A Study on Ensemble Learning for Cervical Cytology Classification

Van-Khanh Tran[1⋆], Thai-Hoc Nguyen[1†], Xuan-Lam Dinh[1], and
Chi-Cuong Nghiem[2]

[1] Institute of Applied Science and Technology - IAST
University of Information and Communication Technology, Thai Nguyen University
{tvkhanh, dxlam}@ictu.edu.vn; thaihocit02@gmail.com
[2] Department of Pathology, Hospital A Thai Nguyen
bscuongbva@gmail.com

Abstract. Cervical cancer remains one of the leading causes of mortality among women, which requires early detection and treatment to mitigate its impact. Recent advancements in medical image classification have demonstrated significant efficacy, with ensemble learning strategies playing a crucial role. Ensemble learning takes advantage of the combined strengths of multiple models to improve classification accuracy by creating a stronger and more accurate predictive model. This study presents an ensemble learning approach incorporating preprocessing techniques, image enhancement methods, and six diverse convolutional neural network (CNN) architectures for the classification of cervical cytology images from a Vietnamese dataset. Our ensemble models demonstrated superior classification performances across various metrics. Moreover, we observed a significant influence of image size variations on model efficacy, highlighting the importance of standardized image preprocessing.

Keywords: Cervical Cytology Screening · Ensemble Learning · Deep Learning.

1 Introduction

Cervical cancer originates in the lower part of the uterus and is often associated with viral infections, carrying a high risk of transmission through sexual contact. It is the second leading cause of death from malignant diseases in women [2]. Regular screening, like other preventive measures, can significantly reduce mortality by enabling early detection [7]. The Pap smear test is a crucial screening procedure for identifying cancerous or precancerous cells in the cervix. It involves gently scraping a sample of cervical cells, spreading them on a glass slide with a solution, and examining the sample under a microscope.

In recent years, several research articles have explored the early detection of cervical cancer using machine learning techniques [1], [13]. These methods are

[†] First batch student in IAST Young Talent Program, 2024
[⋆] Corresponding author: tvkhanh@ictu.edu.vn

typically trained on specific datasets to extract relevant features for classification purposes. Deep neural networks have become popular in computer vision tasks [10], demonstrating remarkable results compared to traditional machine learning algorithms. These architectures exhibit strong predictive capabilities and performance comparable to that of clinical experts [8]. Classifying cervical cancer cell abnormality involves labeling entire images with predefined classes, thereby aiding cytologists in decision-making to enhance diagnostic reliability or automate processes to reduce time.

This study analyzes the impact of ensemble learning techniques [4] on the classification performance of cervical cancer cell abnormality, comparing their effectiveness with individual models. Additionally, it explores how image size influences model performance. These experiments aim to elucidate the strengths and weaknesses of single models trained on datasets with varying image sizes and their complementary roles in ensemble learning.

2 Related Works

In recent years, the application of deep learning to medical image analysis has gained considerable traction, particularly in the classification of cervical cytology images. [2] conducted a comprehensive meta-analysis of cervical cancer screening strategies, emphasizing the potential of automated systems to enhance diagnostic accuracy. Advances in convolutional neural networks (CNNs) have facilitated significant improvements in image-based diagnostics. A study by [11] demonstrated the efficacy of CNNs in detecting cervical abnormalities, achieving high accuracy in differentiating between normal and pathological samples

Recent developments have focused on ensemble learning techniques, which integrate multiple models to boost performance. Authors in [3] explored various ensemble configurations for the classification of cervical cell images, highlighting the benefits of model diversity in achieving robust predictions. Similarly, the work of [20] leveraged ensemble approaches combining different CNN architectures, which resulted in improved classification accuracy and reliability in Pap smear analysis

Building on these advancements, our study integrates six distinct CNN architectures, such as MobileNetV2 [18], InceptionResNetV2 [17], InceptionV3 [15], VGG16 [19], ResNet101 [14], and Xception [5] into an ensemble learning framework to classify cervical cytology images. By incorporating preprocessing and image enhancement techniques, we aim to enhance the model's ability to accurately identify and categorize cervical abnormalities in a Vietnamese dataset.

3 Methods

3.1 Datasets

In this study, we utilized a cervical cancer dataset from hospital A, Thai Nguyen, comprising 15,645 images of five common cervical cell abnormalities such as

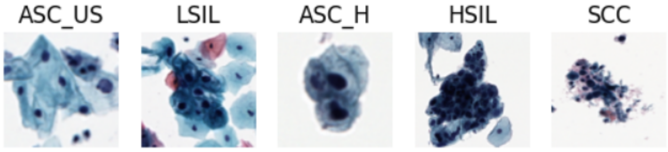

Fig. 1: Five abnormal cell samples in the dataset.

