International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. , No. (Mon-20..)

http://dx.doi.org/10.12785/ijcds/XXXXXX

Ensemble and Transformer Encoder-based Models for the Cervical Cancer Classification Using Pap-smear Images

Maysoon Mohammed Alzahrani[1], Usman Ali Khan[1] and Sultan Abdullah Al-Garni[1]

[1]*Department of Information Systems Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah , Saudi Arabia*

Received Mon. 20, Revised Mon. 20, Accepted Mon. 20, Published Mon. 20

Abstract: Cervical cancer poses a health concern for women globally ranking as the seventh most common disease and the fourth most frequent cancer among women. The classification of cytopathology images is utilized in diagnosing this condition with a focus on automating the process due to potential human errors in manual examinations.This study presents an approach that integrates transfer learning, ensemble learning and a transformer encoder to classify cervical cancer using pap smear images from the SIPaKMeD dataset. By combining these methods human involvement in the classification task is minimized. Initially individual models based on transfer learning are. Their unique characteristics are combined to create an ensemble model. This ensemble model is then input into the proposed transformer encoder specifically utilizing the Vision Transformer (ViT) model.The results highlight the effectiveness of this methodology. The VGG16 model demonstrates accuracy of 97.04% and an F1 score of 97.06% when applied to classifying five categories using the SIPaKMeD dataset. However surpassing this performance is the learning model, with an accuracy of 97.37%. Notably outperforming all models is the transformer encoder model achieving an accuracy of 97.54%.Through the utilization of transfer learning, ensemble learning and the transformer encoder model this research introduces a method, for automating the classification of cervical cancer. The findings underscore the capability of the suggested approach to enhance the precision and effectiveness of diagnosing cancer.
Keywords: Cervical cancer, Pap-smear, hybrid CAD system, ensemble learning, transformer encoder

1. INTRODUCTION

Cervical cancer is the most common and dangerous disease in women, as it is the fourth in the global prevalence level among the types of cancers for females, while it is considered the ninth in the Kingdom of Saudi Arabia [1, 2]. Since 1990, the number of new cases of cervical cancer in Saudi Arabia has climbed by more than 450%, with an average of 358 each year and 179 deaths. Cervical cancer is the most common and dangerous disease in women. Cervical cancer is more dangerous in developing countries. The primary causes of this disease are precocious sexual activity, smoking, sexual activity with multiple partners, early pregnancy, weak immunity, the use of oral anti-pregnancy pills, and poor menstrual cycles [3, 4]. Abnormal menstruation cycle and mild to moderate pain during sex are the most often reported signs of cervical cancer. Cervical cancers are often diagnosed using cytopathology screening, which involves placing cells on a glass slide and observing them under a microscope[5]. The manual examination is still insufficient, leading even specialists to make inaccurate diagnoses. Artificial intelligence (AI) helps diagnose this disease without any manual effort[6, 7]. as shown in the literature review section many researchers have used the Convolutional Neural Networks (CNNs) Deep Learning models along with Ensemble which is a machine learning technique that combines multiple models to improve model performance [8][9]. The Vision Transformer (ViT)has demonstrated the capacity of Transformer-based models as backbone networks in image synthesis and classification tasks, surpassing the performance of traditional pure CNN models [10] [11]. This study utilizes the capabilities of the ensemble learning algorithm and vision transformer encoder to evaluate pap smear images. The next few lines outline the primary contributions of the study:

- A new hybrid transformer model is introduced for the prediction of cervical cancer using pap-smear images.
- Ensemble concatenated transfer learning is implemented using a Self-Attention transformer to decrease the computational complexity and prioritize the most important features.
- A comprehensive empirical study is carried out on five sets of cancer cells utilizing cutting-edge machine learning (ML) and deep learning (DL) technologies.

The paper is divided into several sections. Section 3 provides a Background of the models , Section 3 provides a summary of related works, while Section 4 introduces the

E-mail address: maysoonalz@outlook.sa, ukhan@kau.edu.sa, saalgarni@kau.edu.sa

research techniques and materials used. The experimental results are presented in Section 5, Finally, the conclusion findings are discussed in Section 6.

2. Background

A. Transfer learning and Convolutional Neural Networks

Convolutional Neural Networks (CNNs) and transfer learning are closely related concepts in deep learning. When it comes to processing and analyzing visual data like images, CNNs are a powerful type of neural network architecture. They are composed of multiple layers of convolutional and pooling operations, which enable them to learn hierarchical representations and extract significant features from raw input data automatically[12].

On the other hand, transfer learning refers to the technique of utilizing the knowledge gained from one task to enhance the performance of another related task. This strategy involves the use of pre-trained models trained on large datasets to extract valuable features and representations that can be applied to new, similar tasks.

