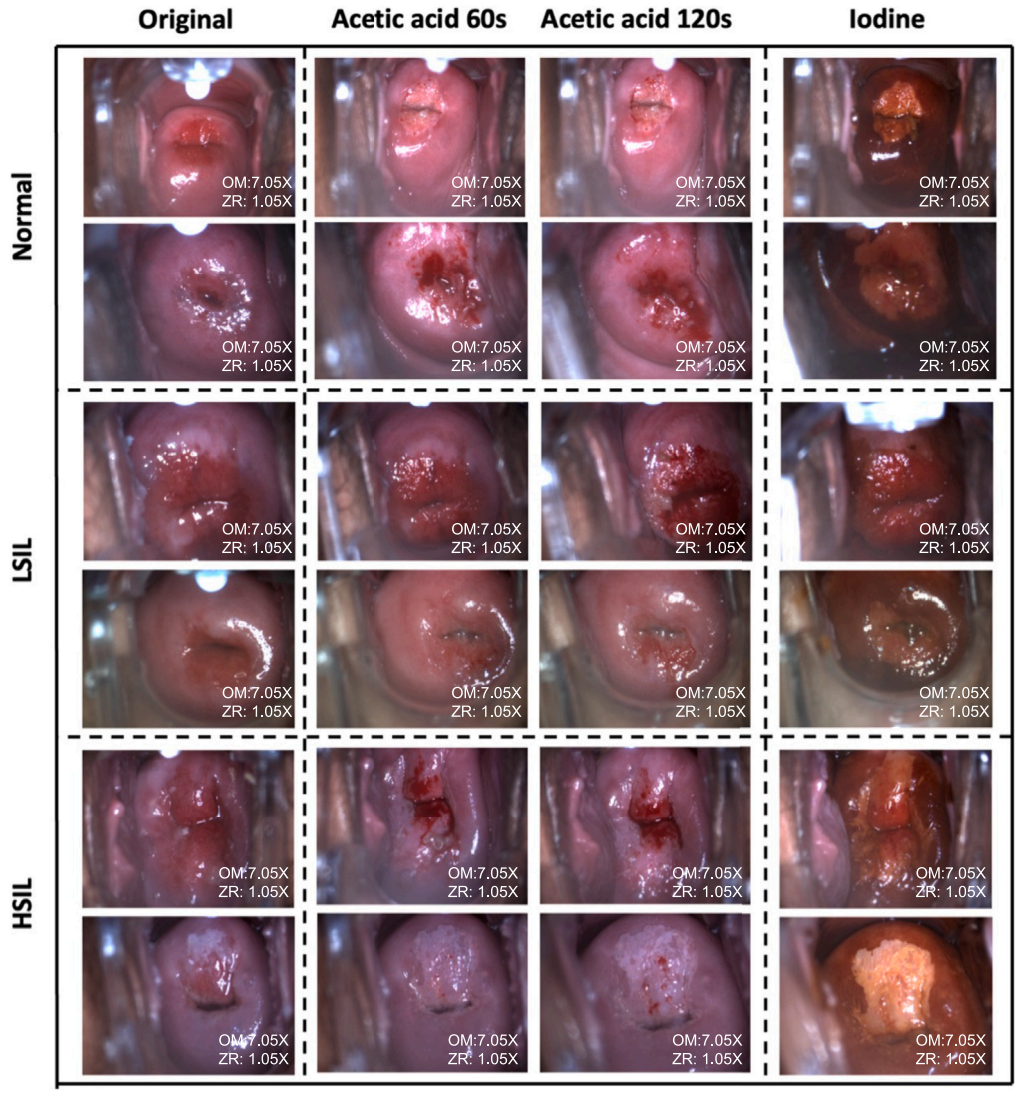


Figure 2. Visualization of the collected multi-modal dataset for cervical lesions classification
OM defines Optical Magnification, and ZR means Zoom Ratio.

access. These datasets usually focus on single-modal data collection and give less attention to multi-modal samples.

To tackle the issues mentioned, we gathered a multi-modal dataset for classifying cervical lesions. Utilizing this dataset, we developed a method called CerviFusionNet for cervical lesions analysis. This method can extract robust features from multi-modal images and uncover the relationships between input samples. To our knowledge, this dataset may be the first publicly available multi-modal colposcopy dataset. It contains 2,273 HPV+ cases, and all of the samples are annotated by experienced doctors. This dataset comprises original images, 60s and 120s acetic acid reactions, iodine staining, diagnoses, and pathology reports. To take full use of multi-modal colposcopy images, the proposed CerviFusionNet employs a hybrid architecture with three feature extraction branches that obtain discriminate features from input samples. Specifically, we have designed a Temporal Encoder Module for the acetic acid reaction data to extract correlations in the time dimension. Then, the Feature Fusion Module fuses these input features and outputs refined representations with more discriminability and robustness. Finally, a simple MLP can map the refined feature into classification results. The details of CerviFusionNet are

Table 1. Basic characteristics of the collected dataset

Pathological results	Number of images (cases * 4)
Normal	253 * 4
LSIL	234 * 4
HSIL	214 * 4

Notes: LSIL: low-grade squamous intraepithelial lesions; HSIL: high-grade squamous intraepithelial lesions.