ASC_US, LSIL, ASC_H, HSIL, and SCC. This dataset includes images of suspected diseased cells captured from various angles in variable-sizes. An overview of the dataset can be seen in Fig. 1.

3.2 Sampling and Preprocessing

To address the variability in image sizes and the inconsistencies in quality present within our dataset, we implemented a comprehensive set of preprocessing techniques designed to optimize the learning process and introduce data diversity. During preprocessing, we utilized batch-wise image augmentation strategies. Specifically, within each training batch, images underwent a series of transformations, including flipping, rotation, and adjustments in brightness, contrast, and scale. These augmentations were systematically applied to enhance the robustness of the model by simulating various imaging conditions and augmenting the dataset with a wider range of visual features.

Furthermore, we standardized the image resolutions to a default input size of 224x224 pixels for most model architectures, except for InceptionResNetV2 and Xception, which used 299x299 pixels [6]. This adjustment was crucial as each model architecture performs optimally with specific input resolutions, impacting model performance significantly if trained and used with mismatched resolutions. We conducted experiments to validate this finding.

Prior to model training, we normalized all images to a pixel intensity range between 0 and 1 to standardize input data and facilitate effective learning. The dataset was splitted into three subsets: training, validation, and test sets with 80%, 10%, and 10% of the data, respectively. The training set was utilized for model training, enabling the networks to learn feature representations. The validation set provided a mechanism for monitoring performance and fine-tuning model parameters, while the test set was reserved for the final evaluation to measure model performance post-training objectively.

3.3 Classification Models

We conducted extensive experiments on a diverse set of classification architectures to ensure reliable results. The following architectures were chosen: MobileNetV2 [18], InceptionResNetV2 [17], InceptionV3 [15], VGG16 [19], ResNet101 [14], and Xception [5]. These models were pretrained on the ImageNet dataset [16]. During models implementation, we kept the weights frozen in most layers and made adjustments in a few layers to adapt the model output with softmax activation and dropout layers to mitigate overfitting. The training process was conducted over 100 epochs, employing the Adam optimizer configured with a learning rate of 1e-4. To prevent overfitting, early stopping and regularization were incorporated during the fine-tuning stages. The training was halted if no improvement in validation accuracy was observed for 10 consecutive epochs, and the best model checkpoints on the dev set was retained. We set a batch size of 16, striking a balance between computational efficiency and the stability of gradient descent. This approach aimed to leverage the strengths of pretrained models while fine-tuning them to effectively classify images of cervical cancer abnormality, ensuring robust performance across different architectures.

3.4 Ensemble Learning

Ensemble learning methods combine individual models, each generating separate predictions, to formulate a consolidated inference. Several ensemble techniques are prevalent, including Boosting, Bagging, Stacking, and Augmenting. In this study, we concentrate on the Stacking technique (Fig. 2), a form of heterogeneous ensemble learning that has demonstrated significant advantages in enhancing overall performance [4]. Stacking involves leveraging diverse base models and combining their outputs through methods such as voting or weighted averaging to produce a unified prediction. This approach, while more intricate, facilitates the integration of various modeling strategies, thereby capturing a broader range of predictive insights and improving the robustness of the final model.

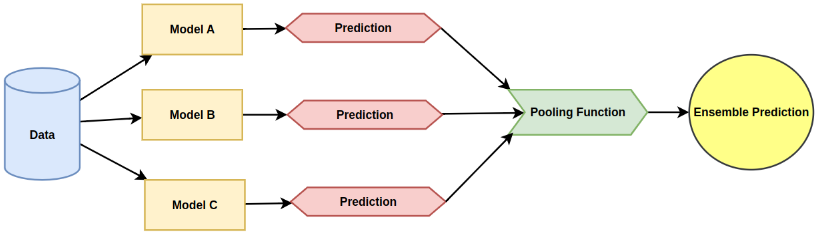

Fig. 2: Ensemble Learning

Given the heterogeneous quality and variable sizes of images in our dataset, our models necessitate inputs of differing dimensions. Our approach involves training distinct models on various input sizes present within the dataset. For each input size, the model demonstrating the highest performance is selected. These high-performing models are subsequently integrated into an ensemble framework, exploiting their complementary strengths to manage the diversity in input sizes and image quality.

This strategy aims to improve the robustness and accuracy of our cervical cancer classification system by effectively harnessing the strengths of multiple models, each trained on different input sizes and image qualities. By doing so, we enhance the system's ability to generalize across the diverse and variable dataset, leading to improved diagnostic performance in real-world scenarios.