The relationship between CNNs and transfer learning lies in the fact that CNNs are often used as the base architecture for transfer learning. Pre-trained CNN models, such as those trained on large-scale image datasets like ImageNet, have learned to recognize various visual patterns and features. These models have already captured and encoded valuable information about the underlying structure of images, making them highly effective feature extractors [13].

The advantages of using transfer learning with CNNs include leveraging pre-existing knowledge, reduced need for large amounts of labeled data, faster training times, and improved generalization capabilities. By using the learned features extracted from pre-trained CNN models, transfer learning empowers the model to perform better on new, unseen data and achieve good results even with limited training samples, thereby improving its generalization capabilities. [14].

In summary, transfer learning and CNNs are closely intertwined concepts in deep learning. CNNs serve as the base architecture for transfer learning, allowing the model to leverage pre-trained knowledge and representations to improve performance on related tasks. Combining CNNs and transfer learning offers advantages such as efficient use of computational resources, reduced need for labeled data, improved generalization, and applicability to domains with limited data [15].

B. Ensemble Learning

Ensemble learning involves using a method in machine learning where multiple models are combined to boost the performance of the model[8][9]. By utilizing the perspectives and collaborative knowledge of each model ensemble learning surpasses the constraints and prejudices of individual models resulting in enhanced overall effectiveness and resilience [16].Techniques, like averaging or voting are commonly used in methods to enhance classification tasks[17].The advantages of learning include increased accuracy and resilience[18][19].

The ensemble aggregates the predictions of the constituent models, often through voting or averaging, to arrive at a final prediction [20]. This approach has been successfully applied in various domains, including image classification [16], disease prediction [21][22], and control of complex systems [23]. By combining the strengths of multiple models, ensemble learning enhances prediction accuracy and generalization capabilities, making it a valuable technique in machine learning and data analysis.

C. The Transformer Model and Self-Attention Mechanism

The Transformer model, introduced by [24], has revolutionized natural language processing and various other domains. At the core of the Transformer model lies the self-attention mechanism, which enables the model to capture dependencies between different positions in a sequence. This mechanism has been widely studied and applied in diverse fields.

The self-attention mechanism allows each position in the input sequence to attend to all other positions, capturing the importance of different positions for understanding the context [24]. It computes attention scores between pairs of positions based on the learned query, key, and value vectors [25]. These attention scores determine the relevance of each position to others, enabling the model to focus on different parts of the sequence during processing [25].

The self-attention mechanism has shown remarkable performance in various tasks. In natural language processing, it has been applied to machine translation [24], text summarization [26], and sentiment analysis [25]. It has also been utilized in computer vision tasks such as image recognition [27], object detection [27], and gaze estimation [27]. Additionally, it has been employed in fields like molecular property prediction [28], speech synthesis, and medical image analysis [29].

Recently, the Transformer model and its self-attention mechanism have gained significant attention in computer vision and natural language processing tasks. The Vision Transformer (ViT), introduced by[30], has demonstrated the capacity of Transformer-based models as backbone networks in image synthesis and classification tasks, surpassing the performance of traditional pure CNN models [31] [32].

Transformer models have gained significant popularity in both natural language processing (NLP) and computer vision tasks.[33]. Starting from the end of 2020, significant advancements have been made in the application of transformers in various domains. Interestingly, some recent studies have demonstrated that transformer-based research has surpassed the performance of CNN-based research in specific areas such as image classification, image detection,

and image segmentation[34]. The Transformer algorithm is applied to the discipline of computer vision using the ViT model. The effect of ViT is superior than CNN when it is trained with a substantial quantity of data[33]. The Transformer models have gained significant popularity in both natural language processing (NLP) and computer vision tasks.[33].

3. RELATED WORK

A number of research papers have been written on the topic of cervical cancer classification using deep learning techniques. One study conducted by [35] utilized transfer learning with pre-trained deep learning models such as InceptionResNetV2, VGG19, DenseNet201, and Xception to enhance the accuracy of cervical cancer diagnosis. The study focused on the SIPaKMeD dataset for their analysis.

Another research paper proposed by [36] suggests using an IDT framework that combines the strengths of deep learning and decision tree-based methods. This framework allowed Convolutional Neural Networks (CNNs) to understand novel categories better while maintaining high accuracy for previously learned categories. The framework was evaluated on different datasets, including MNIST, BreakHis, LBC, and SIPaKMeD, and demonstrated promising results with high accuracy rates ranging from 87% to 98%.

In another study[37], researchers employed an ensemble strategy that utilized three pre-trained CNN architectures (Inception v3, Xception, and DenseNet-169) and considered the base classifiers' confidence when making final predictions. The model proposed in this study demonstrated exceptional performance on both the Mendeley Liquid-Based Cytology (LBC) and SIPaKMeD datasets, achieving accuracy rates of 98.55% and 99.23% respectively.Moreover, the literature [38] uses the SIPaKMeD dataset and the adoption of ViT due to its competitive performance and low inductive bias compared to CNNs. However, training large ViT networks can be computationally intensive. As an alternative, a transfer learning-based strategy is introduced, which leverages pre-trained CNN features and categorizes them using a resource-efficient LSTM network. The CNN-LSTM approach achieves a classification accuracy of 95.80% for 5-class classification.

