Artificial Intelligence Applications And Innovations Vol. 2 No. 1 (2025)

Received 22 November 2025
Revised 18 December 2025
Accepted 10 May 2026
Published online 1 January 2025

Automated Cervical Cancer Detection Using Feature-Fused Deep CNNs and Ensemble Learning

Mahshid. ZamanVaziri[1], Niloofar. Rastin[2*], Shokufeh. Yaraghi[1]

[1] Department of Computer Engineering, Faculty of Engineering, Shahid Ashrafi Esfahani University, Isfahan, Iran
[2] Faculty of Computer Engineering, Iranian eUniversity, Tehran, Iran

*** Corresponding author email address**: niloofar.rastin@iranian.ac.ir

A r t i c l e I n f o

Article type:
Original Research

How to cite this article:
ZamanVaziri, M., Rastin, N., & Yaraghi, S. (2025). Automated Cervical Cancer Detection Using Feature-Fused Deep CNNs and Ensemble Learning. *Artificial Intelligence Applications and Innovations, 2*(1), 74-92.
https://doi.org/10.61838/jaiai.2.1.6

A B S T R A C T

Cervical cancer remains a significant global health concern, ranking as the fourth most common cancer among women worldwide. Early detection through Pap smear screening is vital for improving treatment outcomes, yet manual examination is time-consuming and subject to inter-observer variability. To address this challenge, this paper proposes an automated cervical cancer detection framework based on Deep Learning (DL) and ensemble classification. The proposed method leverages transfer learning to extract discriminative features from three architecturally complementary pre-trained Convolutional Neural Networks (CNNs): InceptionV3, InceptionResNetV2, and MobileNetV2. These networks are fine-tuned on Pap smear images, and their deep feature representations are concatenated to form a unified and enriched feature space. To effectively handle the resulting high-dimensional fused features, multiple classifiers are evaluated, among which a Bagging ensemble with Random Forest (RF) base learners demonstrates the most robust and consistent performance. Experimental evaluations conducted on three benchmark datasets—SIPaKMeD, Herlev, and Mendeley Liquid-Based Cytology—show that the proposed fusion-and-ensemble framework consistently outperforms individual CNN baselines and several state-of-the-art methods. The proposed model achieves 97.25% accuracy, 97.26% precision, 97.28% recall, and a 97.26% F1-score on the SIPaKMeD dataset, along with accuracies of 96.72% and 99.47% on the Herlev and Mendeley LBC datasets, respectively. These results demonstrate the effectiveness and robustness of the proposed approach for automated cervical cancer detection.

Keywords: *Cervical Cancer, Pap Smear Classification, Deep Learning, Feature Fusion, Convolutional Neural Networks, Transfer Learning*

1. Introduction

Cervical cancer is the fourth most prevalent cancer among women worldwide and remains a major public health concern, accounting for more than 340,000 deaths annually [1]. Its development is strongly associated with persistent infection by high-risk human papillomavirus (HPV) subtypes and the oncogenic activity of viral proteins E5, E6, and E7 [1]. Early diagnosis plays a crucial role in reducing mortality, as timely detection of precancerous lesions enables effective intervention. Among screening techniques, cytological examination using the Papanicolaou (Pap smear) test has been one of the most reliable and widely adopted methods for cervical cancer screening [2]. Since its introduction in the 1950s, the Pap smear test has

contributed to a reduction in cervical cancer mortality by approximately 70% [2]. The test involves collecting cervical cell samples and examining them microscopically to identify abnormal cellular changes.

Despite its effectiveness, manual analysis of Pap smear images is time-consuming, labor-intensive, and prone to inter-observer variability. As a result, computer-aided detection (CAD) systems have emerged as a promising alternative to support cytopathologists and improve diagnostic reliability [3]. Recent advances in deep DL, particularly CNNs, have significantly enhanced the performance of CAD systems by enabling automated extraction of discriminative image features [3-7]. These methods reduce human error, increase screening efficiency, and offer consistent diagnostic support in clinical workflows.

