Submitted: 12 July, 2024 Accepted: 13 August, 2024 Published: 15 January, 2025 DOI:10.22514/ejgo.2025.012

ORIGINAL RESEARCH

European Journal of Gynaecological Oncology

Investigating deep learning approaches for cervical cancer diagnosis: a focus on modern image-based models

Ishak Pacal[1,*]

[1]Department of Computer Engineering, Faculty of Engineering, Igdir University, 76000 Igdir, Turkey

***Correspondence**
ishak.pacal@gdir.edu.tr
(Ishak Pacal)

Abstract

Background: Cervical cancer is a leading health concern for women globally, necessitating accurate and timely diagnostic methods. While the Papanicolaou smear (Pap smear) test remains the gold standard for cervical cancer screening, it is time-consuming and prone to human error. This highlights the need for automated diagnostic tools to improve efficiency and accuracy. **Methods**: This study evaluated the performance of deep learning models for automating cervical cancer diagnosis using Pap smear images. A new dataset was constructed by merging the Mendeley Liquid-Based Cytology (LBC) dataset (963 images) and the Malhari dataset (318 images), resulting in 1,281 images. Twenty-seven cutting-edge deep learning models, including Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), were used for classification. Data augmentation and transfer learning techniques were applied to enhance model performance. **Results**: The majority of ViT-based models achieved a high classification accuracy of 99.48%. Among the 13 CNN-based models evaluated, EfficientNetV2-Small was the only model to achieve the same accuracy level. The results demonstrate the superiority of ViT-based models in achieving high diagnostic accuracy. **Conclusions**: Deep learning methods, particularly ViT-based models, show substantial potential in automating cervical cancer diagnosis. These models can enhance diagnostic accuracy, reduce human error, and provide timely results, thereby supporting more efficient and reliable cervical cancer screening practices.

Keywords

Deep learning; CNNs and ViTs; Cervical cancer detection; Artificial intelligence in medicine

1. Introduction

Cervical cancer is considered to be one of the most serious health problems for women in the world, with nearly 570,000 new cases per year, which are the main cause of deaths from cancer [1, 2]. The main reason for this disease is the human papillomavirus (HPV) which causes abnormal tissue in cervical cells [3]. Despite HPV being the major factor, other elements like smoking, Human Immunodeficiency Virus (HIV), contraceptive pill use for a long time, and weak immune system are also significant in the disease's course [4]. Over time, cervical cancer has become the fourth most common cancer in women all around the globe, thereby highlighting the requirements for accurate and early detection mechanisms to improve the treatment results, increase the survival rate of the patients, and lower the healthcare costs [5].

Screening for cervical cancer is the typical route which includes procedures such as the Papanicolaou (Pap) smear test, HPV testing, and liquid-based cytology [6]. In a Pap smear test, the doctor scrapes the cells from the cervix with brushes and then puts them on a glass slide for microscopic examination by a cytopathologist. Every slide in the case includes thousands of cells, and the task is thinking how some cell types become the same looking because of the size and shape of the nuclei and cells [7]. Receiving the correct spiritual diagnose by visually seeing these cells comes from the examination of the cells by the best of their abilities, this sometimes makes them to be in a bit of a rush and they might skip some things on the slide [8]. Hence the number of incorrect diagnoses and delayed treatments may be experienced in some cases.

The testing area might have to handle the increased number of diagnostics that are executed to a great extent, which leads more of the procedure to be faulty and therefore quite contrary to the expectations of the person who is expecting it to be further tested [9]. But, as a matter of fact, the more screening tests are to be done the less probable it is to get things faster and with fewer mistakes on the doctors' side too than the other medical professional's show. Thus, the automation of the diagnostic procedure becomes the far more fundamental issue. Labor-intensive and challenging screening of the same

cells have been identified as the main areas where people are primarily occupied, hence, machine learning techniques have been designed to improve the precision and speed of cytopathologists in the interpretation of the slides [10]. These ways are the mixture of the traditional hand-crafted features together with the artificial intelligence learning algorithms that are capable of detecting cervical cancer on the base of the information received from the histograms constructed from the images of the cells and the cell clusters [11]. A direct result of the automation of the workload of cytology, it has become 20% less time-consuming [12]. Shortly, cervical cancer remains a significant health challenge worldwide, with early detection being crucial for effective treatment and improved patient outcomes. Traditional methods, such as Pap smear tests, have been the gold standard for cervical cancer screening for decades. While effective, these methods have limitations, including variability in test results due to human error, the need for highly skilled cytologists, and the time-consuming nature of manual slide examination [13].

Artificial intelligence, and more specifically deep learning, have made significant breakthroughs that not only overcome the limitations of the practices but also offer more accurate and reliable identification and classification of many types of medical images, cancers as well as cervical cancers [14–18]. The application of deep learning in medical imaging has revolutionized healthcare, particularly in the domain of automated disease detection and diagnosis [19, 20]. Early advancements demonstrated the potential of convolutional neural networks (CNNs) in various medical imaging tasks, including tumor detection, organ segmentation and disease classification [21, 22]. Subsequent studies expanded the scope, applying deep learning to more complex challenges and various modalities such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT) scans, and histopathological images [23–25]. CNNs and ViTs have been utilized to automatically learn and extract relevant features from cervical cell images, leading to improved classification outcomes [26]. CNNs are some of the best tools for merging images, among many, as they are the most suited in medical imaging, where they can learn features at different levels and both low and high spatial frequency automatically. The recent performance of the ViT models that have beaten CNNs in classification methods dishes up an outstanding observation. The ultimate working mechanism of ViTs is the parallel processing of these images using the encoder to divide the input patch. This encoder has a self-attention module that can recognize long-range dependencies between image patches [27].

Considering the literature is examined, the effectiveness of deep learning in cervical cancer can be easily seen. In some review and survey articles, it has been stated that the effectiveness of deep learning in cervical cancer is at the clinical level and in some studies, it has been used in the clinic. Youneszade *et al.* [28] provide a comprehensive review of deep learning applications in cervical cancer diagnosis. They discuss various deep learning architectures and their potential in overcoming the limitations of traditional artificial intelligence techniques and manual screening methods. The study emphasizes the necessity of adopting advanced deep learning techniques to enhance the early detection of cervical cancer, reduce false negatives, and improve overall diagnostic precision. They also highlight the current opportunities and challenges in this field, underscoring the importance of large-scale, high-quality datasets for training effective deep learning models. Sambyal and Sarwar [29] highlight that integrating whole slide imaging (WSI) with deep learning technology has led to significant advancements in the screening and diagnosis of cervical cancer. Their review focuses on the evolution, limitations, and gaps in the use of deep learning algorithms with WSI, analyzing 37 selected studies for methodological insights. They examine popular deep learning techniques and current trends, recommending the application of transfer supervised learning. Jiang *et al.* [5] conducted a review highlighting the critical role of early detection and diagnosis of cervical cancer for effective clinical treatment and management. They examined over 80 publications since 2016, offering a thorough overview of deep learning-based Computer Aided Diagnosis (CAD) methods in cervical cytology screening. The review covers medical and biological knowledge, analyzes public cervical cytology datasets, and discusses image analysis techniques such as cell identification and abnormal cell detection. Their work underscores the effectiveness of deep learning in this field.

In this study, we utilized advanced computational techniques to develop a robust framework for the autonomous classification of cervical cancer cell images. By incorporating leading CNNs and ViTs with innovative data augmentation and transfer learning, our approach aims to achieve high accuracy in identifying and classifying Pap smear images and their abnormalities. The proposed system outperforms state-of-the-art methods using publicly available datasets. This integration of deep learning techniques holds significant potential for enhancing diagnostic accuracy and supporting early detection efforts in cervical cancer. Our key contributions are summarized as follows:

• We enhanced the dataset for deep learning algorithms by combining the publicly available Mendeley Liquid Based Cytology (LBC) and Malhari datasets. This comprehensive dataset supports the training of more reliable models and ensures a diverse representation of cervical cancer cell images.

• Our study evaluated leading-edge ViT-based architectures (Swin, PiT, MobileViT, DeiT3, totaling 14 models) and cutting-edge CNN-based architectures (MobileNetv3, EfficientNetv2, ConvNeXt, InceptionNeXt, totaling 14 models), making this one of the most extensive studies in the field with a total of 28 models tested.

• We utilized advanced data augmentation and transfer learning techniques to enhance the performance of these models. Our approach demonstrated that nearly all ViT-based models, as well as the EfficientNetv2-Small model from the CNN-based models, achieved a high accuracy of 99.45%, exceeding the current benchmarks in the literature.

• We conducted a comparative analysis of CNN and ViT-based models, providing valuable insights into their relative performance and highlighting the strengths and weaknesses of each approach.

• To ensure clinical relevance, we split the combined dataset into three distinct sets (train, validation, test). This structure allowed for a more accurate evaluation of model performance

on unseen test data, simulating real-world scenarios. The thorough preprocessing and splitting strategy ensured that the models were trained, validated and tested under conditions that best reflect clinical applications, thereby enhancing their generalization capabilities and overall reliability.

These contributions emphasize the robustness and efficiency of our proposed approach in delivering reliable and accurate cervical cancer classification, making it a valuable tool in medical diagnostics.

2. Related works

Cervical cancer detection and diagnosis has been significantly influenced by the latest deep learning technology, which has been brought forth. Despite these advancements, several research gaps remain. Models often have limited generalization due to small, homogeneous datasets, underscoring the need for larger, more diverse data. Many studies rely on publicly available datasets, which may not capture real-world clinical variability. Additionally, deep learning models' black-box nature hinders clinical trust, necessitating research on improving model interpretability. Most current studies focus on single-modality data, mainly Pap smear images, whereas integrating multi-modal data, such as patient demographics and clinical history, could enhance diagnostics. Recent developments include using transfer learning for better performance on cervical cancer datasets and developing hybrid models that combine CNNs and ViTs to boost accuracy.

There have been several research papers released that tackle different deep learning methods, both showing how powerful they are and the issues they face. The work of Sambyal and Sarwar [29] brings focus to how deep learning models for cervical cancer diagnosis have been changing over time, convincing us of the rewarding models such as DenseNet and EfficientNet. The treatment of cervical cancer through the use of artificial intelligence is unsettled by the subjectivity of the different databases and the requirement to settle the computational costs together with data shortages. Ahmadzadeh Sarhangi *et al.* [30] perform a critical analysis of CNN making use of cytology and colposcopy images. The impact of the change in the level of use of public datasets and the improvement of diagnosis accuracy are the other two main results highlighted. The paper further leads on the researchers' innovation in ensuring a new diagnostic approach that relies more on computerized analysis and less on manual inspection. Kang *et al.* [8] have an achievement in Raman spectroscopy when combined the very deep learning algorithms that were applied to the most common way of detecting cervical cancer at a very early stage in the experiment. Their findings uncover the potential of new methods to improve sensitive and fast diagnostics, in particular by falling upon the issue of disproportionate genders and overfitting.

Pacal conducted a study demonstrating the MaxCerVixT model's effectiveness both offline and in real-time, highlighting its high performance, low complexity and speed. The study also examined how attention mechanisms impact the model's efficiency. The proposed model shows high performance, significantly enhancing the effective diagnosis of cervical cancer [12]. The study by Gao *et al.* [31] focuses on using deep learning and adversarial networks to predict the likelihood of treatment plan approval for high-dose-rate brachytherapy in cervical cancer. Their approach combines dose prediction and plan-approval networks, enhancing the accuracy and efficiency of the automated treatment planning process. By utilizing adversarial networks to automate the evaluation process, the method aims to eliminate the ambiguity and imprecision associated with subjective planning. Mishra *et al.* [32] used a quantum invasive weed optimization technique combined with deep learning to classify cervical precancerous stages from Pap smear images. Their method includes preprocessing with Gabor filtering, feature extraction using SqueezeNet, and hyperparameter tuning via optimization. They achieved accuracy rates up to 99.09%, demonstrating the method's high effectiveness for cervical cancer screening and diagnosis.