3.5 Pooling Functions

To aggregate ensemble predictions into a unified inference, we investigated multiple methodologies and algorithms, including Voting, Fuzzy Distance [12], K-Nearest Neighbor, Naive Bayes, Decision Tree, Support Vector Machine, and Logistic Regression. Predicted outputs were computed using the softmax function for a given sample. In the Voting method, the final prediction is based on the majority class predicted by the individual models. Fuzzy Distance, a novel ensemble method, minimizes error values between observed and actual values by performing three distance measures: Euclidean, Manhattan (City-Block), and Cosine for each class from their respective best feasible solutions. For KNN, decision-making relied on the consensus of the nearest three predictions. These methods aimed to integrate diverse predictions from multiple models into a unified and reliable prediction for our cervical cancer classification system.

4 Results

To evaluate the models' performance, we use several metrics such as Accuracy, Precision, Recall, and F1 score [9]. The following sections present experimental results. Overall, ensemble learning methods demonstrate robust performance and significant reliability when combining multiple individual models, in which ensemble methods consistently outperformed individual models in terms of accuracy, precision, recall, and F1 score. The results highlight the critical role of input size and data preparation in model training, underlining the need for careful preprocessing to optimize classification outcomes.

4.1 Results of Single Models

Table 1 shows results of individual models trained on variable-sized cervical cell images. We can observe that the input size significantly influenced model performance, highlighting the importance of image standardization for consistent results. Certain models, such as InceptionResNetV2 and Xception, performed better with larger input sizes (224 and 256). Smaller input sizes (128) resulted in lower performance for most models, indicating the importance of higher resolution images for accurate classification. The InceptionV3 model consistently performed well across all input sizes, frequently achieving the highest or second highest accuracy. This suggests that InceptionV3 is robust to changes in input size. VGG16 and ResNet101 generally underperformed compared to InceptionV3, InceptionResNetV2, and Xception, particularly with lower accuracies at input sizes of 128 and 224.

Name	**Accuracy**	**Precision**	**Recall**	**F1 score**	**Input size**
MobileNetV2	61.29	63.42	64.43	62.64	128
InceptionV3	***67.27***	***69.21***	***70.81***	***69.91***	
InceptionResNetV2	64.22	66.60	67.83	66.80	
VGG16	51.36	56.87	54.13	49.13	
ResNet101	66.29	68.65	69.28	68.80	
Xception	**70.00**	**71.90**	**72.19**	**71.91**	
MobileNetV2	65.51	68.84	69.05	68.41	224
InceptionV3	***69.93***	***73.24***	***72.25***	***72.57***	
InceptionResNetV2	**71.62**	**74.36**	**74.12**	**74.04**	
VGG16	61.94	65.22	65.15	64.65	
ResNet101	58.83	61.08	61.91	60.41	
Xception	65.58	67.74	68.75	67.18	
MobileNetV2	65.58	67.83	68.81	67.62	256
InceptionV3	**70.12**	**72.37**	**72.57**	**72.30**	
InceptionResNetV2	***68.76***	***72.10***	70.65	***71.12***	
VGG16	67.20	69.18	70.41	69.75	
ResNet101	63.57	66.58	67.58	66.43	
Xception	67.27	68.66	***70.92***	69.39	

Table 1: Results of individual models on variable-sized images. **Bold** and ***Italic Bold*** indicate **best** and ***second best*** results for each input size.

The performance of individual models varied significantly with changes in input size. For an input size of 128, the Xception model exhibited the highest accuracy at 70.00%, followed by InceptionV3 and ResNet101 with accuracies of 67.27% and 66.29%, respectively. For an input size of 224, InceptionResNetV2 achieved the highest accuracy at 71.62%, while InceptionV3 closely followed with 69.93%. For an input size of 256, InceptionV3 showed the best performance with an accuracy of 70.12%, while InceptionResNetV2 and VGG16 achieved second and third highest accuracies of 68.76% and 67.20%, respectively.

4.2 Results of Ensemble Models

Table 2 shows results of ensemble models on variable-sized cervical cell images. The performance of ensemble methods generally improved with increasing input sizes. For an input size of 128, Logistic Regression achieved the highest accuracy of 73.63%, closely followed by Support Vector Machine (SVM) with an accuracy of 73.11%. For an input size of 224, SVM achieved the highest accuracy of 74.22%, followed by Voting and Logistic Regression with accuracies of 73.83% and 73.76%, respectively. For an input size of 256, Voting, SVM, and Logistic Regression all achieved the highest accuracy of 73.83%.