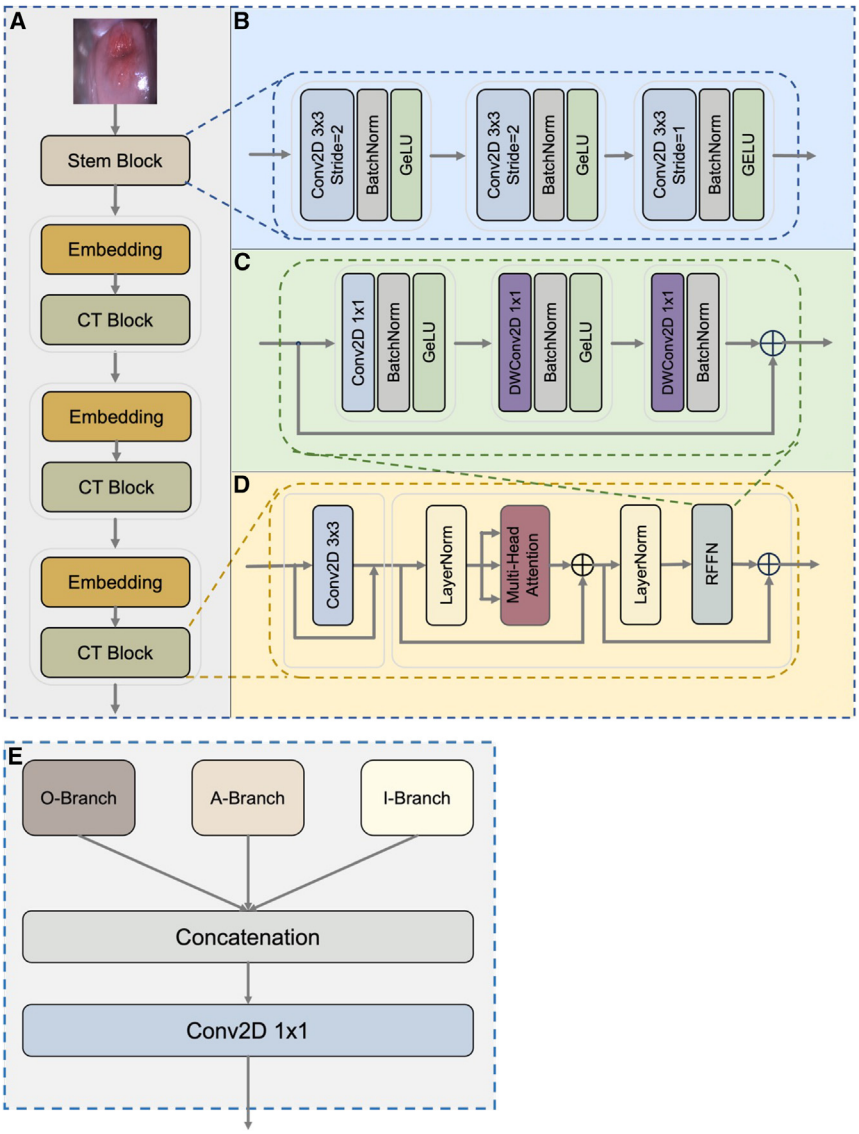


Figure 3. The details of proposed module in CerviFusionNet

(A) The structure of O-Branch in Hybrid Architecture. It forms by hybrid architecture with Linear and CT Block.

(B) The details of the Stem Block, which can make 4X downsizing for the input samples.

(C) The architecture of proposed RFFN module.

(D) Details of the CT Block, which consists of three key modules: Local Feature Unit (LFU), Multi-head Self-attention Encoder (MHSA), and Residual Feedforward Network (RFFD).

(E) The structure of proposed Feature Fusion Module.

shown in Figure 1. In summary, the primary contributions of this paper are as follows.

(1) We introduced the open-source, large-scale multi-modal colposcopy dataset, designed to facilitate the development of deep learning-based cervical lesions classification methods.

(2) Based on multi-modal colposcopy data, we designed the CerviFusionNet model to predict cervical lesions efficiently. Through the proposed Hybrid Backbone, Temporal Encoder Module, and Feature Fusion Module, our model can explore the relationship among the multi-modal information and optimize the model performance.

(3) We conducted extensive experiments to verify the effectiveness of proposed CerviFusionNet and compare it with existing cervical lesions diagnosis systems. The results demonstrate that our proposed approach achieves outstanding performance in terms of accuracy and effectiveness. Moreover, the model's accurate predictions highlight its potential for use in clinical scenarios.

(4) To further verify the robustness and generalization of proposed CerviFusionNet, we also collected an external validation dataset containing 50 cases, then evaluated the pre-trained model on it. The results show that CerviFusionNet still can achieve promising results on cervical lesions prediction. Furthermore, there is only a slight performance difference between our method on external and internal datasets, demonstrating the generalization of our method.

RESULTS

Dataset characteristics

The dataset includes information about 2,273 patients who visited the gynecological colposcopy clinic at the First Teaching Hospital of Tianjin University of Traditional Chinese Medicine between July 2021 and August 2023. After analysis, 701 patients are meeting the requirements for data collection. All patients perform colposcopy examinations and biopsies. Based on the pathological reports, patients were classified into three categories: pathological normal (Normal), low-grade squamous intraepithelial lesion (LSIL), and high-grade squamous intraepithelial lesion (HSIL). Specifically, this dataset includes 253 HPV-positive without cervical lesions cases, 234 LSIL cases, and 214 HSIL cases, totaling 2,804 images. Each patient has four colposcope images, including original colposcopy images (Original), acetic acid reaction time series at 60 and 120 s (Acid-60s, Acid-120s), and iodine staining images (Iodine). Besides, we also collected patients' clinical text information such as age, TCT and HPV status. All of the collected images are JPEG with the resolution of 1280 × 960. Magnification ranges from 1 to 40 times, depending on the patient's condition. Of note, all the colposcopy images were obtained by doctors with more

Table 2. Performance comparison of proposed method versus existing cervical lesion diagnosis systems

Methods	Accuracy	F1-Score	Precision	Recall
AlexNet	0.675	0.6726	0.6731	0.675
VGG-16	0.7083	0.7054	0.7071	0.7083
ResNet-34	0.75	0.749	0.75	0.75
ResNet-101	0.7666	0.7656	0.7656	0.7666
HRNet-W48	0.775	0.7748	0.7757	0.775
Swin	0.7833	0.783	0.7847	0.7833
LSTM	0.5416	0.5418	0.546	0.5416
CerviFusion	**0.8333**	**0.8333**	**0.8545**	**0.8333**

than ten years of experience. The pathological reports are used as ground truth for these collected samples. All the sensitive information was removed, and the data entries were anonymized. Figure 2 shows some examples of our established multi-modal dataset and Table 1 provides detailed information about the collected dataset. The study protocol was approved by the institutional review board (IRB) of the First Teaching Hospital of Tianjin University of Traditional Chinese Medicine (TYLL2023-Z-045). Written content was obtained from all the subjects upon image collection.

Comparison of the CerviFusionNet with different methods

In order to verify the effectiveness of our proposed method (details are shown in Figure 3), we compared it with several existing algorithms. Most of the existing colposcope image classification systems are based on single-modal datasets, which usually employ a complex Convolutional Neural Network (CNN) backbone to extract features and then use Multilayer Perceptron (MLP) to map them into prediction results. Therefore, we utilized classical deep learning methods to reimplement these cervical lesion diagnosis systems, including VGG,[37] ResNet,[38] HRNet,[39] and Swin-Transformer.[40] Specifically, all experiments were conducted using our collected dataset. To ensure fair comparison, we implemented all methods within the same software and hardware environment. The detailed results are presented in Table 2. As we can find that, our proposed CerviFusionNet better than other comparison methods in all evaluation metrics, with accuracy, f1-score, precision, and recall of 0.8333, 0.8333, 0.8545, and 0.8333 respectively. These results indicate that our proposed method achieve excellent performance and practicality in cervical lesions diagnosis.