Attallah [9] has designed a computer-aided diagnostic (CAD) system that combines deep learning and handcrafted descriptors for the diagnosis of cervical cancer. This new approach, known as the hybrid, makes it possible to use different domain features to enhance diagnostic accuracy thus, showing the high performance of complex Artificial Intelligence-based (AI-based) systems together with classical medical image processing. The research that was carried out proved the high performance of diagnostics, thus this case exhibits the potential of a hybrid approach in real-world applications of the medical field. Kalbhor *et al.* [33] present a hybrid methodology for cervical cancer prediction based on Pap smear images using pre-trained deep neural network models for feature extraction. They utilized models like AlexNet, ResNet-18, ResNet-50 and GoogLeNet for extracting features, followed by training different machine learning models on these features. The study found that the Simple Logistic Regression model achieved the highest accuracy of 95.14% with the AlexNet pre-trained model, demonstrating the effectiveness of combining deep learning with traditional machine learning classifiers for improving diagnostic accuracy. Devaraj *et al.* [34] utilized a dataset of cervical smear images to analyze and predict cervical cancer using three advanced deep learning models: ResNet50V2, InceptionV3 and Xception. These models were validated through cross-validation, and their performance was assessed using metrics such as accuracy, precision, recall and F1-score. Among the models, ResNet50V2 demonstrated the highest accuracy. The results indicate that deep learning techniques can accurately classify cervical cancer, significantly improving early diagnosis without the need for invasive procedures.

Ramu *et al.* [35] propose a novel approach to identify cervical cancer risk factors by combining Long Short-Term Memory (LSTM) with an evolutionary technique, Artificial Bee Colony (ABC). Despite some limitations in specificity, their model achieved a high accuracy of 98.68%, outperforming models like Support Vector Machine-Principal Component Analysis (SVM-PCA). Mathivanan *et al.* [36] introduce a groundbreaking methodology using pre-trained deep neural network models (AlexNet, ResNet-101, ResNet-152 and InceptionV3) for feature extraction. Fine-tuning these models with various machine learning algorithms, ResNet-152 achieved an impressive accuracy of 98.08%. The use of the publicly accessible SIPaKMeD dataset enhances the

transparency and reproducibility of their findings. This hybrid approach combines deep learning and machine learning for effective cervical cancer classification, enabling the extraction of intricate image features. Pacal *et al.* [37] present effective techniques for developing a more efficient diagnostic system using advanced deep learning methods. The study applies 40 CNN-based models and over 20 ViT-based models on the SIPaKMeD pap-smear dataset, utilizing data augmentation and ensemble learning to enhance model accuracy. Results show ViT-based models outperforming, and CNN models performing similarly. The study's extensive comparison highlights its potential for clinical implementation.

A comprehensive review of the literature indicates that deep learning technology has substantially advanced the diagnosis and detection of cervical cancer. Numerous studies underscore the high accuracy achieved by deep learning models, thereby diminishing the dependence on manual examination. Research in image processing and data analysis suggests that the utilization of public datasets can improve diagnostic precision. Furthermore, integrating deep learning algorithms with various imaging techniques facilitates the early detection of cancer. These advancements affirm the efficacy of deep learning as a robust tool for the swift and accurate diagnosis of cervical cancer.

3. Materials and methods

3.1 Datasets

Deep learning algorithms require substantial amounts of data to be effective. The performance of these models is heavily dependent on the dataset's quality and size. Small datasets may lead to overfitting and poor generalization, while larger datasets enable better generalization and model performance. In cervical cancer diagnosis, datasets with cytology-based Pap smear images are rare and typically small. This scarcity limits the effectiveness and generalizability of deep learning models, as highlighted in the literature. In this study, we utilized the publicly available Mendeley LBC [38] and Malhari datasets [39] to train and evaluate deep learning algorithms. The Mendeley LBC dataset comprises 963 images, and the Malhari dataset includes 318 images, resulting in a combined total of 1281 images. To enhance the generalization ability of the deep learning models, we combined both datasets as each is small in scale. By merging these datasets, we aimed to create a more robust and comprehensive dataset. Additionally, these datasets are more recent compared to other available datasets. This approach is critical for improving the performance of deep learning models and achieving more accurate predictions.

LBC is a method used in cervical cancer screening and is a variation of the Papanicolaou (Pap) smear test [40]. In LBC, cervical cells are collected and preserved in a liquid medium before being examined under a microscope. The results of this test are typically classified according to the Bethesda System. In this study, the publicly available data consisted of four classes, so the deep learning algorithms were trained on these four classes. The classes used are: NILM (Negative for Intraepithelial Lesion or Malignancy), LSIL (Low-grade Squamous Intraepithelial Lesion), HSIL (High-grade Squamous Intraepithelial Lesion) and SCC (Squamous Cell Carcinoma). NILM indicates a normal result where no cancer or precancerous lesions are found. LSIL denotes low-grade squamous intraepithelial lesions, which are mild cellular abnormalities often associated with low-risk HPV infections and are considered abnormal [41]. HSIL represents high-grade squamous intraepithelial lesions, indicating more severe cellular abnormalities and precancerous conditions, and is also considered abnormal. SCC stands for squamous cell carcinoma, indicating the presence of malignant (cancerous) cells, and is classified as abnormal [42].

3.1.1 Mendeley LBC dataset

The Mendeley LBC dataset contains 963 cervical cytological images collected from three reputable medical diagnostic centers [38]. With consent from 460 participants, these centers provided 613 normal (NILM) and 350 abnormal images, including 113 HSIL, 163 LSIL and 74 SCC images. Scanned at 40× magnification, these images are illustrated in Fig. 1 and summarized in Table 1.

3.1.2 Malhari dataset

The Malhari dataset includes both Pap smear and colposcopy images from the same patients. Patients consented to the use of their data for research and development purposes under a strict confidentiality agreement [39]. This dataset comprises information from 32 patients. Each patient has four colposcopy images and, in most cases, 10 image patches derived from a single Pap test image. The Malhari dataset contains a total of 318 images, categorized into various cervical cancer types. The distribution of images across these categories is as follows: 40 images of HSIL, 80 images of LSIL, 158 images NILM and 40 images of SCC. These categories comprehensively represent different cervical conditions. Table 1 provides an overview of the Malhari dataset. Randomly selected images from each category in the Malhari dataset are shown in Fig. 2.

3.1.3 Combined dataset

Due to the small scale of both the Malhari and Mendeley LBC datasets, we combined these datasets to meet the high data requirements of deep learning algorithms. Initially, each dataset was randomly split into 70% training, 15% validation, and 15% test data. Then, we combined the subsets from both datasets to create a unified dataset. By merging these datasets, we aimed to enhance the model's ability to generalize better on test data and provide more objective results suitable for clinical applications. This merging process resulted in a more robust and comprehensive dataset comprising information from 492 patients. The combined dataset includes a total of 203 HSIL images, 193 LSIL images, 771 NILM images and 114 SCC images. This integration increases the data quantity and diversity, enhancing model generalization to new data. The comprehensive dataset aims to improve training, leading to more accurate and reliable cervical cancer diagnoses. Table 1 details the image distribution across categories.

Fig. 3 displays a set of pie charts showing the class distribution in each dataset used in this study. Each pie chart represents one of the datasets (LBC, Malhari and Combined) and illustrates the proportion of HSIL, LSIL, NILM and SCC

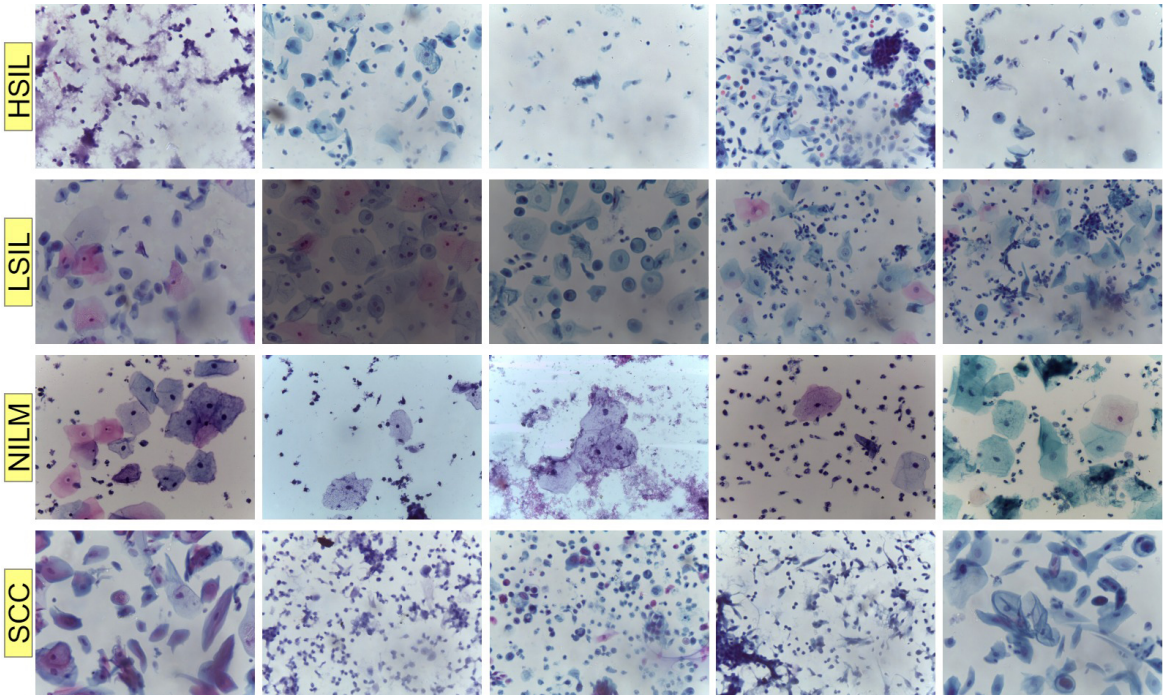

FIGURE 1. Sample images for each class in the Mendeley LBC dataset. HSIL: High-grade Squamous Intraepithelial Lesion; LSIL: Low-grade Squamous Intraepithelial Lesion; NILM: Negative for Intraepithelial Lesion or Malignancy; SCC: Squamous Cell Carcinoma.

TABLE 1. Information on each dataset used in this study.

Dataset	Patient	HSIL	LSIL	NILM	SCC
LBC dataset	460	163	113	613	74
Malhari Dataset	32	40	80	158	40
Combined Dataset	492	203	193	771	114

HSIL: High-grade Squamous Intraepithelial Lesion; LSIL: Low-grade Squamous Intraepithelial Lesion; NILM: Negative for Intraepithelial Lesion or Malignancy; SCC: Squamous Cell Carcinoma; LBC: Mendeley Liquid Based Cytology.