Finally,[39]is a recent research paper utilized state-of-the-art such as Vision Transformers (ViTs) and, Convolutional Neural Networks (CNNs) are employed in this study to classify cervical cancer. The study also utilized data augmentation techniques to enhance the dataset's diversity and ensemble learning techniques to improve model accuracy. The researchers found that the latest ViT-based models outperformed the existing CNN models, offering promising results that could be implemented in clinical settings and ultimately lead to more accurate and timely cancer detection. The study [40]utilized multiple individual models, namely GoogleNet, InceptionResNetV2, VGG16, ResNet50, and Xception. They optimized and froze specific layers as unmodifiable layers. They subsequently combined the high-level characteristics to function as the backend of the convolutional network alongside the self-attention ViT. The model achieved an impressive accuracy of 97.87% when applied to INbreast mammograms.

Cancer class	Number of images	Cells per class
Intermediate,	126	813
Parabasal	108	787
Koilocytotic	238	825
Metaplastic	271	793
Dysketarotic	223	813
Total	**966**	**4049**

TABLE I. Types of cancerous cells in SIPaKMeD dataset

4. MATERIALS AND METHOD

The proposed model for cervical cancer classification is shown in Figure 1, including the ensemble and transformer encoder models, aim to obtain an accurate and effective diagnosis of cervical cancer malignancy. We utilize medical data collected from the SIPaKMeD. The main stages of the proposed models, involve collecting medical benchmark data, preprocessing it, separating the data, applying augmentations, and developing transfer learning models with fine-tuning. The models are then validated and evaluated, and the high-level features are fused. Finally, the models are evaluated again. Preprocessing is necessary to eliminate extraneous information and isolate the potential lesion region of interest (ROIs) or patches. The study uses ensemble transfer learning methodologies and the ViT model, which incorporates the self-attention mechanism.

1) SIPaKMeD Dataset

we utilized the SIPaKMeD dataset, which was produced by researchers associated with the University of Ioannina in Greece in 2018 [41]. This dataset contains 4049 images of women's cells that have been extracted from Pap smear slides.The dataset contains five categories for labeling cancer cells, such as Sup Intermediate, Parabasal, Koilocytotic, Dysketarotic and Metaplastic. To enhance the clarity and quality of images a preprocessing step was carried out on the medical data. This involved adjusting intensities resizing images and eliminating unnecessary information to prepare them for machine learning algorithms [42]. The dataset was divided randomly into training, testing and validation sets. 70% of the images were allocated to the training set while 15% each went to the testing and validation sets. A special summary of the information acquired for this cervical cancerous dataset is furnished in Table I.

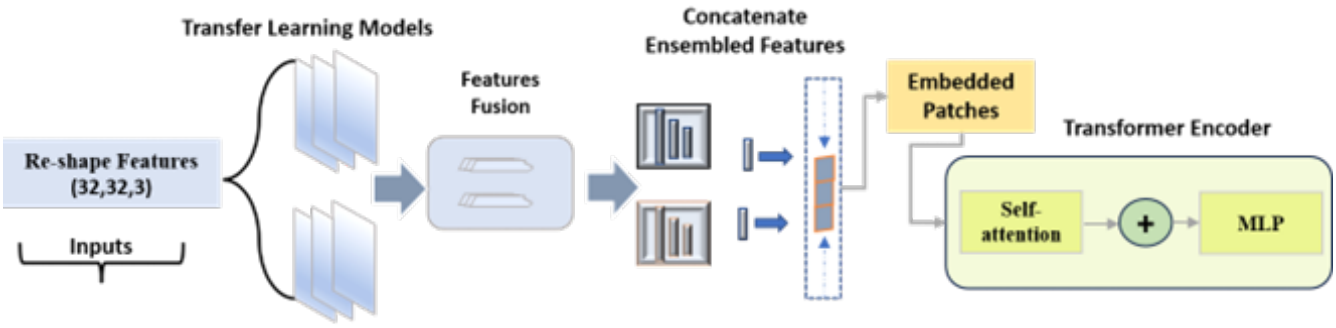

Figure 1. The proposed ensemble and transformer encoder-based models for cervical cancer classification

A. Preprocessing

Medical data preprocessing is essential for adequately preparing trainable medical images. It involves removing unwanted or irrelevant information for AI classifiers, enhancing the spatial resolution and quality of the images, and normalizing and resizing the pixel intensities to ensure they fall within a consistent grayscale range for all images [42] . Moreover, skilled radiologists carefully establish the precise boundaries of the cancerous cell in each pap image, exclusively focusing on these specific areas rather than utilizing the entire image, for the purpose of training AI models.