CNNs are specifically designed to capture hierarchical spatial features through convolutional, pooling, and fully connected layers, making them highly effective for image recognition and medical image analysis tasks [8]. However, training deep CNNs from scratch requires large annotated datasets, which are often unavailable in medical imaging domains. To address this limitation, transfer learning has been widely adopted as an effective solution [9]. In transfer learning, CNNs pre-trained on large-scale datasets such as ImageNet are fine-tuned for domain-specific tasks, allowing robust feature extraction even with limited labeled data. Several studies have successfully applied transfer learning to cervical cancer detection using architectures such as ResNet [10], Inception [11], VGG [12], and DenseNet [13], demonstrating notable performance improvements.

More recently, hybrid frameworks combining multi-CNN feature extraction with feature fusion and ensemble classifiers have gained attention in cervical cytology analysis. While these approaches often improve accuracy, many existing studies rely on heuristic fusion strategies, dimensionality reduction techniques (e.g., Principal Component Analysis (PCA)), fuzzy decision rules, or weighted voting mechanisms. Moreover, the interaction between fused deep features and ensemble classifiers is often underexplored, and the design choices behind fusion and classification strategies are not systematically justified.

In medical imaging scenarios characterized by limited and imbalanced datasets, methodological rigor, principled architectural selection, and systematic validation of design choices are critical for ensuring robustness and clinical applicability. Accordingly, rather than focusing solely on performance improvement, this work emphasizes a carefully justified integration strategy that analyzes the interaction between feature fusion and ensemble classification within a unified and experimentally validated framework.

To address these gaps, this paper proposes a feature-fusion and ensemble-based classification framework for automated cervical cancer detection using Pap smear images. The proposed method employs transfer learning to extract deep features from three architecturally diverse pre-trained CNNs—MobileNetV2, InceptionResNetV2, and InceptionV3—which are fine-tuned to the cervical cytology domain. The extracted features are concatenated to form a unified representation that integrates complementary multi-scale, deep discriminative, and lightweight local features. Unlike conventional fusion approaches that primarily aim to increase feature dimensionality, the proposed strategy emphasizes architectural complementarity and robustness.

For classification, multiple machine learning and ensemble strategies are systematically evaluated, including Support Vector Machines (SVM), Gaussian Naive Bayes, Gradient Boosting, RF, and Bagging with RF as the base estimator. Experimental analysis across three benchmark Pap smear datasets demonstrates that the Bagging–RF ensemble consistently achieves superior and more stable performance when applied to high-dimensional fused deep features. This robustness is particularly well-suited to medical imaging scenarios, where annotated datasets are relatively limited and class distributions are often imbalanced.

The effectiveness of the proposed framework is evaluated on three widely used benchmark Pap smear datasets: SIPaKMeD [14], Mendeley Liquid-Based Cytology (LBC) [15], and Herlev [16]. Comparative experiments against state-of-the-art methods indicate that the proposed approach achieves competitive or superior performance across all datasets, highlighting its robustness and generalizability.

The main contributions of this work lie in the principled design, systematic validation, and robustness-oriented integration of multi-CNN feature fusion and ensemble learning for cervical cytology classification, and are summarized as follows:

1. Architecturally complementary multi-CNN feature fusion: A feature-fusion framework is proposed that integrates deep features extracted from three architecturally

diverse pre-trained CNNs—MobileNetV2, InceptionResNetV2, and InceptionV3. By explicitly leveraging architectural complementarity rather than relying on heuristic weighting or dimensionality reduction, the proposed fusion strategy captures multi-scale, deep discriminative, and lightweight local cytological features within a unified representation.

2. Ensemble-aligned classification strategy for fused deep features: A systematic evaluation of multiple classification and ensemble learning strategies demonstrates that Bagging with RF base learners is particularly effective for high-dimensional fused deep features. The proposed classifier choice is motivated by its variance-reduction capability and robustness under limited-data conditions, which are common in medical image analysis.