The increase in input size generally led to improved performance for most ensemble methods, highlighting the importance of higher resolution images in enhancing classification accuracy. SVM and Logistic Regression consistently performed well across all input sizes, demonstrating robustness and reliability in

Method	**Accuracy**	**Precision**	**Recall**	**F1 score**	**Input size**
Voting	72.72	74.71	75.48	74.74	128
K-Nearest Neighbor	71.81	73.80	74.67	74.20	
Naive Bayes	71.23	71.83	74.26	72.52	
Decision Tree	70.97	73.42	73.99	73.67	
Support Vector Machine	***73.11***	***74.86***	***75.83***	***75.29***	
Logistic Regression	**73.63**	**75.49**	**76.38**	**75.89**	
Voting	***73.83***	76.01	***76.44***	***76.17***	224
K-Nearest Neighbor	73.05	75.32	75.43	75.36	
Naive Bayes	73.44	74.19	76.29	75.01	
Decision Tree	72.72	74.91	75.25	75.05	
Support Vector Machine	**74.22**	**76.48**	**76.69**	**76.57**	
Logistic Regression	73.76	***76.05***	76.23	***76.14***	
Voting	73.44	75.36	76.14	75.70	256
K-Nearest Neighbor	72.72	75.14	75.36	75.23	
Naive Bayes	72.40	73.31	75.40	73.93	
Decision Tree	70.06	72.64	72.91	72.76	
Support Vector Machine	**73.83**	**76.31**	***76.33***	**76.29**	
Logistic Regression	**73.83**	***76.26***	**76.34**	**76.29**	

Table 2: Results of ensemble methods on variable-sized images. **Bold** and ***Italic Bold*** indicate **best** and ***second best*** results for each input size.

classification tasks. SVM achieved top accuracy at input sizes 224 and 256, and second highest at input size 128, while Logistic Regression showed high performance, matching the highest accuracy at input size 256. The Voting method also showed competitive performance, indicating that simple ensemble techniques can be highly effective, particularly at input size 224 where it achieved an accuracy of 73.83%.

5 Conclusion

This paper investigates ensemble learning techniques to improve the classification performance of medical images, focusing specifically on the challenge of identifying abnormalities in cervical cancer. Our study uses stacking as the chosen ensemble learning method, comparing its performance against individual model training approaches. The results highlight the effectiveness of ensemble learning in achieving notable performance gains and the influence of variable-size images on the model's performance. However, despite these advances, achieving optimal model accuracy remains a challenge, with some individual models exhibiting signs of overfitting, likely attributable to dataset quality and adequacy limitations.

Acknowledgments. Dataset used in this study was supported by Project No. ĐT/YD/08/2021 conducted at Hospital A Thai Nguyen, Vietnam.

References

1. Ali, M.M., Ahmed, K., Bui, F.M., Paul, B.K., Ibrahim, S.M., Quinn, J.M., Moni, M.A.: Machine learning-based statistical analysis for early stage detection of cervical cancer. Computers in biology and medicine **139**, 104985 (2021)
2. Arbyn, M., Weiderpass, E., Bruni, L., de Sanjosé, S., Saraiya, M., Ferlay, J., Bray, F.: Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. The Lancet Global Health **8**(2), e191–e203 (2020)
3. Dong, N., Zhao, L., Wu, C.H., Chang, J.F.: Inception v3 based cervical cell classification combined with artificially extracted features. ASC **93**, 106311 (2020)
4. Ganaie, M.A., Hu, M., Malik, A.K., Tanveer, M., Suganthan, P.N.: Ensemble deep learning: A review. Engineering Applications of AI **115**, 105151 (2022)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on CVPR. pp. 770–778 (2016)
6. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
7. Kessler, T.A.: Cervical cancer: prevention and early detection. In: Seminars in oncology nursing. vol. 33, pp. 172–183. Elsevier (2017)
8. Lee, K., Zung, J., Li, P., Jain, V., Seung, H.S.: Superhuman accuracy on the snemi3d connectomics challenge. arXiv preprint arXiv:1706.00120 (2017)
9. Lever, J., Krzywinski, M., Altman, N.: Classification evaluation. Nature (2016)
10. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
11. Mathivanan, S.K., Francis, D., Srinivasan, S., Khatavkar, V., P, K., Shah, M.A.: Enhancing cervical cancer detection and robust classification through a fusion of deep learning models. Scientific Reports **14**(1), 10812 (2024)
12. Pramanik, R., Biswas, M., Sen, S., de Souza Júnior, L.A., Papa, J.P., Sarkar, R.: A fuzzy distance-based ensemble of deep models for cervical cancer detection. Computer Methods and Programs in Biomedicine **219**, 106776 (2022)
13. Rahaman, M.M., Li, C., Yao, Y., Kulwa, F., Wu, X., Li, X., Wang, Q.: Deepcervix: A deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques. Computers in Biology and Medicine (2021)
14. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**, 211–252 (2015)
15. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
16. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
17. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision
18. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML. pp. 6105–6114. PMLR (2019)
19. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: CVPR. pp. 1492–1500 (2017)
20. Zhao, Z., Alzubaidi, L., Zhang, J., Duan, Y., Gu, Y.: A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. Expert Systems with Applications **242**, 122807 (2024)