Among them, the LSTM approach shows the worst performance compared with other approaches, indicating that the CNN or ViT plays an essential role in cervical lesion classification. The system with ResNet-101 performs better than VGG-16 and AlexNet. It suggests that the complex backbone has a stronger capacity for feature extraction and image classification. However, the complex backbone often has large model parameters and computational costs, damaging the inference speed and increasing resource consumption. Compared with ResNet-34, system with ResNet-101 has only a slight advantage in model performance, but the model parameters and computational costs have increased by 110.8% and 104.3%, respectively. When the single-modal model's performance reaches a certain range, increasing the model parameters and computational costs cannot significantly improve the performance. HRNet-W18 performs better than ResNet-101, indicating that the multi-scale feature and global information enhance the model performance in cervical cancer classification. Method with Swin-Transformer gets the best accuracy in the single-modal model. It demonstrates that the Vision Transformer-based structure has more advantages in handling colposcopy images. Based on this analysis, our proposed method combines the strengths of multi-scale features, multi-modal information, and hybrid architecture to achieve significant advantages over existing diagnosis systems.

Validation on external dataset

To evaluate the generalization of our proposed method, we conduct experiments on the external dataset. This dataset forms 50 cases (12 Normal, 21 LSIL, and 17 HSIL) collected by the First Teaching Hospital of Tianjin University of Traditional Chinese Medicine (XiQing District), details are shown in Table S1. The data type and annotation function are consistent with our established multi-modal cervical lesions classification dataset. Then, we validate the model performance with pre-trained CerviFusionNet on the external dataset, and the results are shown in Figure 4. Our proposed method demonstrates impressive accuracy, f1-score, precision, and recall of 0.8000 and 0.8083, 0.8155, and 0.8025, respectively. Additionally, the

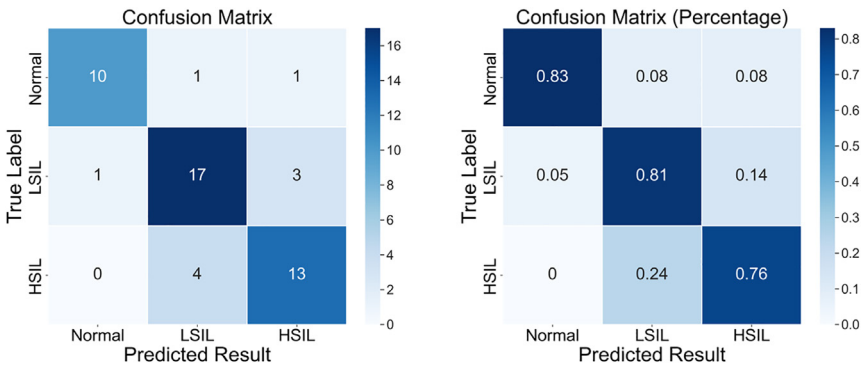


Figure 4. Confusion matrix of CerviFusionNet on the external validation dataset

Table 3. The model performance with different backbone in terms of accuracy, F1-score, precision, and recall. Higher is better

Methods	Accuracy	F1-Score	Precision	Recall
VGG-16	0.7916	0.7916	0.7970	0.7916
ResNet-34	0.8166	0.8155	0.8179	0.8166
Our Method	**0.8333**	**0.8333**	**0.8545**	**0.8333**

performance of CerviFusionNet surpasses existing methods for classifying cervical lesions when tested on an external dataset. The test results of the external dataset show minimal deviation compared to the internal testing data, further demonstrating the robustness and generativity of the proposed model.

Comparison of different backbones in model performance

In this section, we investigate the effectiveness of the proposed hybrid feature extraction module and conduct experiments on our collected data. The designed hybrid feature extraction module can discriminate representations from the input samples. To ensure a fair comparison, we replaced our proposed hybrid architecture with VGG-16 and ResNet-34. Then, these methods would be trained with the same experimental settings. The detailed results are reported in Table 3 and Figure S1A. Compared with CerviFusionNet using VGG-16, the system employing the proposed hybrid structure achieved a relative improvement of 5.3% and 5.3% in accuracy and f1-score, respectively. This indicates that the hybrid structure is suitable for classifying cervical lesions compared to classical deep-learning backbones. To further evaluate the model performance, we conduct extensive experiments to compare the model parameters and computational costs. The results are shown in Table 4 and Figure S2. The CerviFusionNet with hybrid feature extraction module is much small in model parameters (15.7M) and computational costs (2.76G), compared to system with VGG-16 or ResNet-34. According to Tables 2 and 4, the proposed model can better balance model performance and complexity than the comparison models. This demonstrates the potential of the proposed CerviFusionNet as an automated method for cervical lesion classification in clinical applications.

Effectiveness of temporal encoder module

The change of acetic acid reaction sequence over time is an important indicator for doctors in diagnosing cervical lesions. In order to make full use of these temporal features, we introduce the GRU-based Temporal Encoder Module, which can find the global relationship between different frames. To verify the effectiveness of the Temporal Encoder Module, we perform a comparison experiment between the method with TEM and without TEM. It can be seen from Table 5 and Figure S1C, CerviFusionNet with TEM can achieve accuracy and f1-score of 0.8333 and 0.8333, which is better than CerviFusionNet without TEM. These findings emphasize that temporal information is essential for cervical lesions classification, and the TEM can significantly boost the model performance.

Table 4. The results of our proposed Hybrid Architecture compared with VGG-16 and ResNet-34 in model parameters (Params) and computational costs (FLOPs)

Methods	Params (M)	FLOPs (G)
VGG-16	138.36	15.61
ResNet-34	21.80	3.68
Our Method	**15.71**	**2.76**

Table 5. The effectiveness of proposed Temporal Encoder Module (TEM) in terms of accuracy, F1-score, precision, and recall

Methods	Accuracy	F1-Score	Precision	Recall
w.o. TEM	0.8166	0.8163	0.8216	0.8166
w. TEM	0.8333	0.8333	0.8545	0.8333

Notes: *w.o.* TEM means the CerviFusionNet without Temporal Encoder Module, while the *w*. TEM defines the CerviFusionNet with Temporal Encoder Module.

Effectiveness of different modalities in model performance

In order to explore the impact of different modalities data on the proposed method, we designed several experiments to compare the results between them. For this purpose, we conducted extensive experiments to verify the model performance with signal-modal or multi-modal datasets. Of note, the model structure would be adjusted according to the input data type. For instance, if the input data is only the original image, the Feature Extraction Branch only contains the O-Branch, while the A-Branch and I-Branch would be discarded. When the input data contains the iodine staining sample and acetic acid reaction sequences, the Feature Extraction Branch would retain the A-Branch and I-Branch but drop the O-Branch. The detailed results are shown in Table 6. Based on the results, it is evident that training with just one type of data leads to relatively poor results. However, combining two data types enhances the algorithm's performance significantly. The proposed method can achieve the best accuracy when feeding input data with three modalities. This finding highlights the importance of multi-modal data in cervical lesions classification. Additionally, we observed that the acetic acid reaction sequences had a more significant impact on the results compared to the other two data types, which is aligned with doctors' experience.