The patch images are retrieved by isolating only the regions of interest (ROIs) containing cervical malignant cells while ignoring the remaining information in the image. The retrieved image patches are scaled to dimensions of 256×256 pixels.

B. Data splitting, Training, and testing

The data splitting for both datasets includes assigning 70% of SIPaKMeD Pap- smear images from each class to the training set, 15% to the testing set, and 15% to the validation set in random allocation. . The distribution of the SIPaKMeD dataset for five classes is depicted in Table II.

Class	Records	Training	Validation	Testing
Intermediate	831	578	128	125
Parabasal	787	547	121	119
Koilocytotic	825	574	127	124
Metaplastic	813	566	125	122
Dysketarotic	793	552	122	119
Total	4049	2817	623	609

TABLE II. Data splitting for training, validating, and testing, for SIPaKMeD dataset.

C. Augmentation of data training

The data augmentation task is executed with the machine learning imgaug library, which offers a wide range of augmentation approaches. The freshly generated images are stored together with the training images, resulting in a six-fold increase in the quantity of the training data. This leads to improved outcomes. Augmentation is achieved by applying Affine transformations to an image, which include rotation, scaling, translation, shearing, and horizontal and vertical flip operations [43] . CLAHE, or contrast-limited adaptive histogram equalization, improves image contrast by partitioning images into smaller blocks and applying histogram equalization to each block. The image involves modifications in contrast and brightness to accomplish edge identification, Canny edge detection, photometric transformations (Pms), and contrast adaptation.

D. AI MODELS

1) InceptionResNetV2

While InceptionResNetV1 serves as a model for the network design, InceptionV4 forms the basis of the stem [44] . A quick link is located on the far left of every module. It combines inception architecture with residual connections to improve classification outcomes. For the inception modules convolutional operation to work, both the input and the output must be similar to the inception convolutional operation.

2) VGG16

VGG16 is a convolutional neural network (CNN) that is widely regarded as one of the top-performing models for computer vision tasks. The developers of this model assessed the networks and enhanced the depth by employing an architecture which includes extremely small (3×3) convolution filters. This modification resulted in a notable enhancement compared to the previous configurations used in the field. The depth was increased to 16–19 weight layers, resulting in roughly 138 trainable parameters. It can accurately classify 1000 photos over 1000 distinct categories for image classification tasks. Additionally, it offers the advantage of being user-friendly and compatible with transfer learning techniques [45] .

3) ResNet50

ResNet50 is a convolutional neural network that consists of 50 layers and is specifically optimized for tasks related to image recognition. The model underwent training using the ImageNet dataset in order to classify over one thousand

categories, similar to the pre-trained VGG16 model. Moreover, the classification the classification layers have been removed. The process of adding new layers for multiple classifications to the model after freezing. Optimized setups are implemented for the ResNet50 model by freezing 123 layers [46] .

4) ResNet50-V2

The ResNet50-v2 model utilizes pre-activation of weight layers rather than post- activation. The most recent update eliminated the remaining non-linearity and implemented Batch Normalization and ReLU activation on the input before multiplying it with the weight matrix (convolution operation).

5) Xception

Xception is a neural network design comprising layers of convolution organized in a straightforward manner, with each layer distinguished by its depth. It features 36 layers grouped into 14 modules with residual connections forming the core structure for feature extraction, around each module.

6) Ensemble learning

Ensemble learning is an explored area of study that brings together data fusion, data modeling and data mining in a cohesive framework. The central idea behind learning is to gather a diverse range of features through multiple transformations. These features are then used by learning algorithms to make suboptimal predictions. By combining insights from these predictions ensemble learning fosters knowledge discovery and enhances forecasting accuracy through adaptive voting methods[47]. In our research we incorporate the learning features of ResNetV and VGG16 models, for classifying cervical cancer. We remove the classification layer from each model and leverage the final convolution layer to capture its distinctive attributes.

7) Transformer Encoder Model

The primary utility of the vision transformer lies in its ability to effectively recognize objects by leveraging carefully extracted and significant observable features. Furthermore, the self-attention mechanism is employed to achieve remarkable performance while minimizing the reliance on vision-specific assumptions. Unlike traditional CNN models, transformers utilize self-attention to assign varying weights to evaluate the importance of each input data point in an encoder-decoder configuration. In contrast, CNN models primarily analyze the correlations among neighboring pixels within the receptive area defined by the filter size. Therefore, these models are unable to process pixels that are far away. Consequently, in order to tackle this problem, a novel approach was adopted that relied on the attention process. The attention strategy relies on identifying and analyzing the most significant components of the images while disregarding redundant elements [48] . This approach helps to minimize false negative outcomes. The study utilized and refined the vision transformer through the use of an encoder as calculated by equation:

$$Attention(Q, KV), = Softmax(\frac{Qk^t}{dk})v \quad (1)$$

E. Evaluation Strategy

The performance of this approach is evaluated using common metrics such as accuracy, precision, recall, F1-score, and AUC and compared to other existing cervical cancer detection methods.