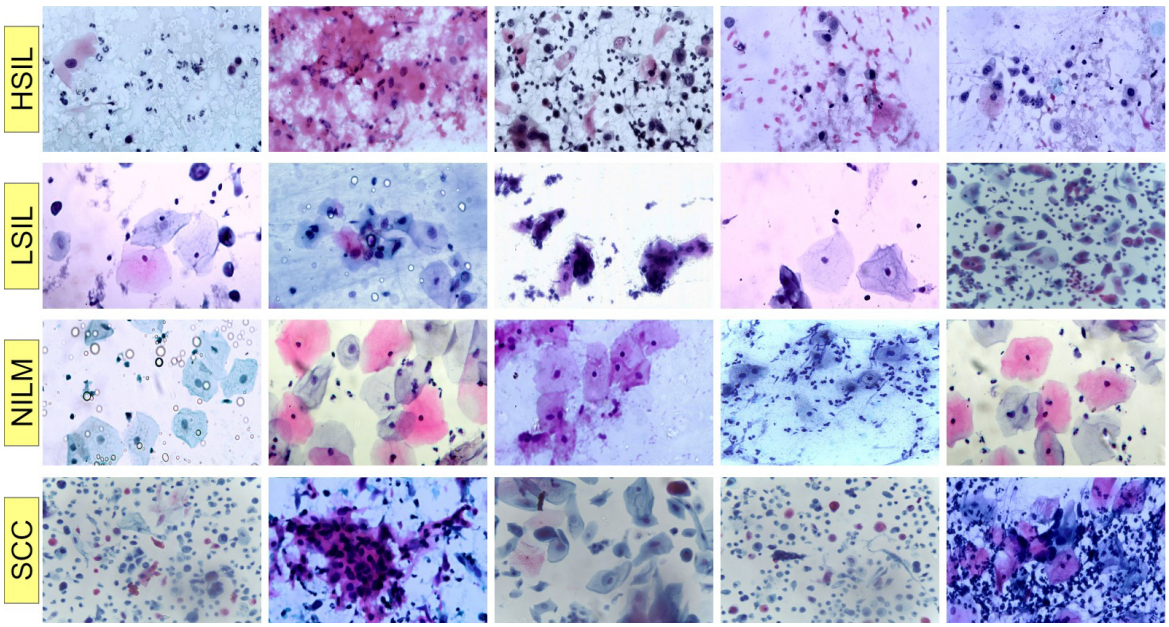

FIGURE 2. Sample images for each class in the Malhari dataset. HSIL: High-grade Squamous Intraepithelial Lesion; LSIL: Low-grade Squamous Intraepithelial Lesion; NILM: Negative for Intraepithelial Lesion or Malignancy; SCC: Squamous Cell Carcinoma.

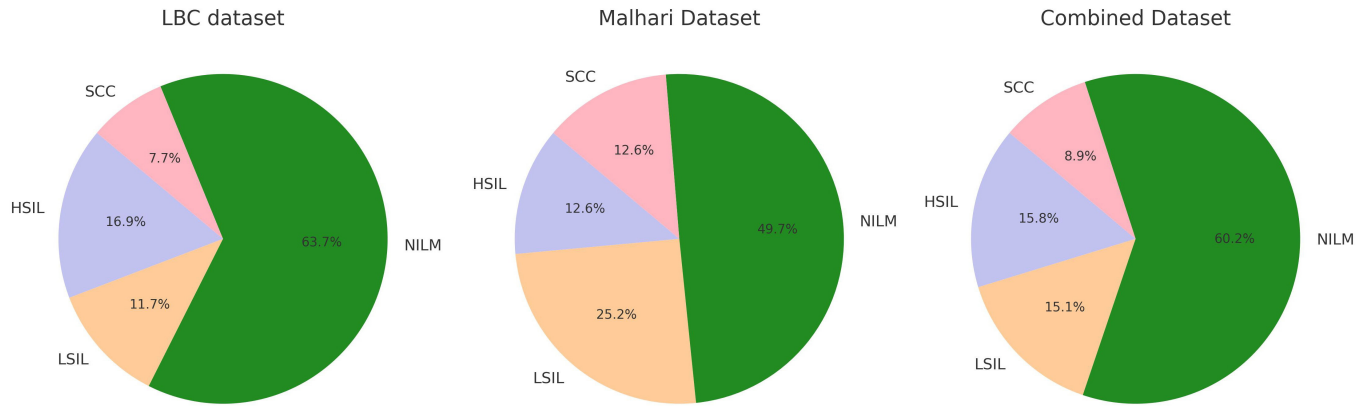

FIGURE 3. A set of pie chart for class-wise distribution of each dataset. HSIL: High-grade Squamous Intraepithelial Lesion; LSIL: Low-grade Squamous Intraepithelial Lesion; NILM: Negative for Intraepithelial Lesion or Malignancy; SCC: Squamous Cell Carcinoma; LBC: Mendeley Liquid Based Cytology.

categories within each dataset. This visual representation provides a clear and comparative view of how each category is distributed across the different dataset. As seen in the Fig. 3, the combined dataset, which includes both the Mendeley LBC and Malhari datasets, offers a comprehensive resource for cervical cancer diagnosis research. However, it's important to acknowledge the class imbalance present in these datasets. Specifically, the HSIL, LSIL, NILM and SCC categories are not evenly distributed, with NILM images making up the majority. Class imbalance can be problematic for deep learning models, as it might lead to biased predictions favoring the majority class. Despite this challenge, the state-of-the-art (SOTA) deep learning models employed in this study have shown significant advancements in addressing these imbalances. By incorporating advanced techniques such as data augmentation, these models can achieve high performance even when dealing with imbalanced datasets.

3.2 Deep learning approaches

Deep learning has greatly improved cervical cancer detection by using computer-aided analysis of Pap smear images. CNNs and ViTs, for example, classify images and detect anomalies with high accuracy. CNNs excel at capturing spatial hierarchies to detect subtle cellular abnormalities, while ViTs use self-attention mechanisms to capture long-range dependencies, complementing CNNs. Pretrained models are fine-tuned on specific cervical cancer datasets, enhancing adaptability and precision. This makes deep learning vital for modern cervical cancer diagnostics, aiding early detection and improving patient outcomes [43].

3.2.1 Convolutional neural networks

Convolutional Neural Networks (CNNs) are a widely used type of artificial neural network, particularly successful in tasks like image recognition and classification. The fundamental principle of CNNs is to process an image through multiple layers, each extracting more complex features [44]. The structure of a CNN consists of various layers performing specific functions as seen in Fig. 4. The first layer is the input layer, which stores raw data. This is followed by the convolutional layer, where dot products between the image patches and filters are computed to produce output volumes. This layer extracts local features from the images. After convolution, activation functions are applied. The next layer, the pooling layer, reduces the computational load by making the output of the previous layer more memory efficient. Pooling layers summarize the feature maps and make them more resistant to small changes in the input. Finally, fully connected layers flatten the output and compute the probabilities for class predictions.

In the convolutional layer, features are extracted by performing dot products between the input image and filters. This process is mathematically represented in Eqn. 1:

$$G[m,n] = (f \times h)[m,n] = \sum_j \sum_k h[j,k] f[m-j, n-k] \quad (1)$$

Here, G represents the feature map, f is the input image, h is the filter matrix, and m and n are the indices of the result matrix. Pooling layers, commonly implemented as max pooling or average pooling, downsample the feature maps. For instance, max pooling takes the maximum value in each pooling region, expressed in Eqn. 2:

$$G\,[i,\ j]\ =\ max_{a\ \leq\ i\ <\ a+H}\ max_{b\ \leq\ j\ <\ b+W}\ X\,[a,\ b] \quad (2)$$

In these equations, G is the pooled feature map, X is the input feature map, H and W are the height and width of the pooling region, and i and j are the indices of the result matrix. Fully connected layers connect every neuron in one layer to every neuron in the next, with outputs calculated in Eqn. 3:

$$y\ =\ f(Wx\ +\ b) \quad (3)$$

In this equation, y is the output vector, x is the input vector, W is the weight matrix, b is the bias vector, and f is the activation function. Activation functions introduce non-linearity into the model, enabling it to learn complex patterns. These mathematical foundations and layer functionalities work together in CNNs to create powerful models capable of recognizing and

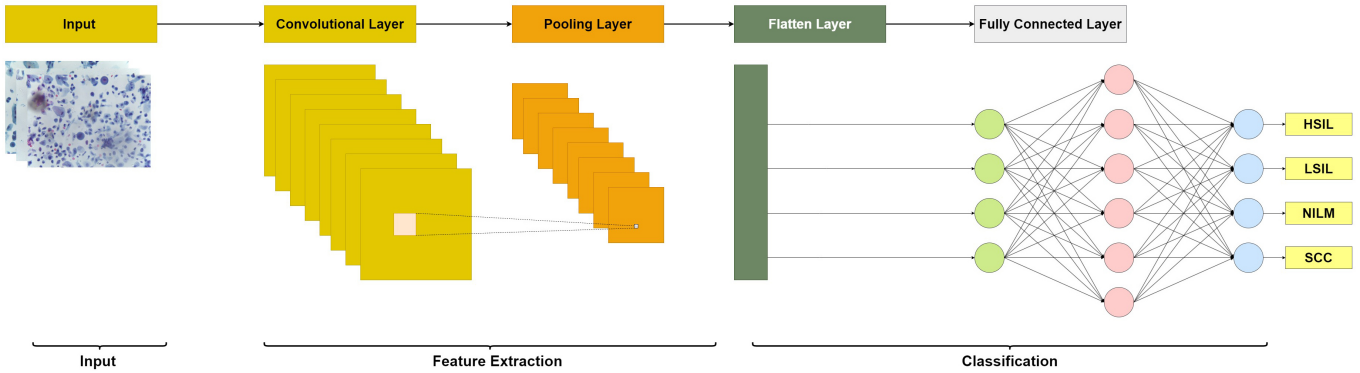

FIGURE 4. General structure of CNN architecture. HSIL: High-grade Squamous Intraepithelial Lesion; LSIL: Low-grade Squamous Intraepithelial Lesion; NILM: Negative for Intraepithelial Lesion or Malignancy; SCC: Squamous Cell Carcinoma.

classifying complex visual patterns.

3.2.2 Vision transformer approaches

Vision Transformers (ViTs), introduced by Dosovitskiy *et al.* [26], extend the Transformer model [45] to image processing by treating an image as a sequence of patches as seen in Fig. 5. Unlike traditional CNNs, ViTs use self-attention mechanisms to capture the global context, eliminating the need for hand-crafted visual features and inductive biases. This method utilizes larger datasets and increased computational power for enhanced performance.

The ViT structure comprises three main components: Patch Embedding, Transformer Encoder and Classification Head. In the Patch Embedding stage, an image is divided into small, fixed-size patches. Each patch is flattened and then projected to a specific dimension using a linear layer. This process represents the image as a series of vectors. Mathematically, patch embedding is expressed in Eqn. 4:

$$x_p \in R^{N\times\left(P^2C\right)}, X = x_p W_p \tag{4}$$

Here, x_p represents the matrix of flattened patches, $N = \frac{HW}{p^2}$ is the number of patches, $P \times P$ is the patch size, and W_p denotes the linear projection weights.

The Transformer Encoder consists of multiple layers, each containing two main components: Multi-Head Self-Attention (MHSA) and a Feed Forward Neural Network (FFN). The Self-Attention (SA) mechanism allows the model to learn long-range dependencies. Each SA block computes the query, key and value matrices and determines the attention scores in Eqn. 5:

$$A = softmax\left(\frac{QK^T}{\sqrt{D_q}}\right), Z = AV \tag{5}$$

Here, $Q = XW_Q$, $K = XW_K$, and $V = XW_V$ are the matrices. Multi-Head Self-Attention (MHSA) combines multiple SA blocks channel-wise to model complex dependencies among different elements in the input sequence, as formulated in Eqn. 6 and Eqn. 7.

$$MHSA(Q, K, V) = [Z_0, Z_1, \ldots, Z_{h-1}]W_O \tag{6}$$

$$Z_i = softmax\left(\frac{QW_{Q_i}\left(KW_{K_i}\right)^T}{\sqrt{\frac{D_q}{h}}}\right)VW_{V_i} \tag{7}$$

Finally, the output of the Transformer Encoder is typically connected to a classification head. This head consists of a fully connected layer that produces the final classification results as formulated in Eqn. 8:

$$Logits = FFN(CLS\ token\ output) \tag{8}$$

Here, the Classification Token (CLS) token output is the result from the token of the Transformer Encoder.