Table 6. The model performance with different combination of data type in terms of accuracy, F1-Score, precision, and recall

Original	Iodine	Acid	Accuracy	F1-Score	Precision	Recall
✓	–	–	0.7166	0.7173	0.7182	0.7166
–	✓	–	0.7333	0.7321	0.7441	0.7333
–	–	✓	0.7833	0.7806	0.7799	0.7833
✓	✓	–	0.7750	0.7736	0.7787	0.7750
✓	–	✓	0.8250	0.8247	0.8267	0.8250
–	✓	✓	0.8166	0.8169	0.8184	0.8166
✓	✓	✓	0.8333	0.8333	0.8545	0.8333

Table 7. The model performance of CerviFusionNet with different single-modal data, including original image (Original), Acetic acid samples of 60s (Acid-60s) and 120s (Acid-120s), and Iodine images (Iodine)

Data Type	Accuracy	F1-Score	Precision	Recall
Original	0.7166	0.7173	0.7182	0.7166
Acid-60s	0.7416	0.7416	0.7416	0.7416
Acid-120s	0.7583	0.7563	0.7587	0.7583
Iodine	0.7333	0.7321	0.7182	0.7166

DISCUSSION

Accurate classification of cervical lesions is crucial for precise patient treatment. Automated cervical lesion diagnosis systems based on AI technology have shown great potential, but they still need to overcome many challenges. In this paper, a three-category classification method for cervical lesions was developed, including Normal, LSIL, and HSIL.

Due to expensive data collection and complex model design, most CAD systems for cervical lesion classification are built on single-modal datasets. However, the single-modal dataset may lose much useful information, making it difficult for existing methods to achieve high accuracy in cervical lesion classification. Some studies indicate that multi-modal CAD systems for cervical lesions classification generally outperform single-modal ones. In the field of colposcopy diagnosis, the various modalities of data contain different information, which contributes differently to the final results. We conducted experiments to explore the contribution of different modalities to the final results. For this purpose, we used different single-modal datasets like original images, iodine staining samples, acetic acid reaction of 60 s, and acetic acid reaction of 120 s. The results of these experiments are presented in Table 7 and Figure S1B. Based on the observations, the model trained with acetic acid reaction samples outperformed other models trained with original images and iodine staining samples. This indicates that the acetic acid reaction samples contain more effective information for classifying cervical lesions. However, it is essential to note that single-modal methods still fall behind multi-modal approaches regarding accuracy and robustness.

The deep learning models have strong feature extraction capabilities, which are very beneficial for analyzing medical images, especially for colposcopy samples. However, interpreting these models remains challenging. Our proposed CerviFusionNet leverages the multi-modal dataset to output precise cervical lesion classifications. Nevertheless, it is hard to determine which features contribute to the final results. To address this problem, we applied Grad-CAM to visualize the extracted features across the three feature extraction branches, and the results are shown in Figure 5. The model focuses on the cervical orifice for pathological Normal cases, while for disease cases (LSIL/HSIL), it looks at different regions beyond just the orifice. In the HSIL case, the model identifies cervical intraepithelial lesions, and its attention range changes over time for acetic acid reactions. These results emphasize that the visualization of attention localization demonstrates that our model successfully extracts the key diagnostic information from the multi-modal inputs, emulating experienced doctors' evaluation process.

Limitations of the study

In this study, we introduced a large-scale public multi-modal colposcopy dataset for cervical lesion classification. Based on this dataset, we developed CerviFusionNet, a framework for multi-class cervical lesion classification. After validating model

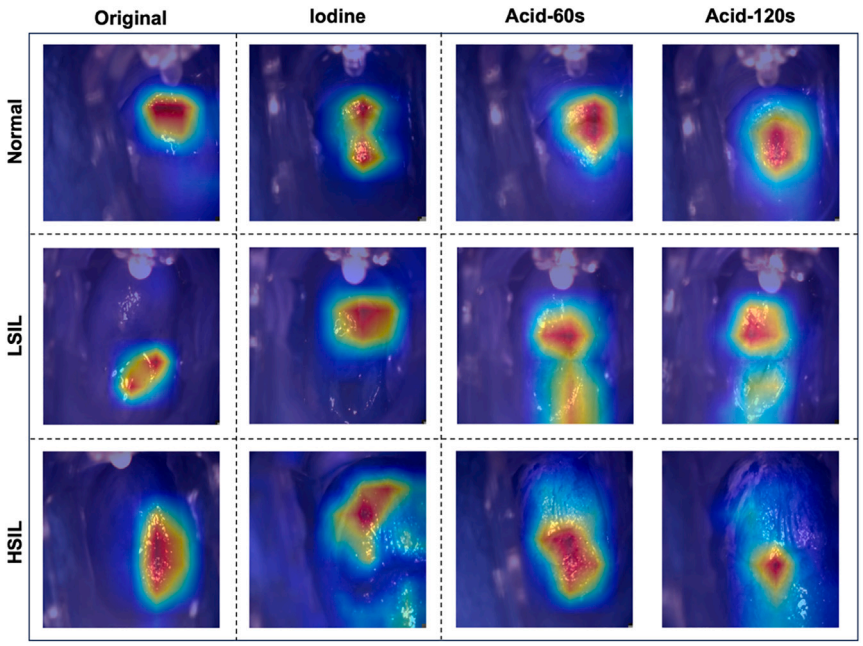


Figure 5. Using Grad-CAM for heatmap visualization highlights areas important for model predictions, enhancing interpretability and explainability

performance on internal and external datasets, we found that CerviFusionNet achieves promising prediction accuracy and exhibits advantages in computational efficiency. However, there are still many challenges that need to be addressed. First, the proposed CerviFusionNet only uses multi-modal image data as input and does not incorporate clinical text information, such as the HVP test, TCT, and age. In future work, we aim to expand CerviFusionNet to leverage more clinical text information along with colposcopy images, which could further enhance diagnostic performance. Second, we can develop Large Vision Model (LVM) to understand the pair of colposcopy image and clinical text, then employ the pretrained LVM to annotate plentiful of unlabeled sample. Currently, the collected multi-modal dataset was collected from a single center. In the next step, we will expand this dataset into a multicenter cohort for research and improve our model to improve generalization performance. Furthermore, the CerviFusionNet for classify hard cases will be studied. We will also explore the application of proposed method cervical lesions classification studies.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, J.M. (email: majing2609@163.com).

Materials availability

Materials are available upon request to J.M. (majing2609@163.com).