1) Accuracy(ACC) $=\frac{(TP+TN)}{(TP+TN+FP+FN)}$
2) Specificity(SPE)$=\frac{TN}{TN+FP}$
3) Sensitivity(SEN)$=\frac{TP}{TP+FN}$
4) F1- score$=\frac{2*(precision\times Sensitivity)}{(precision+Sensitivity)}$

5. RESULTS

1) Individuals Transfer Learning Models

The SIPaKMeD database involved the implementation and comparison of five different pre-trained transfer learning models: InceptionResNetV2, ResNet50, ResNet50-V2, VGG16, and Xception. All layers of the InceptionResNetV2 model were frozen and unable to be trained. However, in the Xception, VGG16, ResNet50, and ResNet50-v2 models, we expect that specific layers will be trainable. Specifically, for the Xception model, we expect training to start at layer 106, and for the VGG16 model, training will begin from layer 17. The training process for ResNet50 and ResNet50-v2 begins at the 143rd layer and continues until the classification layers. The ultimate outcomes include the individual transfer learning models, the fine-tuning of models, and the addition of new hidden layers with varying units. In addition, the Adam optimizer was employed, with a learning rate of 0.001 and a total of 50 epochs for all models. The evaluation predictions for each transfer models on this dataset are displayed in Table III . The VGG16 model demonstrates superior classification performance, as shown

by its outstanding results across all evaluation measures. Specifically, it achieves an overall accuracy of 97.04% and an F1-score of 97.06%. ResNet50 achieves superior prediction performance, with an overall accuracy and F1-score of 96.05% and 96.06%, respectively. Conversely, the InceptionResNetV2 demonstrated a third-order accuracy performance of 94.75% and an F1-score of 94.73%. The Xception model showed the lowest accuracy rate of 83.25% when compared to other transfer models.

According to the confusion matrix evaluation, the VGG16 model demonstrated the highest prediction accuracy, with only 18 cases misclassified out of a total of 609 occurrences across all 5 classes. Out of the entire test set including all classes, the ResNet50 model made incorrect classifications for 28 occurrences. The Xception model showed a higher number of misclassified cases, specifically 94 cases across all classes, with a special focus on the koilocytotic class, which accounted for 41 cases. Figure 2 displays the confusion matrices of each individual transfer learning model conducted in the study.

AI Model	ACC	SEN	SPE	F1-Score	AUC
InceptionResNetV2	0.9475	0.9475	0.9475	0.9473	0.967
ResNet50	0.9606	0.9606	0.9609	0.9605	0.975
ResNet50-V2	0.9163	0.9163	0.9182	0.9160	0.948
VGG16	0.9704	0.9704	0.9710	0.9705	0.982
Xception	0.8325	0.8325	0.8503	0.8335	0.895

TABLE III. Performance evaluation of individual transfer learning models, for the SIPaKMeD dataset.

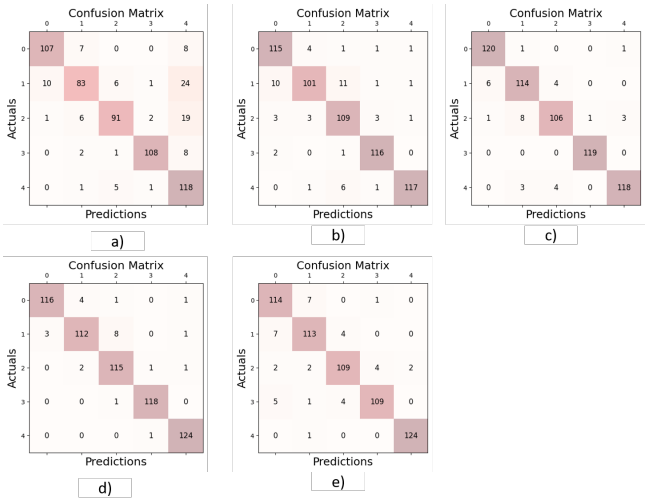

Figure 2. The proposed ensemble and transformer encoder-based models for cervical cancer

2) Ensemble learning vs transformer encoder model

Utilizing the outcomes of the individual transfer models, we combined the vgg16 and resnet50 models to construct the ensemble backbone network for classification. Within this area, we conduct performance assessments through a two-step process. Initially, the ensemble learning model is developed, trained and evaluated separately without involving the ViT. Then we create the encoder model by utilizing the ViT to make the final prediction using the combined high level deep features generated by merging two unique CNN models. The evaluation performance findings of cervical cancer, obtained by the utilization of ensemble learning and the transformer encoder, are succinctly presented in Table IV. The experimental proof shows the transformer encoder model yields the most optimal assessment outcomes when employed in conjunction with the ensemble learning model and others models with 97.54% accuracy and F1-score. Figure 3 illustrates the confusion matrices produced by the ensemble learning and transformer encoder models on the SIPaKMeD dataset. While the Parabasal class remained error-free, the rest of the classes made fifteen and sixteen wrong predictions, respectively, for the ensemble and transformer encoder models.