3. Ablation-driven justification of fusion and classifier design: Comprehensive ablation experiments across three benchmark Pap smear datasets analyze the impact of classifier selection and ensemble design on fused feature representations. The results demonstrate that the proposed fusion configuration, combined with the Bagging with RF ensemble, yields more stable and accurate performance than single classifiers and alternative ensemble methods, providing empirical justification for the overall pipeline design.

4. Robust and partially interpretable decision-making: The use of Random Forest within a bagging ensemble enhances robustness and provides a degree of transparency compared to fully black-box classifiers, owing to its tree-based decision structure. This enables limited qualitative inspection of feature importance and decision behavior while maintaining strong generalization performance.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed methodology. Section 4 presents and discusses experimental results. Section 5 concludes the paper and outlines future research directions.

2. Literature Review

Nguyen et al. [17] proposed a DL approach that leverages transfer learning and feature fusion. They utilized three pre-trained CNN models—InceptionV3, ResNet152, and Inception-ResNetV2, all trained on ImageNet—to extract features from cervical cells. These features were then concatenated and used to train two fully connected layers for classification.

Lin et al. [18] trained four pre-trained CNNs. They trained the CNNs using image data containing both shape and visual features and artificially increased the amount of training data to improve performance. Finally, they combined the outputs from the CNNs to make a final classification.

Dongyao Jia et al. [19] proposed a DL framework for cervical cancer detection that integrates feature extraction from multiple sources. Robust features were obtained using the Grey-Level Co-occurrence Matrix and Gabor filters applied to grayscale images. These features, along with those extracted by a CNN, were combined and input to an SVM for final classification.

Bhatt et al. [20] employed an advanced image resizing technique and used multiple pre-trained CNN architectures, including EfficientNet, VGG, and ResNet, each independently for cervical cell classification. In addition, traditional machine learning classifiers such as RF, K-Nearest Neighbors (KNN), and SVM were also applied. When using CNNs, Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations highlighted regions corresponding to precancerous lesions or cancerous tumors, providing interpretable insights into the model's predictions.

Khampaeria et al. [21] introduced a novel Internet of Health Things framework for cervical cancer diagnosis. For abnormal cervical cell classification, they utilized several well-known pre-trained CNN architectures—InceptionV3, VGG19, SqueezeNet, and ResNet50—as feature extractors, combined with machine learning classifiers such as KNN, Naive Bayes, Logistic Regression (LR), RF, and SVM for the final prediction.

Rahaman et al. [22] proposed a framework based on a hybrid deep feature fusion approach for multiclass cervical cell classification. Features were extracted using four pre-trained CNNs, VGG16, VGG19, XceptionNet, and ResNet50, and then fed into a dense layer for classification.

Zhang et al. [23] employed a three-path collaborative design: two paths extracted deep features from temporal and frequency domains using VGG19, while the third path focused on manual feature extraction and selection. Final predictions were derived by analyzing the correlations among the three paths.

Liu et al. [24] introduced a DL framework for cervical cell classification that combines a CNN with a Vision Transformer (ViT) to capture both fine-grained and global image features. The CNN, based on Xception, produced

local features, while a small DeiT model acted as the ViT to extract global features. These features were then fused through a multi-layer network for final classification.

Hemalatha et al. [25] fused local and global features extracted from a pre-trained DenseNet201 and a Vision Transformer (ViT) enhanced with Shifted Patch Tokenization and Locality Self-Attention using a cervix feature fusion strategy. To reduce redundancy and improve classification performance, a fuzzy feature selection method was employed to select the most discriminative features from the fused feature vector before final classification.

Attallah [26] proposed a framework combining compact versions of MobileNet, DarkNet19, and ResNet18, designed with fewer layers and parameters to reduce classification complexity. Deep features were extracted from the last three layers of each CNN using transfer learning and then fused into a unified feature representation. Feature selection was applied to retain a reduced set of significant features, and the resulting important attributes were fed into various machine learning classifiers for final classification.

Deo et al. [27] proposed a Transformer-based model requiring minimal architectural assumptions regarding input data size. The model leverages a cross-attention technique to iteratively consolidate input data into a compact latent Transformer module, enabling it to handle large-scale inputs.