Data and code availability

- The raw medical images can be made available for research purposes upon reasonable request to the lead author, following appropriate ethical clearance and data sharing agreements.
- Source codes have been deposited at Github. (https://github.com/Li-OmicsLab/CerviFusionNet/tree/main/Source/).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We acknowledge the pathologists for assisting with the data collection. This work was supported by the fund from Macao Polytechnic University (RP/FCA-14/2023), and The Science and Technology Development Funds (FDCT) of Macao (0033/2023/RIB2).

AUTHOR CONTRIBUTIONS

Concept and design: Y.S., K.L., Y.W., and J.M. Acquisition, analysis, or interpretation of the data: Y.S., Q.Z., M.H., X.Z., J.L., and W.M. Drafting of the manuscript: Y.S. and Q.Z. Critical revision of the manuscript for important intellectual content: K.L., Y.W., and J.M. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.111313.

Received: February 18, 2024
Revised: June 10, 2024
Accepted: October 30, 2024
Published: November 2, 2024

REFERENCES

1. Perkins, R.B., Wentzensen, N., Guido, R.S., and Schiffman, M. (2023). Cervical cancer screening: a review. JAMA *330*, 547–558. https://doi.org/10.1001/jama.2023.13174.
2. Singh, D., Vignat, J., Lorenzoni, V., Eslahi, M., Ginsburg, O., Lauby-Secretan, B., Arbyn, M., Basu, P., Bray, F., and Vaccarella, S. (2023). Global estimates of incidence and mortality of cervical cancer in 2020: a baseline analysis of the WHO Global Cervical Cancer Elimination Initiative. Lancet Global Health *11*, e197–e206. https://doi.org/10.1016/S2214-109X(22)00501-0.
3. Xue, P., Wang, J., Qin, D., Yan, H., Qu, Y., Seery, S., Jiang, Y., and Qiao, Y. (2022). Deep learning in image-based breast and cervical cancer detection: a systematic review and meta-analysis. npj Digit. Med. *5*, 19. https://doi.org/10.1038/s41746-022-00559-z.
4. Ginsburg, O., Badwe, R., Boyle, P., Derricks, G., Dare, A., Evans, T., Eniu, A., Jimenez, J., Kutluk, T., Lopes, G., et al. (2017). Changing global policy to deliver safe, equitable, and affordable care for women's cancers. Lacent *389*, 871–880. https://doi.org/10.1016/S0140-6736(16)31393-9.
5. Wentzensen, N., Chirenje, Z.M., and Prendiville, W. (2021). Treatment approaches for women with positive cervical screening results in low-and middle-income countries. Prev. Med. *144*, 106439. https://doi.org/10.1016/j.ypmed.2021.106439.
6. Brisson, M., Kim, J.J., Canfell, K., Drolet, M., Gingras, G., Burger, E.A., Martin, D., Simms, K.T., Bénard, É., Boily, M.-C., et al. (2020). Impact of HPV vaccination and cervical screening on cervical cancer elimination: a comparative modelling analysis in 78 low-income and lower-middle-income countries. Lancet *395*, 575–590. https://doi.org/10.1016/S0140-6736(20)30068-4.
7. Mo, Y., Han, C., Liu, Y., Liu, M., Shi, Z., Lin, J., Zhao, B., Huang, C., Qiu, B., Cui, Y., et al. (2023). HoVer-Trans: Anatomy-Aware HoVer-Transformer for ROI-Free Breast Cancer Diagnosis in Ultrasound Images. IEEE Trans. Med. Imag. *42*, 1696–1706. https://doi.org/10.1109/TMI.2023.3236011.
8. Shen, Y., Shamout, F.E., Oliver, J.R., Witowski, J., Kannan, K., Park, J., Wu, N., Huddleston, C., Wolfson, S., Millet, A., et al. (2021). Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. Nat. Commun. *12*, 5645. https://doi.org/10.1038/s41467-021-26023-2.
9. Jiang, P., Li, X., Shen, H., Chen, Y., Wang, L., Chen, H., Feng, J., and Liu, J. (2023). A systematic review of deep learning-based cervical cytology screening: from cell identification to whole slide image analysis. Artif. Intell. Rev. *56*, 2687–2758. https://doi.org/10.1007/s10462-023-10588-z.
10. Demarco, M., Hyun, N., Carter-Pokras, O., Raine-Bennett, T.R., Cheung, L., Chen, X., Hammer, A., Campos, N., Kinney, W., Gage, J.C., et al. (2020). A study of type-specific HPV natural history and implications for

contemporary cervical cancer screening programs. eClinicalMedicine *22*, 100293. https://doi.org/10.1016/j.eclinm.2020.100293.