AI Model	ACC	SEN	SPE	F1-Score	AUC
Ensemble Learning Model	0.9737	0.9737	0.9738	0.9737	0.984
Transformer Encoder Model	0.9754	0.9754	0.9756	0.9754	0.985

TABLE IV. Performance evaluation of ensemble CNN fusion and Transformer encoder models, for the SIPaKMeD dataset.

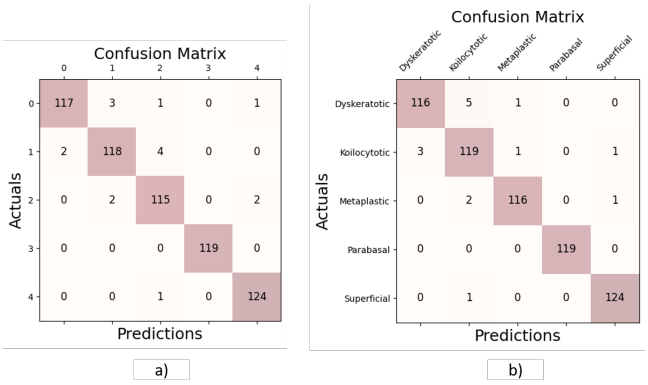

Figure 3. The confusion matrices of ensemble CNN fusion and Transformer encoder models: (a) Ensemble learning model, and (b) Transformer encoder model using the SIPaKMeD dataset

Figure 4 illustrates the comparative performance results of the prediction error for each individual transfer learning model in contrast to the proposed ensemble and transformer encoder models. The VGG16 model has a prediction error of 0.029, which is the closest to the proposed models that have prediction errors of 0.026 and 0.024, respectively. Conversely, the Xception model has the largest degree of error variation in terms of accuracy, with a prediction error of 0.16.

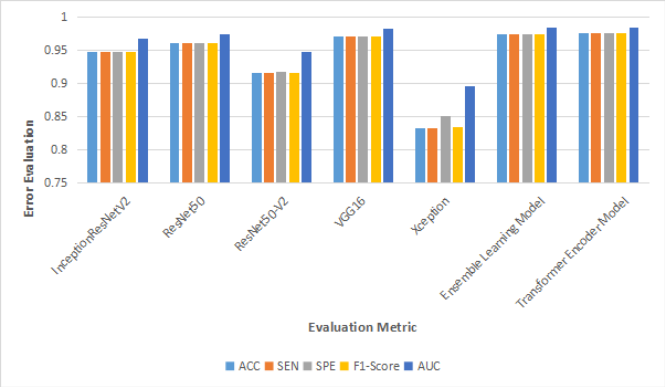

Figure 4. The performance error evaluation for individual transfer learning models with the proposed ensemble and transformer encoder models on SIPaKMeD dataset.

6. Conclusions and Future Work

The study presents an approach that use deep learning to classify cervical cells. The pre-trained transfer learning models, including InceptionResNetV2, ResNet50, ResNet50-V2, VGG16, and Xception, are utilized with fine-tuned and frozen layers. The high-level characteristics of these models are then fused to create an ensemble model, which is subsequently input into the proposed transformer encoder ViT model. The SIPaKMeD were utilized for models training with 5 cervical cancer classes for model training. Based on the performance metrics, it is evident that the suggested transformer encoder ViT model outperforms both the ensemble learning and pre-trained models in terms of classification accuracy. The VGG16 model achieved the highest accuracy of 97.04% and an F1-score of 97.06% for 5-class classification problems in the SIPaKMeD dataset. Nevertheless, the ensemble learning model attained an accuracy of 97.37%, while the transformer encoder model surpassed all other models with an accuracy of 97.54%.And to compare the results obtained here with those of other methods found in the literature, Table V shows that the proposed method achieved the highest accuracy for 5 classes compared with the literature.

References

[1] M. Arbyn, E. Weiderpass, L. Bruni, S. de Sanjosé, M. Saraiya, J. Ferlay, and F. Bray, "Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis," *The Lancet Global Health*, vol. 8, no. 2, pp. e191–e203, 2020.

[2] R. O. Alkhaldi, H. A. Alzahrani, L. A. Metwally, R. Alkhaldi, and H. Alzahrani, "Awareness level about cervical cancer, human papillomavirus (hpv) and corresponding vaccine among women living in the western region of saudi arabia," *Cureus*, vol. 15, no. 4, 2023.

[3] T. Šarenac and M. Mikov, "Cervical cancer, different treatments and importance of bile acids as therapeutic agents in this disease," *Frontiers in Pharmacology*, vol. 10, p. 484, 2019.

