International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 03, ICCT - 2026

AN INTERPRETABLE ATTENTION-BASED INCEPTION-TRANSFORMER FRAMEWORK FOR CERVICAL CANCER CELL CLASSIFICATION USING HERLEV DATASET

Ananthakumari A
Assistant Professor, CSE
Dr. G. U pope College of Engineering,
Sawyerpuram.
kumari.cse2007@gmail.com

Josephine Monisha R
II ME (CSE)
Dr. G. U pope College of Engineering,
Sawyerpuram.
josephinemonisha@gmail.com

***Abstract* - Cervical cancer detection requires accurate automated systems. This work proposes a deep learning framework for cervical cell classification using the Herlev dataset containing 9,650 images across five classes. Images are preprocessed through resizing and normalization, along with data augmentation techniques such as rotation, zoom, and flipping. The model employs InceptionV3 for feature extraction with partial fine-tuning, followed by a Multi-Head Self-Attention module to enhance feature representation. Global Average Pooling and a multilayer perceptron with dropout are used for classification. A 5-fold cross-validation strategy is applied with an 80:20 train-test split. The model is trained using the Adam optimizer with a learning rate of 0.0001 and categorical cross-entropy loss, along with Early Stopping and learning rate scheduling. The proposed model achieves an average accuracy of 98.69% with a standard deviation of 0.35%. Performance metrics show high precision, recall, and F1-score, while Grad-CAM improves interpretability by highlighting important regions.**

Keywords* - *Cervical Cancer Classification, Herlev Dataset, InceptionV3, Attention Mechanism, Deep Learning, Medical Image Analysis, Grad-CAM, Computer-Aided Diagnosis

I. INTRODUCTION

Cervical cancer is one of the leading causes of cancer-related deaths among women worldwide, and early detection plays a crucial role in reducing mortality. Conventional diagnostic methods such as Pap smear testing rely on manual examination of cervical cell images, which is time-consuming and highly dependent on expert interpretation. This often leads to variability in diagnosis and delays in treatment.

With the advancement of deep learning, automated image classification systems have shown significant potential in medical diagnosis. Convolutional Neural Networks (CNNs) have been widely used for extracting spatial features from medical images. However, traditional CNN-based models are limited in capturing global contextual relationships and often lack interpretability.

To overcome these limitations, this work proposes an attention-enhanced deep learning framework for cervical cell classification using the Herlev dataset. The model combines InceptionV3 for feature extraction with a Multi-Head Self-Attention mechanism to improve feature representation. Additionally, Grad-CAM is used to visualize important regions influencing predictions, enhancing model transparency.

The proposed model is evaluated using a 5-fold cross-validation strategy to ensure robustness. Experimental results demonstrate high accuracy and reliable classification performance across all classes.

The main contributions of this work are summarized as follows:

- Development of a hybrid CNN and attention-based model for cervical cell classification
- Integration of Multi-Head Self-Attention to capture global dependencies
- Application of Grad-CAM for interpretability
- Achievement of high accuracy using robust validation techniques

II. RELATED WORK

Automated cervical cancer detection has gained significant attention with the advancement of machine learning and deep learning techniques. Early approaches relied on traditional machine learning algorithms using handcrafted features such as texture, shape, and color. These methods typically employed classifiers like Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN). However, their performance was limited due to the dependency on manually extracted features and lack of robustness.

With the emergence of deep learning, Convolutional Neural Networks (CNNs) have become the dominant approach for medical image classification. Architectures such as VGGNet, ResNet, and Inception have demonstrated strong performance by automatically learning hierarchical feature representations. Among these, InceptionV3 has been widely adopted due to its efficient architecture and ability to capture multi-scale features.

Transfer learning has further improved performance by utilizing pre-trained models on large datasets such as ImageNet. This approach reduces training time and enhances generalization, especially when dealing with limited medical datasets like the Herlev dataset.

Recent research has focused on improving model performance using attention mechanisms. Transformer-based models and Multi-Head Self-Attention have shown the ability to capture long-range dependencies and contextual relationships within images. These methods

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 03, ICCT - 2026

enhance feature representation beyond what traditional CNNs can achieve.