3.3 Data augmentation and preprocessing

Deep neural networks require a large amount of input data to train effectively. Data augmentation is a technique that can significantly improve the generalization capability of these models by artificially expanding the training dataset [46]. This method provides several benefits over other training approaches in deep learning: it helps prevent overfitting by increasing the dataset size, enhances performance in areas where data is limited, reduces the need for manual data labeling, and makes the model more robust to variations in input data, such as changes in lighting, orientation, and scale. Common data augmentation techniques include flipping, color space augmentation, cropping, rotation, translation and noise injection.

In the context of cervical cancer diagnosis using Pap smear images, data augmentation plays a crucial role. Obtaining a sufficient number of labeled Pap smear images for accurate diagnosis is challenging due to privacy issues and the labor-intensive nature of manual labeling. Given the limited size of the cervical cancer dataset, we employed various basic data augmentation techniques including rotation, scaling, flipping, noise injection, shear and translation. Additionally, meth-

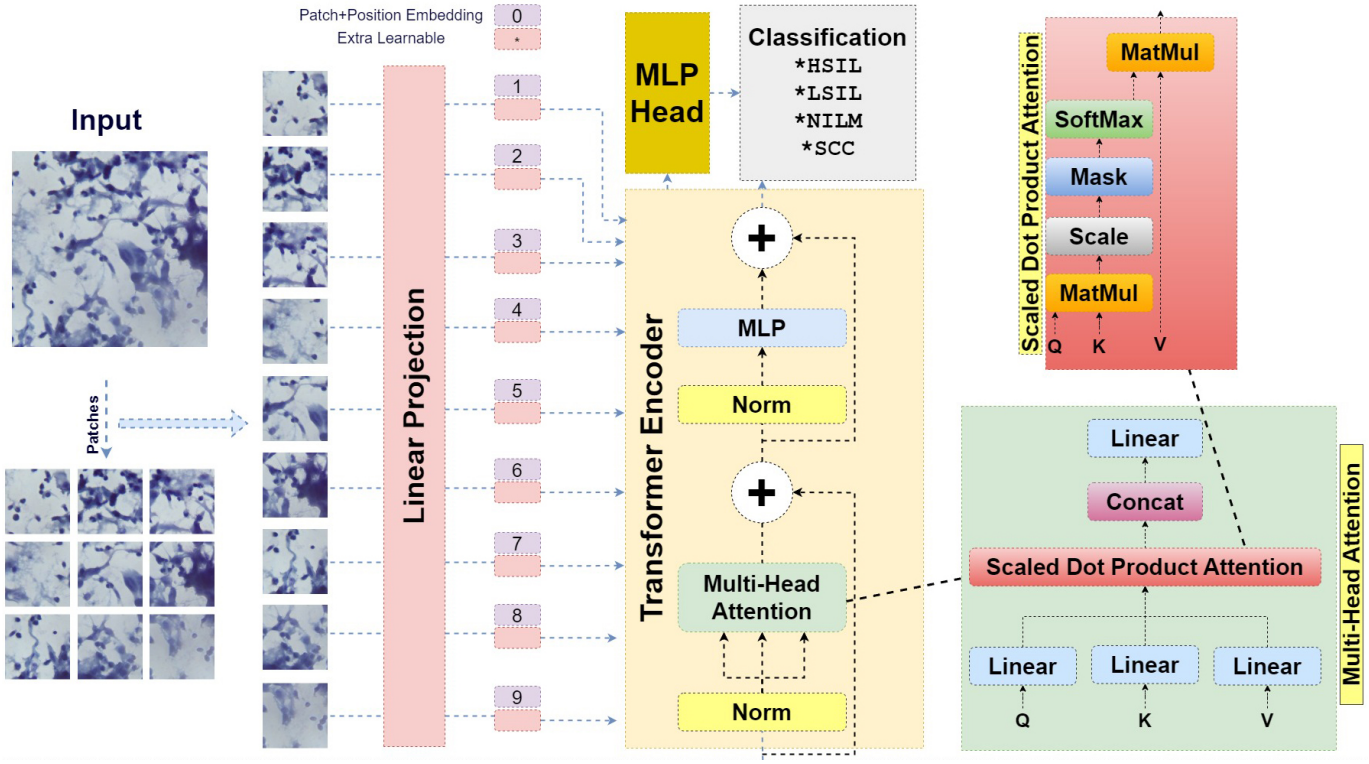

FIGURE 5. General structure of ViT architecture. HSIL: High-grade Squamous Intraepithelial Lesion; LSIL: Low-grade Squamous Intraepithelial Lesion; NILM: Negative for Intraepithelial Lesion or Malignancy; SCC: Squamous Cell Carcinoma; MLP: Multilayer Perceptron; *: Extra Learnable (Parameter).

ods such as brightness and contrast adjustment, zooming and random cropping were used to further diversify the training samples.

Data augmentation was performed online for all samples during training to enhance the diversity and robustness of the training data. This strategy ensured that the model was exposed to a wide range of variations, simulating real-world differences in Pap smear images, such as varying angles, sizes and lighting conditions of cervical cells. Moreover, we utilized state-of-the-art deep learning techniques to effectively address the data imbalance issue, ensuring that the model could generalize well across all classes despite the inherent class imbalance in the dataset.

To ensure a fair comparison across all models for cervical cancer classification, we standardized the input size to 224 × 224 pixels. This uniformity allows each model to process Pap smear images of the same dimensions, facilitating consistent performance evaluation. The combined dataset, detailed in Table 2, was divided into training, validation, and test sets in proportions of 70%, 15% and 15%, respectively. The dataset comprises four classes: HSIL, LSIL, NILM and SCC, totaling 1281 images. Each class was proportionally split to maintain balanced training and evaluation phases. Specifically, HSIL had 203 images (142 for training, 30 for validation, 31 for testing), LSIL had 193 images (135 for training, 29 for validation, 29 for testing), NILM had 771 images (540 for training, 116 for validation, 115 for testing), and SCC had 114 images (80 for training, 17 for validation, 17 for testing). This careful distribution ensures that each phase has a representative sample from each class.

TABLE 2. Data preprocessing of combined dataset.

Class	Total	Train (70%)	Validation (15%)	Test (15%)
HSIL	203	142	30	31
LSIL	193	135	29	29
NILM	771	540	116	115
SCC	114	80	17	17
Total	1281	897	192	192

HSIL: High-grade Squamous Intraepithelial Lesion; LSIL: Low-grade Squamous Intraepithelial Lesion; NILM: Negative for Intraepithelial Lesion or Malignancy; SCC: Squamous Cell Carcinoma.

Dividing the dataset into these three parts is essential for assessing the generalization ability of deep learning models in cervical cancer detection. The training set is used to develop the model, the validation set helps fine-tune its parameters, and the test set, which remains unseen during training, provides an objective measure of the model's performance. This method allows for a more accurate evaluation of how well the model can generalize to new, unseen Pap smear images, ensuring its effectiveness and reliability in practical applications. This comprehensive preprocessing and splitting strategy guarantee that the models are trained, validated and tested in conditions that best simulate real-world scenarios, thereby enhancing their generalization capabilities and overall performance in cervical cancer diagnosis.

3.4 Transfer learning

Improving the sensitivity and speed of deep learning models is essential for accurate and efficient cervical cancer diagnosis. Transfer learning is one of the most effective techniques to achieve this. It involves taking a model pre-trained on a large dataset, like ImageNet and fine-tuning it for a related, specific task [47]. For detecting cervical cancer, transfer learning offers significant benefits. By fine-tuning a pre-trained model, we can tailor the general features and patterns it learned from ImageNet to accurately identify cervical cancer cells. This method utilizes the power of previously acquired knowledge, enabling the model to classify or recognize cervical cancer images more accurately and quickly. Transfer learning not only saves time and effort in developing high-performance models but also enhances their effectiveness in complex tasks like image recognition. Instead of starting from scratch, which is both time-consuming and resource-intensive, transfer learning allows us to repurpose and refine existing models. This approach has been successful in various fields, including image identification and natural language processing, especially when training data is limited. In the context of cervical cancer diagnosis, transfer learning maximizes the model's potential by utilizing knowledge from a large dataset. This results in improved sensitivity and faster processing times, leading to more reliable and effective diagnostic outcomes, making it an invaluable tool in medical imaging.

3.5 Proposed approach

The proposed method for classifying cervical cancer in Pap smear images involves a detailed multi-stage process aimed at improving accuracy and reliability as illustrated in Fig. 6. First, a comprehensive dataset is created by merging the Mendeley LBC and Malhari datasets. This combined dataset undergoes various preprocessing steps such as resizing and splitting into training, validation, and test sets. Advanced data augmentation techniques are then applied to increase the dataset's diversity, addressing the challenges of limited data and enhancing the model's ability to generalize.

In the transfer learning phase, the final layers of pre-trained models are fine-tuned specifically for cervical cancer classification. This approach utilizes cutting-edge architectures from both ViT and CNN. Advanced ViT models like Swin, DeiT3, PiT and MobileViT, along with state-of-the-art CNN models such as EfficientNetV2, MobileNetV3, ConvNeXt and InceptionNeXt are employed. The selection of these 28 models was based on several criteria: The chosen models represent the latest advancements in CNN and ViT architectures known for their superior performance in image classification tasks. By including a variety of models, we aimed to cover a broad spectrum of architectural innovations, from traditional CNNs to more recent transformer-based approaches. Models were selected based on their proven robustness and ability to generalize well across different datasets in previous studies. The models were chosen for their scalability, enabling them to handle large datasets and complex medical imaging tasks effectively. Models with a track record of success in medical imaging, particularly in cancer diagnosis, were prioritized to ensure relevance and applicability to our study.

To ensure a fair comparison, all models are trained with default hyperparameters, enabling an unbiased determination of the best-performing model. The training phase is carefully conducted, with each model being trained on the augmented dataset. After training, the models are evaluated on an independent test set that was set aside earlier to ensure unbiased performance metrics. The evaluation includes metrics such as accuracy, precision, recall and F1-Score, providing a comprehensive assessment of each model's performance.