11. Fan, J., Fu, Y., Peng, W., Li, X., Shen, Y., Guo, E., Lu, F., Zhou, S., Liu, S., Yang, B., et al. (2023). Multi-omics characterization of silent and productive HPV integration in cervical cancer. Cell Genom. *3*, 100211. https://doi.org/10.1016/j.xgen.2022.100211.
12. Solomon, D., Davey, D., Kurman, R., Moriarty, A., O'Connor, D., Prey, M., Raab, S., Sherman, M., Wilbur, D., Wright, T., Jr., et al. (2002). The 2001 Bethesda System: terminology for reporting results of cervical cytology. JAMA *287*, 2114–2119. https://doi.org/10.1001/jama.287.16.2114.
13. Zhu, X., Li, X., Ong, K., Zhang, W., Li, W., Li, L., Young, D., Su, Y., Shang, B., Peng, L., et al. (2021). Hybrid AI-assistive diagnostic model permits rapid TBS classification of cervical liquid-based thin-layer cell smears. Nat. Commun. *12*, 3541. https://doi.org/10.1038/s41467-021-23913-3.
14. Cheng, S., Liu, S., Yu, J., Rao, G., Xiao, Y., Han, W., Zhu, W., Lv, X., Li, N., Cai, J., et al. (2021). Robust whole slide image analysis for cervical cancer screening using deep learning. Nat. Commun. *12*, 5639. https://doi.org/10.1038/s41467-021-25296-x.
15. Hu, L., Bell, D., Antani, S., Xue, Z., Yu, K., Horning, M.P., Gachuhi, N., Wilson, B., Jaiswal, M.S., Befano, B., et al. (2019). An Observational Study of Deep Learning and Automated Evaluation of Cervical Images for Cancer Screening. J. Natl. Cancer Inst. *111*, 923–932. https://doi.org/10.1093/jnci/djy225.
16. Organization, W.H. (2013). WHO guidelines for screening and treatment of precancerous lesions for cervical cancer prevention (World Health Organization). https://www.who.int/publications/i/item/9789240030824.
17. Melnikow, J., Henderson, J.T., Burda, B.U., Senger, C.A., Durbin, S., and Weyrich, M.S. (2018). Screening for Cervical Cancer With High-Risk Human Papillomavirus Testing: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. JAMA *320*, 687–705. https://doi.org/10.1001/jama.2018.10400.
18. Verdoodt, F., Jentschke, M., Hillemanns, P., Racey, C.S., Snijders, P.J.F., and Arbyn, M. (2015). Reaching women who do not participate in the regular cervical cancer screening programme by offering self-sampling kits: A systematic review and meta-analysis of randomised trials. Eur. J. Cancer *51*, 2375–2385. https://doi.org/10.1016/j.ejca.2015.07.006.
19. Li, Y., Chen, J., Xue, P., Tang, C., Chang, J., Chu, C., Ma, K., Li, Q., Zheng, Y., and Qiao, Y. (2020). Computer-aided cervical cancer diagnosis using time-lapsed colposcopic images. IEEE Trans. Med. Imag. *39*, 3403–3415. https://doi.org/10.1109/TMI.2020.2994778.
20. Zhang, Y., Yin, Y., Liu, Z., and Zimmermann, R. (2021). A spatial regulated patch-wise approach for cervical dysplasia diagnosis. In Proceedings of the AAAI Conference on Artificial Intelligence, pp. 733–740. https://doi.org/10.1609/aaai.v35i1.16154.
21. Habib, Z., Mughal, M.A., Khan, M.A., Hamza, A., Alturki, N., and Jamel, L. (2024). A novel deep dual self-attention and Bi-LSTM fusion framework for Parkinson's disease prediction using freezing of gait: a biometric application. Multimed. Tool. Appl. *83*, 80179–80200. https://doi.org/10.1007/s11042-024-18906-5.
22. Ullah, M.S., Khan, M.A., Almujally, N.A., Alhaisoni, M., Akram, T., and Shabaz, M. (2024). BrainNet: a fusion assisted novel optimal framework of residual blocks and stacked autoencoders for multimodal brain tumor classification. Sci. Rep. *14*, 5895. https://doi.org/10.1038/s41598-024-56657-3.
23. Sha, Y., Meng, W., Zhai, X., Xie, C., and Li, K. (2024). Accurate Facial Landmark Detector via Multi-scale Transformer. In Pattern Recognition and Computer Vision, pp. 278–290. https://doi.org/10.1007/978-981-99-8469-5_22.
24. Sha, Y., Zhai, X., Li, J., Meng, W., Tong, H.H., and Li, K. (2023). A novel lightweight deep learning fall detection system based on global-local attention and channel feature augmentation. Interdiscip. Nurs. Res. *2*, 68–75. https://doi.org/10.1097/NR9.0000000000000026.
25. Sha, Y. (2021). Efficient Facial Landmark Detector by Knowledge Distillation. In IEEE International Conference on Automatic Face and Gesture Recognition, pp. 1–8. https://doi.org/10.1109/FG52635.2021.9667036.
26. Sha, Y., Meng, W., Luo, G., Zhai, X., Tong, H.H.Y., Wang, Y., and Li, K. (2024). MetDIT: Transforming and Analyzing Clinical Metabolomics Data with Convolutional Neural Networks. Anal. Chem. *96*, 2949–2957. https://doi.org/10.1021/acs.analchem.3c04607.
27. Jabeen, K., Khan, M.A., Hameed, M.A., Alqahtani, O., Alouane, M.T.H., and Masood, A. (2024). A novel fusion framework of deep bottleneck residual convolutional neural network for breast cancer classification from mammogram images. Front. Oncol. *14*, 1347856. https://doi.org/10.3389/fonc.2024.1347856.
28. Ullah, M.S., Khan, M.A., Masood, A., Mzoughi, O., Saidani, O., and Alturki, N. (2024). Brain tumor classification from MRI scans: a framework of hybrid deep learning model with Bayesian optimization and quantum theory-based marine predator algorithm. Front. Oncol. *14*, 1335740. https://doi.org/10.3389/fonc.2024.1335740.
29. Kumar, Y., Koul, A., Singla, R., and Ijaz, M.F. (2023). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J. Ambient Intell. Hum. Comput. *14*, 8459–8486. https://doi.org/10.1007/s12652-021-03612-z.
30. Chen, X., Pu, X., Chen, Z., Li, L., Zhao, K.N., Liu, H., and Zhu, H. (2023). Application of EfficientNet-B0 and GRU-based deep learning on classifying the colposcopy diagnosis of precancerous cervical lesions. Cancer Med. *12*, 8690–8699. https://doi.org/10.1002/cam4.5581.
31. Tan, M., and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, pp. 6105–6114. https://arxiv.org/abs/1905.11946.
32. Yu, Y., Ma, J., Zhao, W., Li, Z., and Ding, S. (2021). MSCI: A multistate dataset for colposcopy image classification of cervical cancer screening. Int. J. Med. Inf. *146*, 104352. https://doi.org/10.1016/j.ijmedinf.2020.104352.
33. Li, J., Hu, P., Gao, H., Shen, N., and Hua, K. (2024). Classification of cervical lesions based on multimodal features fusion. Comput. Biol. Med. *177*, 108589. https://doi.org/10.1016/j.compbiomed.2024.108589.
34. Zhang, Y., Yin, Y., Zhang, Y., Liu, Z., Wang, Z., and Zimmermann, R. (2023). Prototypical Cross-domain Knowledge Transfer for Cervical Dysplasia Visual Inspection. In Proceedings of the 31st ACM International Conference on Multimedia, pp. 1504–1514. https://doi.org/10.1145/3581783.3612000.
35. Fan, Y., Ma, H., Fu, Y., Liang, X., Yu, H., and Liu, Y. (2022). Colposcopic multimodal fusion for the classification of cervical lesions. Phys. Med. Biol. *67*, 135003. https://doi.org/10.1088/1361-6560/ac73d4.
36. Yuan, C., Yao, Y., Cheng, B., Cheng, Y., Li, Y., Li, Y., Liu, X., Cheng, X., Xie, X., Wu, J., et al. (2020). The application of deep learning based diagnostic system to cervical squamous intraepithelial lesions recognition in colposcopy images. Sci. Rep. *10*, 11639. https://doi.org/10.1038/s41598-020-68252-3.
37. Simonyan, K., and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of 3rd International Conference on Learning Representations, pp. 1–14. https://doi.org/10.48550/arXiv.1409.1556.
38. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.
39. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al. (2021). Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. *43*, 3349–3364. https://doi.org/10.1109/TPAMI.2020.2983686.
40. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted

windows. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986.

41. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2010.11929.