In addition to accuracy, interpretability has become an important aspect in medical applications. Techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) are used to visualize regions of interest in input images, helping clinicians understand model decisions.

Despite these advancements, challenges remain in balancing high accuracy with interpretability and computational efficiency. This work addresses these challenges by combining InceptionV3 with a Transformer-based attention module and Grad-CAM visualization, providing both high performance and explainability.

III. METHODOLOGY

A. Dataset Description

The dataset used in this study is the Herlev cervical cell dataset, which consists of 9,650 microscopic images categorized into five classes: Dyskeratotic, Koilocytotic, Metaplastic, Parabasal, and Superficial-Intermediate. The dataset is well-balanced, with each class containing approximately 20% of the total samples, ensuring unbiased model training.

B. Data Preprocessing

All images are resized to 224 × 224 pixels to match the input requirements of the InceptionV3 model. Pixel values are normalized to the range [0,1] by dividing by 255 to ensure stable training.

To improve model generalization and reduce overfitting, data augmentation techniques are applied using an ImageDataGenerator. The augmentation includes:

- Rotation up to 30 degrees
- Zoom transformation up to 20%
- Horizontal flipping

These transformations increase dataset diversity and help the model learn robust features.

C. Label Encoding

The categorical class labels are converted into numerical format using label encoding. Subsequently, one-hot encoding is applied to transform labels into a binary class matrix suitable for multi-class classification using softmax activation.

D. Cross-Validation Strategy

A 5-fold cross-validation approach is used to evaluate model performance. The dataset is split into five equal parts, where in each iteration:

- 80% of data is used for training
- 20% is used for validation

This process ensures that every sample is used for both training and validation, improving reliability and reducing bias.

E. Feature Extraction using InceptionV3

A pre-trained InceptionV3 model is used as the backbone for feature extraction. The model is initialized with ImageNet weights and the top classification layers are removed.

To retain learned features while adapting to the new dataset:

- Initial layers are frozen
- Last 50 layers are fine-tuned

This allows the model to learn domain-specific features while maintaining general feature representations.

F. Global Feature Representation

The output feature maps from InceptionV3 are passed through a Global Average Pooling layer. This reduces spatial dimensions and converts the feature maps into a compact feature vector, reducing computational complexity and overfitting.

G. Transformer-Based Attention Module

To enhance feature representation, a Multi-Head Self-Attention mechanism is applied.

The feature vector is reshaped and passed through:

- Multi-Head Attention (4 heads, key dimension = 64)
- Residual connection
- Layer Normalization

This module captures global dependencies and relationships between features, improving classification performance.

H. Classification Layer

The refined feature vector is passed through a Multilayer Perceptron consisting of:

- Dense layer (256 neurons, ReLU activation)
- Dropout (0.2)
- Dense layer (128 neurons, ReLU activation)
- Dropout (0.2)

Finally, a softmax layer is used to classify the input into five categories.

I. Model Training

The model is trained using:

- Optimizer: Adam
- Learning rate: 0.0001
- Loss function: Categorical Crossentropy

To improve training efficiency and prevent overfitting:

- EarlyStopping is used to stop training when validation performance stops improving
- ReduceLROnPlateau is used to decrease the learning rate when validation loss stagnates

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 03, ICCT - 2026

J. Workflow Summary

The overall workflow of the proposed system is as follows:

Image Input → Preprocessing → InceptionV3 Feature Extraction → Global Average Pooling → Attention Module → Fully Connected Layers → Softmax Output

IV. PROPOSED ARCHITECTURE

A. Architectural Overview

The proposed architecture is a hybrid deep learning model that combines Convolutional Neural Networks (CNN) and Transformer-based attention mechanisms for accurate cervical cell classification. The model is designed to extract both local spatial features and global contextual relationships from medical images.

The architecture consists of three major stages:

- Feature Extraction using InceptionV3
- Attention Enhancement using Multi-Head Self-Attention
- Classification using Fully Connected Layers

B. Block Diagram

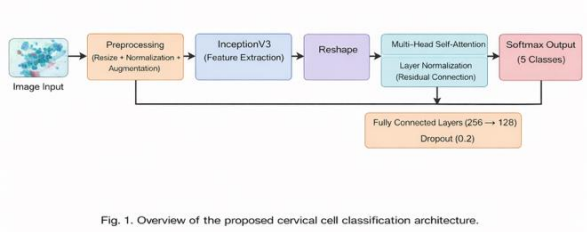


Fig. 1. Overview of the proposed cervical cell classification architecture.