The final step involves thoroughly comparing the models based on these performance metrics to identify the most effective model for cervical cancer classification. This holistic approach aims to develop an efficient and reliable classification system for the early detection of cervical cancer, which is crucial for improving patient outcomes and reducing mortality rates. By incorporating advanced data augmentation, transfer learning, and state-of-the-art model architectures, this

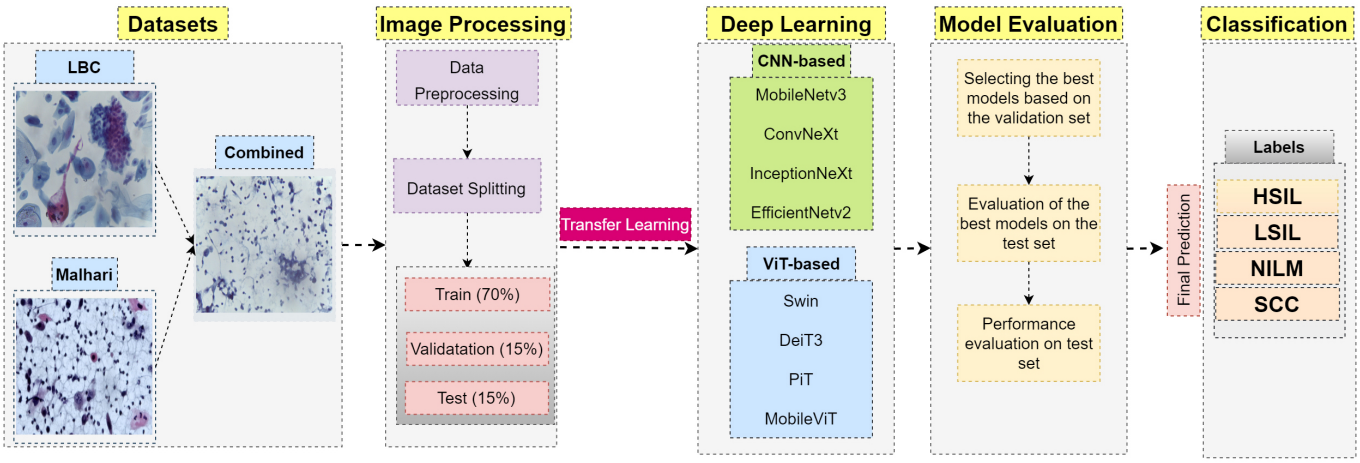

FIGURE 6. Proposed approach for robust classification of cervical pap smear images. HSIL: High-grade Squamous Intraepithelial Lesion; LSIL: Low-grade Squamous Intraepithelial Lesion; NILM: Negative for Intraepithelial Lesion or Malignancy; SCC: Squamous Cell Carcinoma; LBC: Mendeley Liquid Based Cytology; CNN: Convolutional Neural Networks; ViT: Vision Transformers.

approach offers a robust framework for cervical cancer diagnosis, promising significant advancements in early detection and treatment planning.

4. Results

4.1 Experimental setup

The model training and evaluation process is a critical component of developing robust and accurate diagnostic models for cervical pap smear images. In our study, we employed a comprehensive training regime incorporating transfer learning, data augmentation, and hyperparameter optimization to achieve optimal performance.

We trained our models for a maximum of 400 epochs with a batch size of 16 using the Stochastic Gradient Descent (SGD) optimizer with a learning rate (lr) of 0.01 and a momentum factor of 0.9 to accelerate convergence and avoid local minima. Additionally, we applied a weight decay (L2 regularization) of 1×10^{-4} to prevent overfitting. An early stopping mechanism was also employed, where training was halted if there was no improvement in validation loss for 10 consecutive epochs. This ensured that the model training stopped at the optimal point to prevent overfitting. The learning rate was decayed periodically using a StepLR scheduler with a step size of 30 epochs and a gamma factor of 0.1. The input size for the images was set to 224×224 pixels, which is standard for many deep learning models. The hyperparameters, including the learning rate, weight decay, momentum, step size and gamma, were optimized using the settings provided in the timm library, which offers a robust set of tools for fine-tuning deep learning models.

The models were trained and evaluated using the latest deep learning frameworks and libraries, leveraging advanced techniques to enhance their accuracy and reliability in diagnosing cervical cancer. Specifically, we used an RTX 3090 GPU, Ubuntu 22.04 operating system, CUDA 12.1 (NVIDIA Corporation, Santa Clara, CA, USA), cuDNN 8.9 (NVIDIA Corporation, Santa Clara, CA, USA), PyTorch 2.4.0 (Meta Platforms, Inc., Menlo Park, CA, USA) and Python 3.11 (Python Software Foundation, Beaverton, OR, USA). The system setup included an Intel i5 13th generation processor and 32GB of DDR5 RAM, ensuring sufficient computational power and memory for training deep learning models.

4.2 Evaluation metrics

To assess the classification performance of various methods, we utilized widely recognized metrics such as precision, recall, F1-score and accuracy. These metrics are essential for evaluating the effectiveness of models in binary classification tasks. Precision measures the proportion of true positive predictions out of all positive predictions, reflecting the model's confidence in its positive predictions. Recall indicates the model's ability to correctly identify all actual positive cases, highlighting its sensitivity or true positive rate. The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both aspects. Accuracy represents the overall correctness of the model, indicating the proportion of true results (both positive and negative) among all cases examined. For multi-class classification problems, we adopt a macro-averaging approach, where each metric is computed independently for each class and then averaged to provide a comprehensive evaluation. Precision and recall offer insights into the model's confidence and discriminative ability, respectively, while the F1-score and accuracy serve as comprehensive evaluation metrics. All these metrics yield values within the range of 0 to 1, with higher values indicating better performance. By employing these metrics, we can thoroughly compare the performance of cervical cancer classification methods, ensuring a detailed and rigorous assessment of their effectiveness. These formulas are depicted in Eqn. 9–Eqn. 12.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1\text{-}Score = \frac{(2 \times Precision \times Recall)}{(Precision + Recall)} \tag{12}$$

4.3 Results of CNN-based algorithms

In this section, we present the results of various CNN-based algorithms applied to the classification of cervical cancer in Pap smear images. The models evaluated include ConvNeXt [48], InceptionNeXt [49], EfficientNetV2 [50] and MobileNetV3 [51], each in different configurations such as Tiny, Small, Medium and Large. The results, as shown in Table 3, provide a comprehensive overview of the capabilities of these CNN-based algorithms in accurately identifying and classifying cervical cancer cells from Pap smear images.

As shown in Table 3, all evaluated CNN-based models demonstrated impressive performance on previously unseen test data from the combined dataset, achieving over 97% accuracy in classifying cervical cancer from Pap smear images as depicted in Fig. 7. This high level of accuracy highlights these models' strong ability to detect subtle cellular abnormalities indicative of cervical cancer. The models tested include various architectures like ConvNeXt, InceptionNeXt, EfficientNetV2 and MobileNetV3, each examined in different configurations to provide a comprehensive assessment. Within the ConvNeXt family, the ConvNeXt-Base model stood out with the highest performance, achieving an accuracy of 98.96% and an F1-score of 0.9833. In contrast, the ConvNeXt-Small model had the lowest performance in this group, with an accuracy of 97.40% and an F1-score of 0.9614. The InceptionNeXt models (Tiny, Small, Base) all showed similarly high performance, with accuracies of 98.44% and F1-scores ranging from 0.9777 to 0.9780, indicating consistent effective-

TABLE 3. Results of CNN-based models on combined dataset.

Model	Total Parameters (Million)	Accuracy	Precision	Recall	F1-score
ConvNeXt-Tiny	27.82	0.9844	0.9753	0.9811	0.9780
ConvNeXt-Small	49.46	0.9740	0.9604	0.9645	0.9614
ConvNeXt-Base	87.57	0.9896	0.9840	0.9833	0.9833
InceptionNeXt-Tiny	25.76	0.9844	0.9818	0.9747	0.9777
InceptionNeXt-Small	47.08	0.9844	0.9818	0.9747	0.9777
InceptionNeXt-Base	83.61	0.9844	0.9753	0.9811	0.9780
EfficientNet2-Small	20.18	0.9948	0.9978	0.9914	0.9945
EfficientNet2-Medium	52.86	0.9896	0.9840	0.9833	0.9833
EfficientNet2-Large	117.24	0.9896	0.9900	0.9828	0.9860
MobileNetv3-Small-075	1.02	0.9792	0.9622	0.9854	0.9734
MobileNetv3-Small-100	1.52	0.9896	0.9839	0.9831	0.9831
MobileNetv3-Large-075	2.72	0.9792	0.9730	0.9600	0.9661
MobileNetv3-Large-100	4.21	0.9792	0.9570	0.9790	0.9672

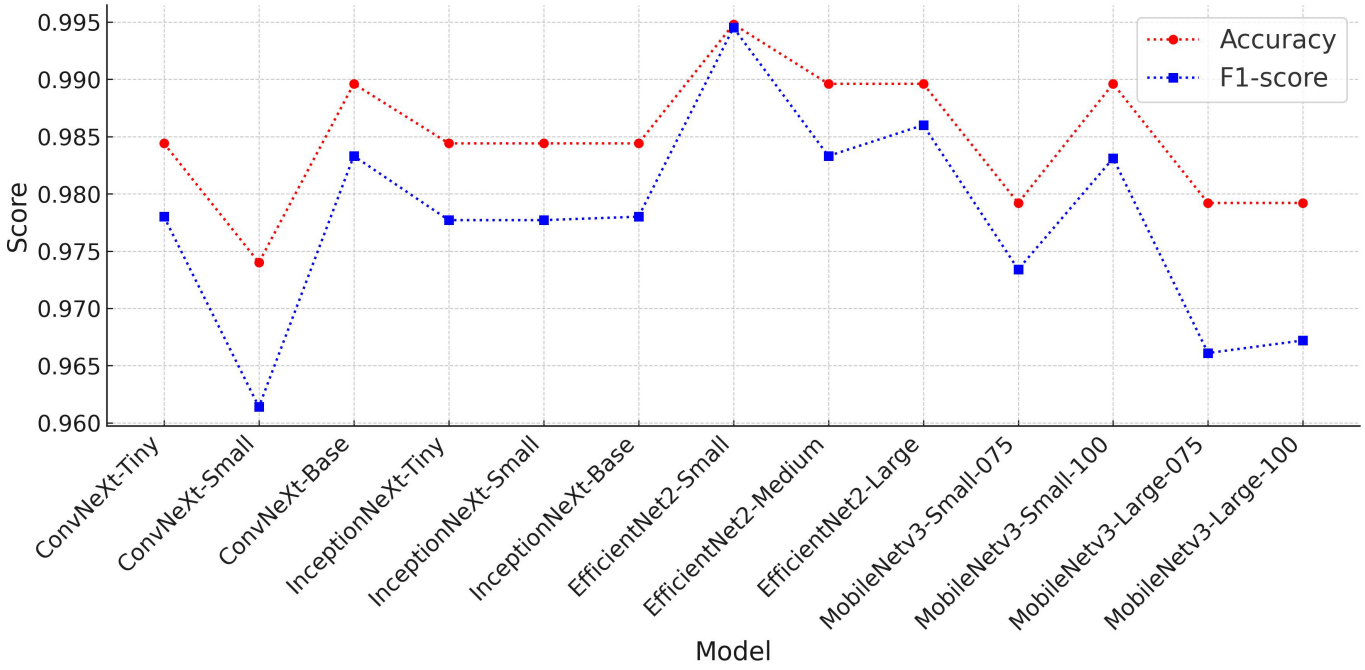

FIGURE 7. Accuracy and F1-score of CNN-based models on combined cervical dataset.

ness across different sizes. In the EfficientNetV2 family, the EfficientNet2-Small model excelled with outstanding scores across all metrics (Accuracy: 99.48%, Precision: 99.78%, Recall: 99.14%, F1-score: 0.9945). EfficientNet2-Large, despite having the highest parameter count, also performed very well, achieving an accuracy of 98.96% and an F1-score of 0.9860. The MobileNetV3 models proved that even smaller models could achieve high performance. The MobileNetv3-Small-100 achieved the highest accuracy at 98.96%, while MobileNetv3-Large-100 followed closely with an accuracy of 97.92%. Model complexity is determined by the total number of parameters (in millions), including weights and biases. More parameters usually improve performance but can increase overfitting. The MobileNetv3-Small-075 model, with 1.02 million parameters, achieves 0.9792 accuracy and 0.9734 F1-score. The EfficientNet2-Large model, with 117.24 million parameters, has 0.9896 accuracy and 0.9860 F1-score. EfficientNet2-Small, with 20.18 million parameters, achieves the highest accuracy (0.9948) and F1-score (0.9945). This shows that more parameters do not always guarantee the best performance for the cervical cancer dataset.