[4] W. Liu, C. Li, N. Xu, T. Jiang, M. M. Rahaman, H. Sun, X. Wu, W. Hu, H. Chen, C. Sun *et al.*, "Cvm-cervix: A hybrid cervical pap-smear image classification framework using cnn, visual transformer and multilayer perceptron," *Pattern Recognition*, vol. 130, p. 108829, 2022.

[5] A. Gençtav, S. Aksoy, and S. Önder, "Unsupervised segmentation and classification of cervical cell images," *Pattern recognition*, vol. 45, no. 12, pp. 4151–4168, 2012.

[6] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[7] R. C. Maron, M. Weichenthal, J. S. Utikal, A. Hekler, C. Berking, A. Hauschild, A. H. Enk, S. Haferkamp, J. Klode, D. Schadendorf *et al.*, "Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks," *European Journal of Cancer*, vol. 119, pp. 57–65, 2019.

[8] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996.

[9] P. Bühlmann and T. Hothorn, "Boosting algorithms: regularization, prediction and model fitting," *Statistical Science*, vol. 22, 2007.

[10] T. Li, F. Zhang, G. Xie, X. Fan, Y. Gao, and M. Sun, "A high speed reconfigurable architecture for softmax and gelu in vision transformer," *Electronics Letters*, vol. 59, no. 5, p. e12751, 2023.

[11] T. Mustaqim, C. Fatichah, and N. Suciati, "Deep learning for the detection of acute lymphoblastic leukemia subtypes on microscopic images: A systematic literature review," *IEEE Access*, 2023.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[13] J. Li, W. Wu, D. Xue, and P. Gao, "Multi-source deep transfer neural network algorithm," *Sensors*, vol. 19,

Refrence	Accurecy %	classes	Dataset/s	Approach
[35]	96.63	5	SIPaKMeD	DenseNet201
[36]	93.00	5	SIPaKMeD	IDT framework
[37]	95.43	5	SIPaKMeD	ensemble
[38]	95.80	5	SIPaKMeD	ViT-CNN
[39]	92.04	2	SIPaKMeD	Rd
proposed model	97.37	5	SIPaKMeD	Ensemble Learning Model
proposed model	97.54	5	SIPaKMeD	Transformer Encoder Model

TABLE V. Comparison with previous work

p. 3992, 2019.

[14] L. Sun, X. Fan, S. Huang, S. Luo, L. Zhao, X. Chen, Y. He, and X. Suo, "Research on classification method of eggplant seeds based on machine learning and multispectral imaging classification eggplant seeds," *Journal of Sensors*, vol. 2021, pp. 1–9, 2021.

[15] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," *Sensors*, vol. 17, p. 425, 2017.

[16] C. Ju, A. F. Bibaut, and M. J. v. d. Laan, "The relative performance of ensemble methods with deep convolutional neural networks for image classification," *Journal of Applied Statistics*, vol. 45, pp. 2800–2818, 2018.

[17] X. Gao, A. A. Ali, H. S. Hassan, and E. M. Anwar, "Improving the accuracy for analyzing heart diseases prediction based on the ensemble method," *Complexity*, vol. 2021, pp. 1–10, 2021.

[18] S. Lee, M. Kim, S. Shin, S. Baek, S. Park, and Y. Jeong, "Ensemble-guided model for performance enhancement in model-complexity-limited acoustic scene classification," *Applied Sciences*, vol. 12, p. 44, 2021.

[19] A. B. Majumder, S. Gupta, and D. Singh, "An ensemble heart disease prediction model bagged with logistic regression, naïve bayes and k nearest neighbour." *Journal of Physics: Conference Series*, vol. 2286, p. 012017, 2022.

[20] L. Li, A. J. Blomberg, R. A. Stern, C. Kang, S. Papatheodorou, Y. Wei, M. Liu, A. Peralta, C. L. Vieira, and P. Koutrakis, "Predicting monthly community-level domestic radon concentrations in the greater boston area with an ensemble learning model," *Environmental Science Amp; Technology*, vol. 55, pp. 7157–7166, 2021.

[21] S. Rajaraman, S. Sornapudi, P. O. Alderson, and L. Folio, "Analyzing inter-reader variability affecting deep ensemble learning for covid-19 detection in chest radiographs," *Plos One*, vol. 15, p. e0242301, 2020.

[22] C. Wu, M. Hwang, T. Huang, Y. J. Chen, Y. Chang, T. Ho, J. Huang, K. Hwang, and W. Ho, "Application of artificial intelligence ensemble learning model in early prediction of atrial fibrillation," *BMC Bioinformatics*, vol. 22, 2021.

[23] Z. Wu, A. Tran, D. Rincón, and P. D. Christofides, "Machine learning-based predictive control of nonlinear processes. part i: theory," *AIChE Journal*, vol. 65, 2019.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, F. Morgan, and J. Brew, "Transformers: state-of-the-art natural language processing," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.