C. Feature Extraction Stage

The input cervical cell image is passed through the InceptionV3 model, which is pre-trained on ImageNet. This stage extracts rich hierarchical features such as edges, textures, and complex patterns relevant to cell structures.

To adapt the model to the medical domain:

- Early layers are frozen to preserve general features
- Deeper layers are fine-tuned to learn dataset-specific patterns

D. Attention Mechanism

The extracted feature vector is enhanced using a Multi-Head Self-Attention module.

Key operations include:

- Reshaping the feature vector into sequence format
- Applying multi-head attention (4 heads)
- Adding residual connection
- Applying layer normalization

This allows the model to:

- Focus on important regions
- Capture global dependencies
- Improve feature relationships

E. Classification Layer

The attention-enhanced features are passed through a Multilayer Perceptron (MLP) consisting of:

- Dense layer with 256 neurons
- Dropout layer (0.2)
- Dense layer with 128 neurons
- Dropout layer (0.2)

Finally, a softmax layer outputs probabilities for the five cervical cell classes.

F. Key Advantages of the Architecture

The proposed architecture offers several advantages:

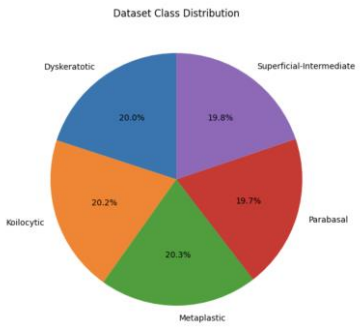
- Combines CNN and Transformer strengths
- Captures both local and global features
- Reduces overfitting using dropout and augmentation
- Improves interpretability through attention and Grad-CAM
- Achieves high accuracy with efficient computation

G. Data Flow Description

The input image undergoes preprocessing before being fed into the InceptionV3 network for feature extraction. The resulting feature maps are converted into a compact vector using Global Average Pooling. This vector is then refined using a Multi-Head Self-Attention mechanism, which enhances important feature relationships. The refined features are passed through fully connected layers for classification, and the final output is generated using a softmax activation function.

V. RESULTS AND PERFORMANCE ANALYSIS

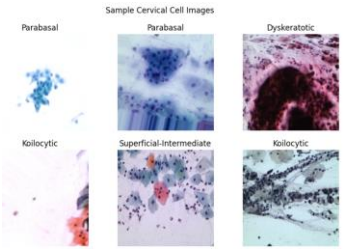
A. Dataset Distribution



The dataset consists of 9,650 cervical cell images distributed across five classes: Dyskeratotic, Koilocytotic, Metaplastic, Parabasal, and Superficial-Intermediate. The distribution is nearly uniform, with each class contributing approximately 20% of the total dataset. This balanced distribution helps in reducing bias during training and ensures fair model evaluation.

B. Sample Input Images

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 03, ICCT - 2026



Sample cervical cell images from different classes are used to visualize the diversity of the dataset. Each class exhibits distinct morphological characteristics, which are learned by the model during training.

C. Model Performance

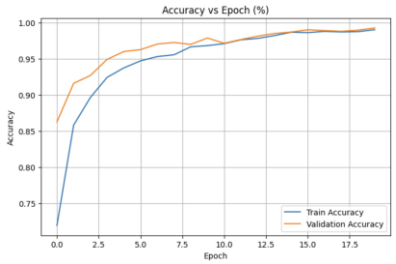
```
========== FINAL MODEL PERFORMANCE ==========
Average Accuracy: 98.69%
Std Deviation: 0.35%
```

The performance of the proposed model is evaluated using 5-fold cross-validation.

- Average Accuracy: 98.69%
- Standard Deviation: 0.35%

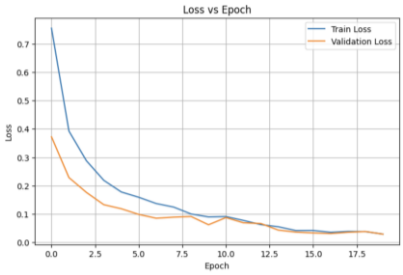
These results indicate that the model achieves high accuracy with consistent performance across different folds, demonstrating strong generalization capability.