As seen in Fig. 7, the CNN-based models demonstrate strong performance in cervical cancer classification, as reflected by their high accuracy, precision, recall and F1-scores. ConvNeXt and InceptionNeXt models stand out due to their innovative architectural design that enhances feature extraction and model efficiency. Specifically, the ConvNeXt-Tiny and EfficientNet2-Small models show outstanding accuracy, achieving 0.9844 and 0.9948 respectively. EfficientNet2-Small also excels with a near-perfect F1-score of 0.9945,

highlighting its balance between precision and recall. The advantages of CNN-based models lie in their ability to capture spatial hierarchies in images, making them highly effective for medical image analysis where detecting subtle variations is crucial. These models benefit from extensive pretraining on large datasets, which aids in achieving robust performance even with the relatively smaller datasets used in this study. Furthermore, their scalable architecture allows for fine-tuning to specific tasks, enhancing their adaptability and application in diverse diagnostic scenarios.

4.4 Results of ViT-based algorithms

This section presents the performance outcomes of various ViT-based models used for classifying cervical cancer from Pap smear images. The models tested include Swin [52], DeiT3 [53], MobileViT [54] and PiT [55], each in different configurations to thoroughly evaluate their effectiveness. The results, summarized in Table 4, demonstrate the high accuracy and reliability of these advanced algorithms in detecting cervical cancer, highlighting their significant potential in medical diagnostics.

As seen in Table 4, all evaluated ViT-based models demonstrated exceptional performance on an unseen test dataset, which was previously separated from the combined dataset. This high level of accuracy underscores the robust capability of these models to detect subtle cellular abnormalities indicative of cervical cancer. The evaluation includes various architectures such as Swin, DeiT3, MobileViT and PiT, each tested in different configurations to provide a comprehensive assessment as depicted in Fig. 8.

Within the Swin family, all configurations (Tiny, Small, Base, Large) achieved an impressive accuracy of 99.48%. Notably, the Swin-Large model exhibited a slightly higher F1-score of 0.9945 compared to the other Swin models, which had an F1-score of 0.9888. For the DeiT3 models, the DeiT3-Large model achieved outstanding results with an accuracy of 99.48%, precision of 0.9978, recall of 0.9919, and an F1-score of 0.9948, making it one of the top performers. Both DeiT3-Small and DeiT3-Medium also performed exceptionally well, with accuracies of 98.96% and 99.48%, respectively, and F1-scores of 0.9863 and 0.9888.

The MobileViT family demonstrated that even the smallest models could achieve high performance, with MobileViT-xxS attaining an accuracy of 97.92% and an F1-score of 0.9664. The MobileViT-S model performed exceptionally well, matching the highest F1-scores of 0.9947. In the PiT family, the PiT-Tiny model had a slightly lower performance with an accuracy of 98.44% and an F1-score of 0.9838, while PiT-Small and PiT-Base both achieved high accuracies of 99.48%, with F1-scores of 0.9947 and 0.9945, respectively.

Each architecture family displayed strong performance, with Swin models showing consistent high accuracy across all configurations, and the Swin-Large model achieving the highest F1-score within its family. The DeiT3-Large model, with its near-perfect scores, stood out as one of the most effective models in this evaluation. MobileViT models, particularly the MobileViT-S, demonstrated that high performance could be achieved with fewer parameters. Similarly, the PiT models exhibited robust performance, with PiT-Small and PiT-Base achieving some of the highest scores across all metrics.

Considering ViT-based models in terms of parameters, the MobileViT-xxS model has the fewest parameters (0.95 million) and achieves 0.9792 accuracy and 0.9664 F1-score. In contrast, the DeiT3-Large model, with 303.35 million parameters, achieves 0.9948 accuracy and 0.9948 F1-score. Similarly, the Swin-Tiny model (27.52 million) and Swin-Large model (195 million) both reach 0.9948 accuracy. These results show that more parameters do not always guarantee better performance.

As seen in Fig. 8, ViT-based models in delivering reliable and accurate cervical cancer classification, making them valuable tools in medical diagnostics. The parameter counts of these models also highlight their efficiency, with models like

TABLE 4. Results of ViT-based models on combined dataset.

Model	Total Parameters (Million)	Accuracy	Precision	Recall	F1-score
Swin-Tiny	27.52	0.9948	0.9861	0.9919	0.9888
Swin-Small	48.84	0.9948	0.9861	0.9919	0.9888
Swin-Base	86.75	0.9948	0.9861	0.9919	0.9888
Swin-Large	195.00	0.9948	0.9978	0.9914	0.9945
DeiT3-Small	21.68	0.9896	0.9836	0.9892	0.9863
DeiT3-Medium	38.34	0.9948	0.9861	0.9919	0.9888
DeiT3-Base	85.82	0.9948	0.9917	0.9978	0.9947
DeiT3-Large	303.35	0.9948	0.9978	0.9919	0.9948
MobileViT-xxS	0.95	0.9792	0.9664	0.9664	0.9664
MobileViT-xS	1.93	0.9896	0.9778	0.9898	0.9834
MobileViT-S	4.94	0.9948	0.9917	0.9978	0.9947
PiT-Tiny	4.59	0.9844	0.9936	0.9747	0.9838
PiT-Small	22.89	0.9948	0.9917	0.9978	0.9947
PiT-Base	72.74	0.9948	0.9978	0.9914	0.9945

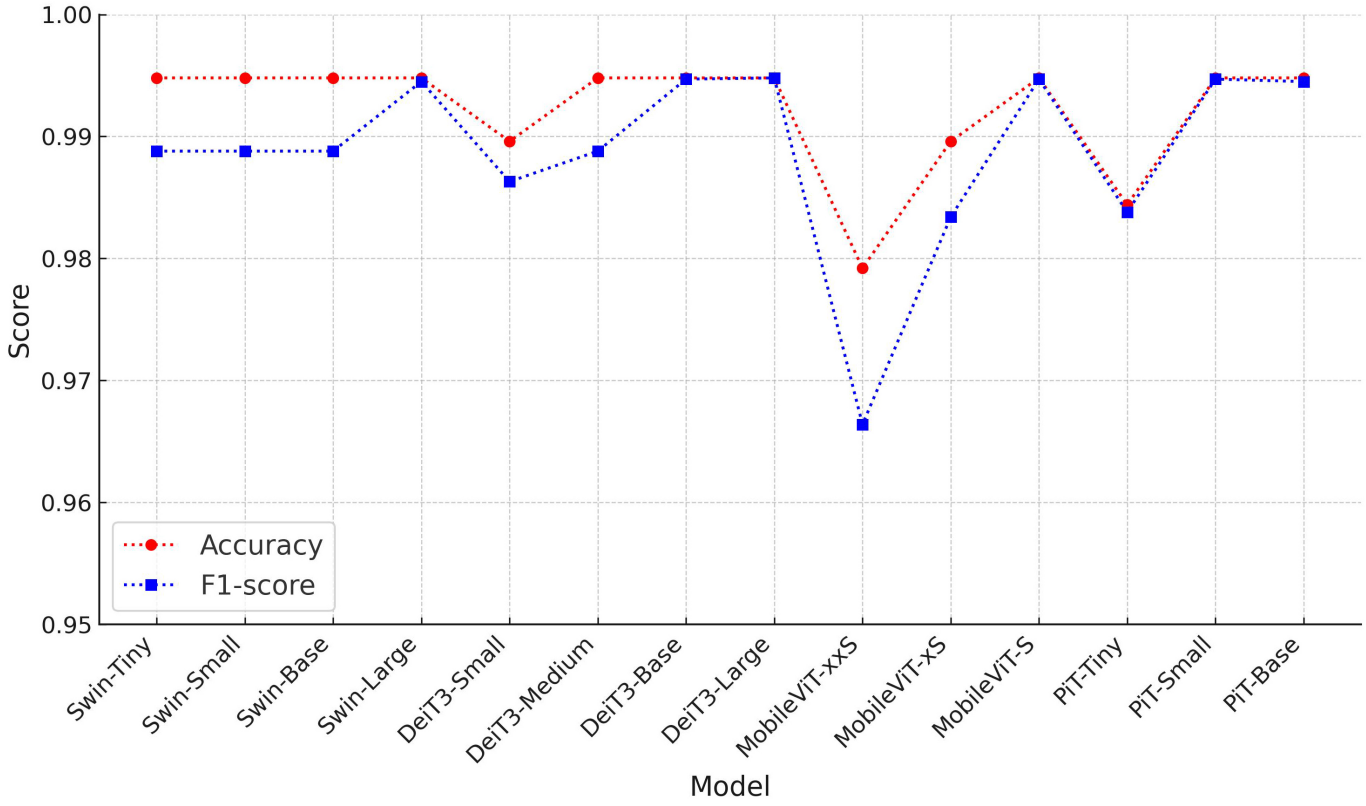

FIGURE 8. Accuracy and F1-score of ViT-based models on combined cervical dataset.

MobileViT achieving excellent performance despite having fewer parameters, indicating their potential for efficient and scalable deployment in clinical settings. Among all models (both CNN-based and ViT-based), the DeiT3-Large model provided the highest F1-score (0.9948), illustrating the model's superiority in balancing accuracy, precision, recall and computational complexity. The use of ViT-based models for cervical cancer classification from Pap smear images offers a promising approach to enhance diagnostic accuracy. The results suggest that these models, particularly the DeiT3-Large, can effectively identify cancerous cells with high precision and recall. This aligns with ongoing research and development in the field of medical diagnostics, where the integration of advanced machine learning models can significantly improve patient outcomes. Among all models (both CNN-based and ViT-based), the DeiT3-Large model provided the highest accuracy, with many models achieving the same accuracy value of 99.48%. The DeiT3-Large model also had the highest F1-score (99.48%). The classification report showcasing the class-wise performance of the DeiT3-Large model is presented in Table 5.

Table 5 highlights the impressive performance of the DeiT3-Large model in diagnosing cervical cancer, breaking down its accuracy for each class. For the HSIL class, the model achieved a perfect precision of 100% and a recall of 96.77%, leading to an F1-score of 98.36%. It excelled in the LSIL and SCC classes with flawless scores of 100% across precision, recall and F1-score. For the NILM class, the model showed 99.14% precision and 100% recall, achieving an F1-score of 99.57%. When looking at the overall performance, the DeiT3-Large model shines with an average precision of 99.78%, recall of 99.19%, and an F1-score of 99.48% on a macro level. The weighted averages for all these metrics are also consistently high at 99.48%. These outstanding results demonstrate that the DeiT3-Large model is exceptionally effective and reliable for cervical cancer diagnosis, delivering accurate predictions across all categories. The confusion matrix for the DeiT3-Large model, which demonstrated the highest performance in terms of accuracy, precision and F1-score, is shown in Fig. 9.

The confusion matrix for the DeiT3-Large model illustrates its performance in classifying cervical cancer images into four categories: HSIL, LSIL, NILM and SCC. The matrix shows that the model accurately classified 30 HSIL, 29 LSIL, 115 NILM and 17 SCC images, with only one misclassification where an HSIL image was predicted as NILM. This high level of accuracy, particularly the perfect classification in the NILM and SCC categories, demonstrates the robustness of the DeiT3-Large model. The visualization highlights the model's strong diagnostic capability, suggesting its potential for reliable clinical application in cervical cancer diagnosis.