42. Li, Y., Hu, J., Wen, Y., Evangelidis, G., Salahi, K., Wang, Y., Tulyakov, S., and Ren, J. (2023). Rethinking Vision Transformers for MobileNet Size and Speed. In Proceedings of the IEEE Conference on International Conference on Computer Vision, pp. 16889–16900. https://doi.org/10.1109/ICCV51070.2023.01549.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited dataset		
CerviFusionNet	This paper	https://github.com/Li-OmicsLab/CerviFusionNet/
Multi-modal cervical lesions dataset	This paper	https://github.com/Li-OmicsLab/CerviFusionNet/tree/main/Dataset/
Source code	This paper	https://github.com/Li-OmicsLab/CerviFusionNet/tree/main/Source/
Software and algorithms		
Python (version: 3.10)	Python software	https://pytorch.org/
PyTorch (version: 1.10.0)	PyTorch software	https://pytorch.org/
Cuda (version: 11.6.0)	Nvidia	https://developer.nvidia.com/
Numpy (version: 1.23.5)	Numpy package	https://numpy.org/
Pandas (version: 1.21.5)	Pandas package	https://pandas.pydata.org/
GradCam	Jacob et al.	https://github.com/jacobgil/pytorch-grad-cam/

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Ethical statement

The study was approved by the Institutional Review Board (IRB) of First Teaching Hospital of Tianjin University of Traditional Chinese Medicine (TYLL2023-Z-045). The written consent was obtained from all the subjects upon image collection.

Datasets

The dataset includes information about 2,273 patients who visited the gynecological colposcopy clinic at the First Teaching Hospital of Tianjin University of Traditional Chinese Medicine between July 2021 and August 2023. After analysis, 701 patients are meeting the requirements for data collection. All patients perform colposcopy examinations and biopsies. Based on the pathological reports, patients were classified into three categories: pathological normal (Normal), low-grade squamous intraepithelial lesion (LSIL), and high-grade squamous intraepithelial lesion (HSIL). Specifically, this dataset includes 253 HPV-positive without cervical lesions cases, 234 LSIL cases, and 214 HSIL cases, totaling 2,804 images. Each patient has four colposcope images, including original colposcopy images (Original), acetic acid reaction time series at 60 and 120 s (Acid-60s, Acid-120s), and iodine staining images (Iodine). Besides, we also collected patients' clinical text information such as age, TCT and HPV status. All of the collected images are JPEG with the resolution of 1280 × 960. Magnification ranges from 1 to 40 times, depending on the patient's condition. Notable, all the colposcopy images were obtained by doctors with more than ten years of experience. The pathological reports are used as ground truth for these collected samples. All the sensitive information was removed, and the data entries were anonymized.

Implementation details

During the training phase, all input images need to be rescaled to the resolution of 224 × 224. The data augmentation technologies are employed to improve the model's robustness, including random rotation, occlusion, scaling, horizontal flipping, and blurring. The total epochs are 300, and the mini-batch size is 128. We choose Adam as the optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and the weight decay of 0.0001. The initial learning rate is set to 10^{-4} and reduced 10^{-6} following the cosine annealing schedule. The implementation of our method is based on PyTorch with one NVIDIA Tesla A100 GPU.

We conduct all the experiments on the collected dataset. The dataset is split 80%/20% for training/testing by patient ID. We adopted 5-fold cross-validation to verify the prediction performance of the proposed model, which can reduce the impact of partition randomness on the obtained results. We utilize the Cross Entropy Loss (CE) to compare the prediction results by our proposed method with ground truth. The CE Loss function can be defined as:

$$\mathcal{L}_{ce} = -\sum_{i} y_i \log(p_i) \quad \text{(Equation 1)}$$

where the y_i and p_i represents the ground truth and prediction result, respectively.

METHOD DETAILS

Image preprocessing

Following the standard convention for computer vision models, all the images used in the experiments are normalized using the mean and standard deviation.

Labels preprocessing

Each case collected should be annotated according to the pathological results. If the case has no cervical lesions, it should be marked as '0'; if it is LSIL, it should be marked as '1'; if it is HSIL, it should be marked as '2'.

Proposed method

Overview

Colposcopy can effectively reduce the incidence of cervical cancer, but it requires experienced doctors. Existing CAD systems for cervical lesions classification via colposcopy images have made considerable progress, which could potentially lessen the workload of colposcopy doctors. However, most classification systems are based on single-modal data, making it challenging to obtain high-accuracy predictions in real-world scenarios. Therefore, we propose a cervical lesions classification system based on deep learning and multi-modal data named CerviFusionNet. Figure 1 shows that the proposed method takes advantage of the original image, acetic acid reaction sequence, and iodine images to classify cervical lesions into Normal, LSIL, and HISL. The proposed CerviFusionNet consists of three parts: Stem Block, Hybrid Architecture, and Prediction Head. The Stem Block can extract multi-scale detailed information from input samples and reduce the resolution quickly. Hybrid Architecture is responsible for extracting robustness representations following the Stem Block. Specifically, the Hybrid Architecture can leverage the advantages of CNN and ViT for feature extraction. The Prediction Head is formed by the Feature Fusion Module and a group of MLP. The Feature Fusion Module is employed to fuse these extracted features and generate refined representations, forcing the model to pay more attention to some important information and enhance the feature discriminability. Finally, the fused representations are mapped into cervical lesions classification results by MLP.

Stem Block

The computational costs of deep learning models are closely linked to their operational efficiency. Cervical lesion classification requires highly efficient models, so it is important to prioritize computational costs. Additionally, the Hybrid Architecture, combining with CNN and ViT, needs to address ViT's limitation in encoding local information. Inspired by previous works, we designed the Stem Block to address these issues. The specifics of the Stem Block are illustrated in Figure 3B. As we can see, the Stem Block is formed by three convolutional layers. We arrange two convolutional layers with the stride of 2 for downsizing, then utilize one convolutional layer with the stride of 1 to get discriminated representations. As a result, the Stem Block can downsize the input sample by four times. The fast down-sampling strategy in the Stem Block significantly reduces the computational costs and optimizes the operating efficiency. Notably, we use one set of Stem Block weights to process data from all three modalities without any adjustments.

Hybrid Architecture

In order to simulate the diagnostic process of experienced colposcopy doctors, the proposed CerviFusionNet should encode three types of input samples simultaneously, including original images, acetic acid reaction sequence, and iodine staining samples. It is challenging for classic backbones to process these multi-modal data efficiently. To address this, we have developed a Hybrid Architecture using CNN and ViT to extract distinct representations from multi-modal colposcopy images.