D. Accuracy Analysis



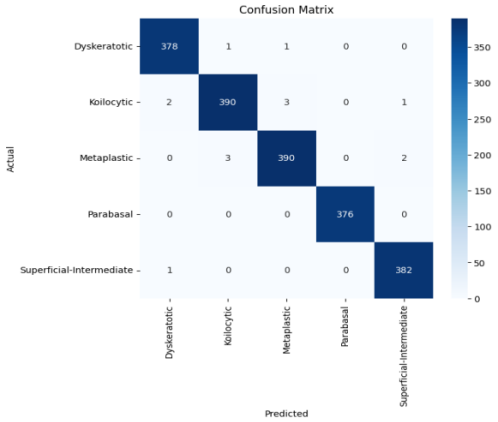
The accuracy curve shows a steady increase in both training and validation accuracy over epochs. The convergence of the two curves indicates minimal overfitting and effective learning.

E. Loss Analysis



The loss curve shows a consistent decrease during training, indicating stable optimization. The validation loss closely follows the training loss, further confirming that the model is not overfitting.

F. Confusion Matrix



The confusion matrix illustrates the classification performance across all classes. Most predictions lie along the diagonal, indicating correct classification. Misclassifications are minimal, showing the effectiveness of the model.

H. Classification Report

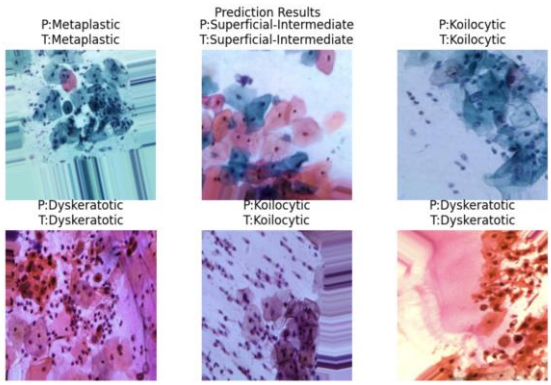
The classification report shows high precision, recall, and F1-score for all classes, with values close to 1.00. This confirms that the model performs consistently across different categories without bias.

```
========== CLASSIFICATION REPORT ==========
                          precision    recall  f1-score   support

             Dyskeratotic       0.99      0.99      0.99       380
              Koilocytic        0.99      0.98      0.99       396
             Metaplastic        0.99      0.99      0.99       395
              Parabasal         1.00      1.00      1.00       376
Superficial-Intermediate        0.99      1.00      0.99       383

                accuracy                            0.99      1930
               macro avg        0.99      0.99      0.99      1930
            weighted avg        0.99      0.99      0.99      1930
```

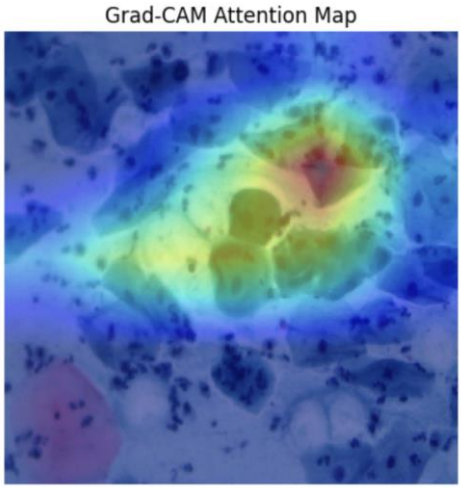
I. Prediction Results



Sample predictions demonstrate that the model accurately classifies cervical cell images. The predicted labels closely match the true labels, indicating strong model reliability.

J. Grad-CAM Visualization

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 03, ICCT - 2026



Grad-CAM visualization highlights the important regions in the input images that influence model predictions. This improves interpretability and helps in understanding how the model makes decisions, which is critical in medical applications.