5. Discussion

In this study, we examined the effectiveness of deep learning approaches in diagnosing cervical cancer, focusing particularly on the use of modern image-based models such as CNNs and ViTs. By combining two publicly available datasets, Mendeley LBC and Malhari, we created a more comprehensive dataset and evaluated our models solely on test data to achieve clinically more applicable results. We enhanced the training and performance of each model through advanced data augmentation techniques and transfer learning. Our results show that nearly all ViT-based models and CNN-based models

TABLE 5. Class-wise performance of the best model (DeiT3-Large).

Class	Precision	Recall	F1-score	Number of test images
HSIL	1.0000	0.9677	0.9836	31
LSIL	1.0000	1.0000	1.0000	29
NILM	0.9914	1.0000	0.9957	115
SCC	1.0000	1.0000	1.0000	17
Macro average	0.9978	0.9919	0.9948	192
Weighted average	0.9948	0.9948	0.9948	192

HSIL: High-grade Squamous Intraepithelial Lesion; LSIL: Low-grade Squamous Intraepithelial Lesion; NILM: Negative for Intraepithelial Lesion or Malignancy; SCC: Squamous Cell Carcinoma.

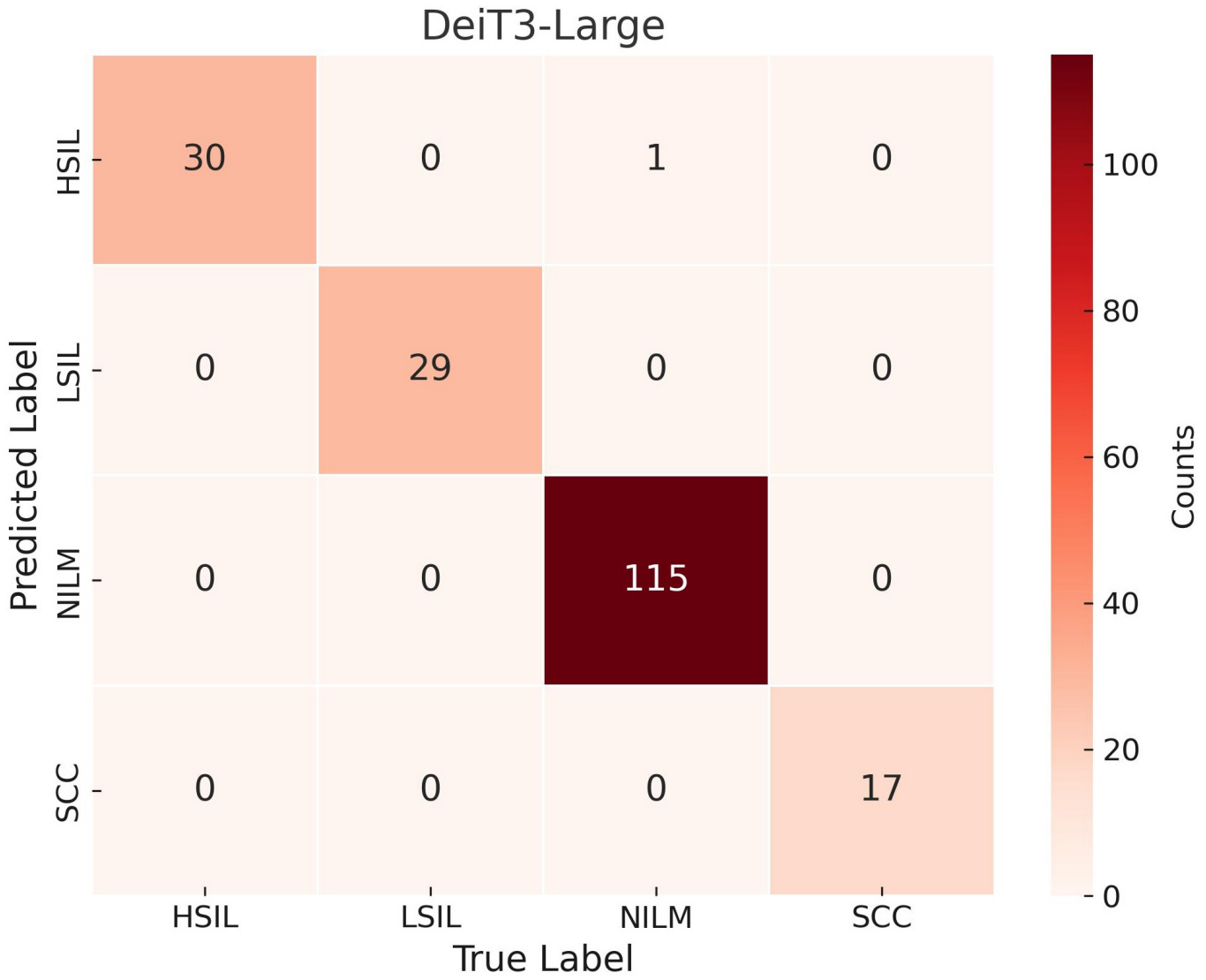

FIGURE 9. Confusion matrix of DeiT3-Large model. HSIL: High-grade Squamous Intraepithelial Lesion; LSIL: Low-grade Squamous Intraepithelial Lesion; NILM: Negative for Intraepithelial Lesion or Malignancy; SCC: Squamous Cell Carcinoma.

like EfficientNetv2-Small achieved high classification metrics, reaching 99.45% accuracy, with precision and recall rates exceeding 99%. These results were consistent across both the Mendeley LBC and Malhari datasets, indicating robust model performance. However, we encountered specific challenges such as class imbalance and variations in image quality. The advanced data augmentation techniques and transfer learning applied helped mitigate these issues, enhancing the model's generalization capabilities and performance consistency across different datasets. This highlights the potential of deep learning to improve cervical cancer diagnosis, offering a reliable and scalable solution for clinical applications. The success of these models is attributed to ViTs' ability to capture long-range dependencies and contextual information more effectively, and the efficiency of architectures like EfficientNetv2-Small in balancing computational demands with performance. The

impact of data augmentation and transfer learning techniques on model generalization is noteworthy. These techniques addressed issues common in medical image classification tasks, such as limited data access and class imbalance. The high performance achieved indicates that our approaches can classify cervical cancer reliably and accurately, thereby reducing the workload of cytopathologists and improving early diagnosis rates.

While our models achieved high accuracy, precision, recall and F1-scores, the interpretability of deep learning models remains a challenge. Future research should focus on developing methods to make these models more interpretable to clinicians, enhancing trust and facilitating their integration into clinical workflows. Additionally, integrating multi-modal data, such as combining Pap smear images with patient demographics and clinical history, could provide a more comprehensive diagnostic approach and further improve model performance. Moreover, the adaptability of our model presents opportunities for extension to other types of cancer and medical imaging tasks. For instance, our approach could be tailored to detect breast cancer, lung cancer or skin cancer by training the model on relevant datasets. Similarly, the model could be applied to different imaging modalities, such as MRI, CT scans, or ultrasound, to diagnose a broader range of conditions. By utilizing transfer learning and fine-tuning, our model can be adapted to these new tasks, potentially enhancing diagnostic accuracy and clinical outcomes across various medical fields. Exploring these potential applications and conducting further studies will not only validate our model's robustness but also contribute to the advancement of deep learning in healthcare. This expanded scope underscores the importance of continuous research to address current limitations and unlock new possibilities for improving patient care through innovative technology.

6. Limitations and future directions

This study has several limitations that must be acknowledged. Firstly, the combined dataset of Mendeley LBC and Malhari, while enhancing data diversity, remains relatively small compared to real-world clinical datasets. This limitation may restrict the generalizability of the findings to larger and more heterogeneous populations. Additionally, class imbalance within the dataset, particularly the overrepresentation of NILM images, could have influenced the model's performance despite the application of data augmentation techniques.

Secondly, the study focused exclusively on Pap smear images for classification, without integrating multi-modal data such as patient demographics or clinical history. Incorporating these additional data types in future studies could provide a more holistic and accurate diagnostic framework. Another significant limitation is the black-box nature of deep learning models, which may hinder clinical adoption due to the lack of interpretability and transparency. To address these limitations, future research should focus on expanding the dataset to include a more diverse and representative sample. Developing explainable AI methods will also be crucial to enhance trust and usability among clinicians. Additionally, exploring the integration of multi-modal data, such as combining cytology images with clinical and demographic information, could further improve diagnostic accuracy and applicability. These directions will not only enhance the robustness of deep learning models but also facilitate their integration into clinical workflows for cervical cancer diagnosis.

7. Conclusions

In this study, we employed advanced deep learning techniques, implementing 28 different models, including leading CNNs and ViTs, to classify cervical cancer from Pap smear images. To strengthen our dataset, we combined two publicly available datasets, Mendeley LBC and Malhari, and evaluated our models exclusively on test data to ensure fairer and more clinically relevant results. We also applied advanced data augmentation techniques and transfer learning to improve the training and performance of each model. Our experimental results showed that all deep learning models achieved over 97% accuracy on the test data. Among these, nearly all ViT-based models and only a few CNN-based models, such as EfficientNetv2-Small, achieved high classification metrics, with accuracy reaching up to 99.48%. This underscores the immense potential of deep learning in improving cervical cancer diagnosis, offering a reliable and scalable solution for clinical applications. The use of advanced ViT models like Swin, DeiT3, PiT and MobileViT, along with state-of-the-art CNN models such as EfficientNetV2, MobileNetV3, ConvNeXt and InceptionNeXt, represents a significant advancement in the field, enhancing both diagnostic accuracy and efficiency. These models have shown to be highly effective in pre-clinical stages and suggest that with the availability of more publicly accessible datasets, further research in this area will be encouraged.

Future research should focus on integrating multi-modal data, such as combining Pap smear images with patient demographics and clinical history, to provide a more comprehensive diagnostic approach. Additionally, exploring the interpretability of deep learning models and their integration into clinical workflows will be crucial for practical applications. Developing larger and more diverse datasets will also be essential for improving model generalization and reliability in real-world settings.

ABBREVIATIONS

CAD, Computer-Aided Diagnosis; CNN, Convolutional Neural Network; DeiT, Data-efficient Image Transformer; HSIL, High-grade Squamous Intraepithelial Lesion; LSIL, Low-grade Squamous Intraepithelial Lesion; NILM, Negative for Intraepithelial Lesion or Malignancy; SCC, Squamous Cell Carcinoma; SGD, Stochastic Gradient Descent; SOTA, State-of-the-Art; ViT, Vision Transformer; WSI, Whole Slide Imaging; Pap smear, Papanicolaou smear; LBC, Liquid Based Cytology; HPV, human papillomavirus; HIV, Human Immunodeficiency Virus; MRI, Magnetic Resonance Imaging; CT, Computed Tomography; AI, Artificial Intelligence; ABC, Artificial Bee Colony; LSTM, Long Short-Term Memory; SVM-PCA, Support Vector Machine-Principal Component Analysis; MHSA, Multi-Head Self-Attention; FFN, Feed Forward Neural Network; SA, Self-Attention; CLS, Classification Token; lr, learning rate.

AVAILABILITY OF DATA AND MATERIALS

The Mendeley LBC dataset and the Malhari dataset can be accessed at the following links, respectively: (Mendeley LBC Dataset) (https://data.mendeley.com/datasets/zddtpgzv63/4) and (Malhari Dataset) (https://data.mendeley.com/datasets/m5kxdj7m36/1).

AUTHOR CONTRIBUTIONS

IP—sole author responsible for the conception, design, execution, analysis and writing of the study.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

ACKNOWLEDGMENT

We would like to thank TÜSEB for their financial support and scientific contributions.

FUNDING

This work was supported by the grant provided by TÜSEB under the "2023-C1-YZ" call and Project No: "33934". Experimental computations were carried out on the computing units at Igdir University's Artificial Intelligence and Big Data Application and Research Center.