The details of Hybrid Architecture are shown in Figures 1 and 3A. Specifically, Hybrid Architecture follows the Stem Block, which consists of three feature extraction branches, including O-Branch, A-Branch, and I-Branch. Among them, O-Branch for processing original images, A-Branch should handle acetic acid reaction sequence, and I-Branch for extracting features from iodine staining samples. Notable, the original image and iodine staining sample are single-frame data. Therefore, the structure of O-Branch and I-Branch are similar. However, the A-Branch must extract robust representation from acid reaction sequences and establish the relationship between different frames. Compared with I-Branch or O-Branch, the structure of A-Branch is more complex, including multiple feature extraction modules and a temporal encoder module. Specifically, the Temporal Encoder Module (TEM) comprises bidirectional Gated Recurrent Units (GRU) that output latent embedding and contain information incorporated from different frames in the acetic acid reaction sequence. We take the O-Branch as an example to describe the detailed workflow of these feature extraction branches, which are shown in Figures 3A and 3D. As we can see, the proposed CTBlock is the main component of the O-Branch, which is used to tackle both local redundancy and global dependency for effective and efficient medical image-related tasks, especially for the classification of cervical lesions. Precisely, our CTBlock consists of three key modules: Local Feature Unit (LFU), Multi-head Self-attention Encoder (MHSA), and Residual Feedforward Network (RFFD). The following sections will provide detailed descriptions of the LFU, MHSA, and RFFD. Besides, we also report the details of TEM.

Local feature unit (LFU). Data augmentation techniques, such as rotation and shift, are widely used in computer vision tasks to enhance visual feature diversity and boost model generalization. However, Transformer-based approaches require position encoding to determine the absolute position of each input patch, which damages the translation-invariance and rotation-invariance of the input image. Ignoring the space-invariance of image may affect the model performance and reduce prediction accuracy. Additionally, ViT often emphasizes extracting global information while overlooking local representation and structure relationships within the

patch. In order to make full use of the capacity of the Transformer-based structure, we introduce the LFU to get the local features and detailed representations, which can enhance the performance of the Vision Transformer Encoder. Given an input feature, the LFU extracted representations can be defined as follows:

$$LFU(X) = Conv(X) + X \quad \text{(Equation 2)}$$

where the $X \in \mathbb{R}^{H\times W\times c}$, $H\times W$ and C defines the resolution and the dimension of input feature map respectively. The $Conv(\cdot)$ denotes the standard Convolution layer.

Multi-head self-attention encoder (MHSA). We choose the proposed LFU is used to extract the local features and detailed representations, which are essential but hard to be explored by ViT directly. Following the LFU, we employ Transformer encoder to capture the long-range relationships between different feature maps. Specifically, the CNN module to extract the low-level feature $F_c \in \mathbb{R}^{H_c\times W_c\times N_c}$, where the N_c defines the dimension of F_c. Then, these feature maps are flattened into sequences $F_s \in \mathbb{R}^{N_c\times L}$, where $L = H_c \times W_c$. Finally, F_s go through several multi-head self-attention encoders to get the refined representations. In our setting, the input feature maps should project into queries (Q), keys (K), and values (V), the self-attention operation can be calculated as follow:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad \text{(Equation 3)}$$

where, d_k is the dimension of the query and key, which is essential for maintaining gradient stability during model training. The Transformer module generates high-level representations $F_t \in \mathbb{R}^{d\times L}$, which contains global information and learned potential dependency. Self-attention operations allow the model to focus on highly relevant features for the final results by exploring intrinsic connections between features. The MHSA contains several parallel self-attention heads, which can be defined as follows:

$$Head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad \text{(Equation 4)}$$

$$MHSA(Q,K,V) = Concat(Head_1, Head_2, \ldots, Head_n)W_{MHSA} \quad \text{(Equation 5)}$$

Where, the W_i^Q, W_i^K, W_i^V, and W_{MHSA} are the matrices of MHSA.

Refine feedforward network (RFFN). Following ViT,[41,42] the FeedForward Network (FFN) is placed behind the MHSA, which contains two linear layers and a GeLU activation function. In this paper, we propose a scheme with some Convolution layer instead of classical Feedforward Network (FFN) named Refine Feedforward Network (RFFN), which is demonstrated in Figure 3C. This change allows the model to incorporate local information, improving performance. The addition of Convolutional layers helps the vision transformer to capture detailed local information, which is beneficial to accurate cervical lesions prediction. The definition of RFFN is shown as follow:

$$RFFN(X) = Conv(F(X)) + X \quad \text{(Equation 6)}$$

$$F(X) = DWConv(DWConv(X)) \quad \text{(Equation 7)}$$

where the $DWConv(\cdot)$ represents the Depth-wise Convolution for get local information with limited model parameters and computations costs.

Temporal encoder module. Acetowhite opacity is significantly correlated with cervical lesions grades in cervical screening. The doctors repeatedly compare the original image with the acetic acid reaction sequence and analyze the changes in the acetowhite epithelium to make a preliminary judgment. Therefore, the acetic acid reaction process is an important indicator for colposcopy doctors to judge the degree of cervical lesions. Based on this assumption, we propose the Temporal Encoder Module, formed by two GRU. The temporal encoder module is placed after the feature extraction module in the A-branch. It is responsible for producing a latent vector that contains temporal dependency between input sequences.

Prediction head

The input sample goes through three feature extraction branches, which generate a set of feature maps: F_O, F_A, and F_I. These maps define the representations created from O-Branch, A-Branch, and I-Branch, respectively. Since the input images belong to the same patient, there exists some correspondence between these input feature maps. Therefore, we utilize the Convolution layer to fuse these input features and output refined representations with more discriminability and robustness, which is beneficial to optimizing model classification performance. As shown in Figure 3E, three feature maps are initially concatenated and then processed with a Convolution layer with the kernel size of 1. The specific calculation process is shown as follows:

$$F_{fuse} = Conv(Cat(F_o, F_A, F_I)) \quad \text{(Equation 8)}$$

where, the F_{fuse} defines the refined feature. Then, F_{fuse} should passed through an MLP to obtain cervical lesion results, which include Normal, LSIL, and HSIL.

QUANTIFICATION AND STATISTICAL ANALYSIS

We evaluate model performance by four commonly used metrics like accuracy, F1-score, precision, and recall. Definitions of these metrics are provided below.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad \text{(Equation 9)}$$

$$Precision = \frac{TP}{TN+FP} \quad \text{(Equation 10)}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{(Equation 11)}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad \text{(Equation 12)}$$

where *TP* stands for the true positive, which means the patient belongs to the positive category and is correctly classified. *FN* is the false negative, which refers to the patient in the positive category but is predicted to be negative. Similarly, *TN* and *FP* represent true negatives and false positives, respectively. Their definitions are similar to those of *TP* and *FN* mentioned above. The F1 score is an evaluation metric that combines precision and recall, providing a balanced reflection of the model's overall performance. In addition, we use the number of parameters (Params) and floating-point operations (FLOPs) to measure the complexity of deep learning model.