K. Overall Observation

The proposed model achieves high accuracy with strong generalization and interpretability. The integration of attention mechanisms enhances feature representation, while Grad-CAM provides visual explanations, making the system suitable for real-world medical diagnosis.

VI. DISCUSSION

The experimental results demonstrate that the proposed hybrid architecture effectively improves cervical cell classification performance by combining convolutional feature extraction with attention-based learning. The use of InceptionV3 enables the model to capture rich spatial features, while the Multi-Head Self-Attention module enhances global contextual understanding, leading to more accurate predictions.

The high classification accuracy of 98.69% and low standard deviation of 0.35% indicate that the model generalizes well across different data splits. The balanced dataset further contributes to consistent performance across all classes, as reflected in the confusion matrix and classification report.

One of the key strengths of the proposed model is its ability to maintain both high accuracy and interpretability. The integration of Grad-CAM provides visual explanations by highlighting important regions in the input images, making the model more transparent and reliable for medical applications. This is particularly important in healthcare, where understanding the reasoning behind predictions is essential.

The training and validation curves show stable convergence with minimal overfitting, which confirms the effectiveness of data augmentation, dropout regularization, and learning rate optimization techniques. The use of cross-validation further strengthens the reliability of the results.

However, the study has certain limitations. The model is evaluated only on the Herlev dataset, which may not fully represent real-world clinical variability. Additionally, while the model performs well in controlled conditions, deployment in real-time clinical environments may require further optimization and validation.

Future improvements can focus on evaluating the model on larger and more diverse datasets, optimizing computational efficiency for real-time applications, and integrating the system into clinical workflows for practical usage.

VII. CONCLUSION AND FUTURE WORK

This work presents an attention-enhanced deep learning framework for cervical cell classification using the Herlev dataset. The proposed model integrates InceptionV3 for feature extraction with a Multi-Head Self-Attention mechanism to capture both local and global features. The use of Global Average Pooling and fully connected layers with dropout further improves classification performance and reduces overfitting.

The model is evaluated using a 5-fold cross-validation strategy and achieves an average accuracy of 98.69% with a standard deviation of 0.35%. The results demonstrate strong generalization and consistent performance across all classes. The confusion matrix and classification report confirm high precision, recall, and F1-score values.

In addition to high accuracy, the model provides interpretability through Grad-CAM visualization, which highlights important regions influencing predictions. This enhances the reliability of the system and makes it suitable for medical applications.

Future work will focus on extending the model to larger and more diverse datasets, improving computational efficiency for real-time deployment, and integrating the system into clinical decision-support tools. Further research can also explore advanced attention mechanisms and lightweight architectures for improved performance in resource-constrained environments.

VIII. ACKNOWLEDGMENT

The authors would like to express their gratitude to their institution for providing the necessary resources and support to carry out this research work. The authors also thank their mentors and peers for their valuable guidance and suggestions throughout the project.

REFERENCES

[1] [1] H. Yan, X. Shen, P. Tao, L. Jin, and Y. Zhang, “Deep learning models for cervical cancer subtyping using whole-slide images,” Frontiers in Oncology, vol. 15, 2025.
[2] [2] M. I. H. Siddiqui et al., “Accelerated and accurate cervical cancer diagnosis using ensemble deep learning models,” Biomedical Signal Processing and Control, 2025.
[3] [3] A. M. Al-Hejri et al., “A hybrid vision transformer with ensemble CNN framework for cervical cancer classification,” BMC Medical Informatics and Decision Making, 2025.
[4] [4] G. K. Ameta et al., “Cervical cancer prediction using deformable kernel CNN and transformer-based attention,” Scientific Reports, vol. 15, 2025.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 03, ICCT - 2026

[5] [5] H. Mbelwa et al., “A systematic review on computer vision-based methods for cervical cancer detection,” Informatics in Medicine Unlocked, 2025.
[6] [6] B. Vazquez et al., “Machine and deep learning for the diagnosis and prognosis of cervical cancer: A review,” Diagnostics, vol. 15, no. 12, 2025.
[7] [7] B. Z. Wubineh et al., “Deep learning-based segmentation and classification of cervical cancer images,” 2026.
[8] [8] S. Simaiya et al., “A hybrid deep learning framework for high-precision cervical cancer detection,” IET Image Processing, 2026.