[26] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," 2019.

[27] J. Chen, J. Ma, X. Wang, L. Huang, and Y. Li, "Gaze estimation based on swin transformer," *Sixth International Conference on Advanced Electronic Materials, Computers, and Software Engineering (AEMCSE 2023)*, 2023.

[28] C. Liu, Y. Sun, R. L. Davis, S. T. Cardona, and P. Hu, "Abt-mpnn: an atom-bond transformer-based message-passing neural network for molecular property prediction," *Journal of Cheminformatics*, vol. 15, 2023.

[29] R. Ghosh and F. Bovolo, "An fft-based cnn-transformer encoder for semantic segmentation of radar sounder signal," *Image and Signal Processing for Remote Sensing XXVIII*, 2022.

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: transformers for image recognition at scale," 2020.

[31] T. Li, F. Zhang, G. Xie, X. Fan, Y. Gao, and M. Sun, "A high speed reconfigurable architecture for softmax

and gelu in vision transformer," *Electronics Letters*, vol. 59, 2023.
[32] T. Mustaqim, C. Fatichah, and N. Suciati, "Deep learning for the detection of acute lymphoblastic leukemia subtypes on microscopic images: a systematic literature review," *IEEE Access*, vol. 11, pp. 16 108–16 127, 2023.
[33] 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.
[34] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.
[35] 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.
[36] W. Mousser, S. Ouadfel, A. Taleb-Ahmed, and I. Kitouni, "Idt: An incremental deep tree framework for biological image classification," *Artificial Intelligence in Medicine*, vol. 134, p. 102392, 2022.
[37] A. Manna, R. Kundu, D. Kaplun, A. Sinitca, and R. Sarkar, "A fuzzy rank-based ensemble of cnn models for classification of cervical cytology," *Scientific Reports*, vol. 11, no. 1, p. 14538, 2021.
[38] R. Maurya, N. N. Pandey, and M. K. Dutta, "Vision-cervix: Papanicolaou cervical smears classification using novel cnn-vision ensemble approach," *Biomedical Signal Processing and Control*, vol. 79, p. 104156, 2023.
[39] F. B. Akyol and O. ALTUN, "Detection of cervix cancer from pap-smear images," *Sakarya University Journal of Computer and Information Sciences*, vol. 3, no. 2, pp. 99–111, 2020.
[40] A. M. Al-Hejri, R. M. Al-Tam, M. Fazea, A. H. Sable, S. Lee, and M. A. Al-Antari, "Etecadx: Ensemble self-attention transformer encoder for breast cancer diagnosis using full-field digital x-ray breast images," *Diagnostics*, vol. 13, no. 1, p. 89, 2022.
[41] 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.
[42] F. Pérez-García, R. Sparks, and S. Ourselin, "Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Computer Methods and Programs in Biomedicine*, vol. 208, p. 106236, 2021.
[43] M. M. Rahaman, C. Li, Y. Yao, F. Kulwa, X. Wu, X. Li, and Q. Wang, "Deepcervix: A deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques," *Computers in Biology and Medicine*, vol. 136, p. 104649, 2021.
[44] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.
[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[46] R. M. Al-Tam, A. M. Al-Hejri, S. M. Narangale, N. A. Samee, N. F. Mahmoud, M. A. Al-Masni, and M. A. Al-Antari, "A hybrid workflow of residual convolutional transformer encoder for breast cancer classification using digital x-ray mammograms," *Biomedicines*, vol. 10, no. 11, p. 2971, 2022.
[47] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science*, vol. 14, pp. 241–258, 2020.
[48] C. C. Ukwuoma, Z. Qin, M. B. B. Heyat, F. Akhtar, O. Bamisile, A. Y. Muaad, D. Addo, and M. A. Al-Antari, "A hybrid explainable ensemble transformer encoder for pneumonia identification from chest x-ray images," *Journal of Advanced Research*, vol. 48, pp. 191–211, 2023.

Maysoon Mohammed Alzahrani Maysoon Alzahrni is a Master's student in Computer Information Systems at King Abdulaziz University. Her research interests lie in AI and Machine Learning with Python Programming, Data Science, Data Mining.

Usman Ali Khan Usman Ali Khan is an Associate Professor in the Information Systems Department. He obtained his Ph.D. in Information Technology with a specialization in Software Engineering from Integral University in India in 2007. His research interests include AI and Machine Learning with Python Programming, Data Science, Data Mining, and Software Engineering-Project Management

Sultan Abdullah Al-Garni Sultan Algarni is an Assistant Professor in the Information Systems Department. He holds a PhD in Computer Science with a specialization in Information Security from King Abdulaziz University (KAU) in Saudi Arabia, which he obtained in 2022. His research interests span several areas, including Information Security, Cybersecurity, Networking, Software Defined Network (SDN), Internet of Things (IoT) systems, Blockchain Technology, and Artificial Intelligence