CONFLICT OF INTEREST

The author declares no conflict of interest.

REFERENCES

[1] Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. CA: A Cancer Journal for Clinicians. 2024; 74: 12–49.

[2] Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. CA: A Cancer Journal for Clinicians. 2023; 73: 17–48.

[3] Bedell SL, Goldstein LS, Goldstein AR, Goldstein AT. Cervical cancer screening: past, present, and future. Sexual Medicine Reviews. 2020; 8: 28–37.

[4] Scarinci IC, Garcia FA, Kobetz E, Partridge EE, Brandt HM, Bell MC, *et al*. Cervical cancer prevention: new tools and old barriers. Cancer. 2010; 116: 2531–2542.

[5] Jiang P, Li X, Shen H, Chen Y, Wang L, Chen H, *et al*. A systematic review of deep learning-based cervical cytology screening: from cell identification to whole slide image analysis. Artificial Intelligence Review. 2023; 56: 2687–2758.

[6] Rerucha CM, Caro RJ, Wheeler VL. Cervical cancer screening. American Family Physician. 2018; 97: 441–448.

[7] Mishra GA, Pimple SA, Shastri SS. An overview of prevention and early detection of cervical cancers. Indian Journal of Medical and Paediatric Oncology. 2011; 32: 125–132.

[8] Kang Z, Liu J, Ma C, Chen C, Lv X, Chen C. Early screening of cervical cancer based on tissue Raman spectroscopy combined with deep learning algorithms. Photodiagnosis and Photodynamic Therapy. 2023; 42: 103557.

[9] Attallah O. Cervical cancer diagnosis based on multi-domain features using deep learning enhanced by handcrafted descriptors. Applied Sciences. 2023; 13: 1916.

[10] Sahoo P, Saha S, Mondal S, Seera M, Sharma SK, Kumar M. Enhancing computer-aided cervical cancer detection using a novel fuzzy rank-based fusion. IEEE Access. 2023; 11: 145281–145294.

[11] Liu W, Li C, Xu N, Jiang T, Rahaman MM, Sun H, *et al*. CVM-Cervix: a hybrid cervical Pap-smear image classification framework using CNN, visual transformer and multilayer perceptron. Pattern Recognition. 2022; 130: 108829.

[12] Pacal I. MaxCerVixT: a novel lightweight vision transformer-based approach for precise cervical cancer detection. Knowledge-Based Systems. 2024; 289: 111482.

[13] Chen W, Gao L, Li X, Shen W. Lightweight convolutional neural network with knowledge distillation for cervical cells classification. Biomedical Signal Processing and Control. 2022; 71: 103177.

[14] Pacal İ. Deep learning approaches for classification of breast cancer in ultrasound (US) images. Journal of the Institute of Science and Technology. 2022; 12: 1917–1927.

[15] Pacal I. A novel Swin transformer approach utilizing residual multi-layer perceptron for diagnosing brain tumors in MRI images. International Journal of Machine Learning and Cybernetics. 2024; 15: 3579–3597.

[16] Sahoo P, Sharma SK, Saha S, Mondal S. A federated multi-stage lightweight vision transformer for respiratory disease detection. In Luo B, Cheng L, Wu ZG, Li H, Li C (eds.) Communications in Computer and Information Science (pp. 300–311). Springer: Singapore. 2023.

[17] Sahoo P, Saha S, Mondal S, Sharma N. COVID-19 detection from lung ultrasound images using a fuzzy ensemble-based transfer learning technique. International Conference on Pattern Recognition. 2022; 5170–5176.

[18] Sahoo P, Saha S, Sharma SK, Mondal S, Gowda S. A multi-stage framework for COVID-19 detection and severity assessment from chest radiography images using advanced fuzzy ensemble technique. Expert Systems with Applications. 2024; 238: 121724.

[19] Sahoo P, Saha S, Mondal S, Chowdhury S, Gowda S. Vision transformer-based federated learning for COVID-19 detection using chest X-ray. In Tanveer M, Agarwal S, Ozawa S, Ekbal A, Jatowt A (eds.) Communications in Computer and Information Science (pp. 77–88). Springer: Singapore. 2023.

[20] Ding W, Wang H, Huang J, Ju H, Geng Y, Lin CT, *et al*. FTransCNN: fusing transformer and a CNN based on fuzzy logic for uncertain medical image segmentation. Information Fusion. 2023; 99: 101880.

[21] Sahoo P, Sharma SK, Saha S, Jain D, Mondal S. A multistage framework for respiratory disease detection and assessing severity in chest X-ray images. Scientific Reports. 2024; 14: 12380.

[22] Attallah O. Skin-CAD: explainable deep learning classification of skin cancer from dermoscopic images by feature selection of dual high-level CNNs features and transfer learning. Computers in Biology and Medicine. 2024; 178: 108798.

[23] Attallah O. Acute lymphocytic leukemia detection and subtype classification via extended wavelet pooling based-CNNs and statistical-texture features. Image and Vision Computing. 2024; 147: 105064.

[24] Burukanli M, Yumuşak N. COVID-19 virus mutation prediction with LSTM and attention mechanisms. The Computer Journal. 2024; 67: 2934–2944.

[25] Pacal I, Celik O, Bayram B, Cunha A. Enhancing EfficientNetv2 with global and efficient channel attention mechanisms for accurate MRI-Based brain tumor classification. 2024. Available at: https://doi.org/10.1007/s10586-024-04532-1 (Accessed: 12 July 2024).

[26] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, *et al*. An image is worth 16×16 words: transformers for image recognition at scale. To be published in ArXiv. 2020. [Preprint].

[27] Rodrigo F, Alexandre P, Marta S, Anselmo DP, Ishak P, António C. Enhancing image annotation with object tracking and image retrieval: a systematic review. 2024. Available at: https://doi.org/10.1109/ACCESS.2024.3406018 (Accessed: 12 July 2024).

[28] Youneszade N, Marjani M, Pei CP. Deep learning in cervical cancer diagnosis: architecture, opportunities, and open research challenges. IEEE Access. 2023; 11: 6133–6149.

[29] Sambyal D, Sarwar A. Recent developments in cervical cancer diagnosis using deep learning on whole slide images: an overview of models, techniques, challenges and future directions. Micron. 2023; 173: 103520.

[30] Ahmadzadeh Sarhangi H, Beigifard D, Farmani E, Bolhasani H. Deep learning techniques for cervical cancer diagnosis based on pathology and colposcopy images. Informatics in Medicine Unlocked. 2024; 47: 101503.

[31] Gao Y, Gonzalez Y, Nwachukwu C, Albuquerque K, Jia X. Predicting treatment plan approval probability for high-dose-rate brachytherapy of cervical cancer using adversarial deep learning. Physics in Medicine & Biology. 2024; 69: 095010.

[32] Mishra AK, Gupta IK, Diwan TD, Srivastava S. Cervical precancerous lesion classification using quantum invasive weed optimization with deep learning on biomedical pap smear images. Expert Systems. 2024; 41: e13308.

[33] Kalbhor M, Shinde S, Joshi H, Wajire P. Pap smear-based cervical cancer detection using hybrid deep learning and performance evaluation. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. 2023; 11: 1615–1624.

[34] Devaraj S, Madian N, Menagadevi M, Remya R. Deep learning approaches for analysing papsmear images to detect cervical cancer. Wireless Personal Communications. 2024; 135: 81–98.

[35] Ramu K, Ananthanarayanan A, Josephson PJ, Paul NRR, Tumuluru P, Divya C, *et al*. Augmenting cervical cancer analysis with deep learning classification and topography selection using artificial bee colony optimization. SN Computer Science. 2024; 5: 703.

[36] Mathivanan SK, Francis D, Srinivasan S, Khatavkar V, K P, Shah MA. Enhancing cervical cancer detection and robust classification through a fusion of deep learning models. Scientific Reports. 2024; 14: 10812.

[37] Pacal I, Kılıcarslan S. Deep learning-based approaches for robust classification of cervical cancer. Neural Computing and Applications. 2023; 35: 18813–18828.

[38] Hussain E, Mahanta LB, Borah H, Das CR. Liquid based-cytology Pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions. Data in Brief. 2020; 30: 105589.

[39] Kalbhor M, Shinde SV. Malhari dataset. 2023. Available at: https://doi.org/10.17632/M5KXDJ7M36.1 (Accessed: 12 July 2024).

[40] Lasch L, Rathat G, Du Thanh A. Squamous cell carcinoma antigen elevation in cervical cancer follow-up: the forest hiding the tree. European Journal of Gynaecological Oncology. 2018; 39: 324–326.

[41] Yoshida H, Yamamoto M, Shigeta H. Successful treatment of uterine cervical carcinoma with extensive vaginal lesions using laparoscopic surgery: a case report. European Journal of Gynaecological Oncology. 2020; 41: 629–633.

[42] Monika EG, Radosław S, Natalia S, Bartosz B, Stefan S. Synchronous ovarian, endometrial and cervical cancer: case report. European Journal of Gynaecological Oncology. 2021; 42: 1093–1094.

[43] Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015; 521: 436–444.

[44] Kunduracioglu I, Pacal I. Advancements in deep learning for accurate classification of grape leaves and diagnosis of grape diseases. Journal of Plant Diseases and Protection. 2024; 131: 1061–1080.

[45] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, *et al*. Attention is all you need. Advances in Neural Information Processing Systems. 2017; 5999–6009.

[46] Lubbad M, Karaboga D, Basturk A, Akay B, Nalbantoglu U, Pacal I. Machine learning applications in detection and diagnosis of urology cancers: a systematic literature review. Neural Computing and Applications. 2024; 36: 6355–6379.

[47] Pacal I, Alaftekin M, Zengul FD. Enhancing skin cancer diagnosis using swin transformer with hybrid shifted window-based multi-head self-attention and SwiGLU-based MLP. Journal of Imaging Informatics in Medicine. 2024. Available at: https://doi.org/10.1007/s10278-024-01140-8 (Accessed: 12 July 2024).

[48] Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. A ConvNet for the 2020s. 2022. Available at: http://arxiv.org/abs/2201.03545 (Accessed: 12 July 2024).

[49] Yu W, Zhou P, Yan S, Wang X. InceptionNeXt: when inception meets ConvNeXt. 2023. Available at: http://arxiv.org/abs/2303.16900 (Accessed: 12 July 2024).

[50] Tan M, Le QV. EfficientNetV2: Smaller Models and Faster Training. 2021. Available at: https://arxiv.org/abs/2104.00298v3 (Accessed: 12 July 2024).

[51] Howard A, Sandler M, Chen B, Wang W, Chen LC, Tan M, *et al*. Searching for mobileNetV3. Proceedings of the IEEE International Conference on Computer Vision. Institute of Electrical and Electronics Engineers Inc.: Seoul, Korea (South). 2019.

[52] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, *et al*. Swin transformer: hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021; 10012–10022.

[53] Touvron H, Cord M, Jégou H. DeiT III: revenge of the ViT. In Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T (eds.) Lecture Notes in Computer Science (pp. 516–533). Springer: Cham. 2022.

[54] Mehta S, Rastegari M. MobileViT: light-weight, General-purpose, and mobile-friendly vision transformer. 2021. Available at: http://arxiv.org/abs/2110.02178 (Accessed: 12 July 2024).

[55] Heo B, Yun S, Han D, Chun S, Choe J, Oh SJ. Rethinking spatial dimensions of vision transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021; 11936–11945.

How to cite this article: Ishak Pacal. Investigating deep learning approaches for cervical cancer diagnosis: a focus on modern image-based models. European Journal of Gynaecological Oncology. 2025; 46(1): 125-141. doi: 10.22514/ejgo.2025.012.