Qian et al. [28] introduced a novel ensemble approach using three Inception networks as base models, whose outputs are combined through a weighted voting scheme. Misclassified samples from the ensemble are used to create a new training set for a linear classifier, serving as a meta-learner for the final predictions. The method also incorporates a multi-level ensemble framework to further boost performance.

Liu et al. [29] employed Local Binary Patterns (LBP) to extract texture features from cervical cell images. Unlike standard LBP, which uses a fixed neighborhood radius and ignores spatial relationships, their method incorporates adaptive neighborhood radii along with a spatially adjacent histogram strategy to capture richer feature information.

Fekri Ershad [30] proposed a method for classifying cell images that prioritizes diagnostic accuracy while maintaining low computational complexity and robustness to rotations and grayscale variations. The technique extracts both textural and statistical features from the nucleus and cytoplasm. Feature extraction leverages a selected subset of Haralick descriptors, global significance measures, and time-series features. Classification is then performed using machine learning algorithms.

Chauhan and Singh [31] analyzed the performance of CNN models in classifying multi-class LBC whole-slide images (WSI). Their analysis focused on variations in channel depth within the convolution layers. Three CNN models were implemented, each featuring two convolution layers with varying channel depths and two pooling layers. The model with the greatest channel depth achieved the best performance.

Yaman and Tuncer [32] developed a deep feature extraction approach, called Exemplar Pyramid, for detecting cervical cancer in Pap-smear images. Their method employed transfer learning with the DarkNet architecture and applied Neighborhood Component Analysis to identify the most relevant features from the extensive set generated by the pyramid structure. The selected features are subsequently classified using an SVM.

Macancela [33] proposed a deep reinforcement learning approach to address pathologist shortages in cervical cancer screening. This method utilized LBC Pap smear images within a training environment. Agents were trained to navigate these images and identify cells through a reward–penalty system. A pre-trained neural network further enhanced cell classification for malignancy. This approach aimed to develop a fully automated Papanicolaou analysis system, ultimately reducing reliance on pathologists in underserved areas.

Haryanto et al. [34] proposed a CNN-based model for classifying cervical cell images, using the AlexNet architecture as the backbone. Initially, the network was implemented without any padding. Subsequent experiments introduced zero-pixel padding around the images to evaluate its effect on classification performance. The results demonstrated that incorporating padding into AlexNet led to a statistically significant improvement in accuracy.

Win et al. [35] first segmented nuclei using a shape-based iterative method and separated overlapping cytoplasm with a marker-controlled watershed approach. From the segmented regions, three key features were extracted, and RF was applied for feature selection. Finally, classification was performed using a bagging ensemble that combined five individual classifiers.

Manna et al. [36] proposed an ensemble-based cervical cancer classification framework for Pap smear images by integrating three pre-trained CNN architectures, namely Inception v3, Xception, and DenseNet169. Their method employed a fuzzy rank-based fusion strategy that incorporates nonlinear functions on the decision scores and explicitly accounts for the confidence of each base classifier, rather than relying on simple fusion schemes.

Pramanik et al. [3] applied three pre-trained CNNs, InceptionV3, MobileNetV2, and Inception ResNetV2, for cervical cancer classification, enhancing them with additional layers to capture data-specific features. They introduced a novel ensemble strategy to combine predictions from these models. For instances with multiple outputs, Euclidean, Manhattan, and Cosine distances were computed for each class using each model's optimal solutions, and these distances were then defuzzified via the product rule to generate the final classification.

Raza et al. [37] proposed a DL framework for cervical cancer classification that combines Neural Feature Extraction (NFE) with the AutoInt model. They adapted a pre-trained VGG16 by removing its top layer and adding global average pooling to function as the NFE module. The AutoInt model was then applied to capture complex feature interactions. Finally, the extracted features were classified using various machine learning algorithms, with KNN achieving the highest accuracy.

Sharma and Parvathi [38] proposed a hybrid DL architecture to balance precision and recall—an important challenge in medical image analysis. The framework integrates DenseNet201 for feature reuse and InceptionV3 for multi-scale analysis, fuses the extracted features, and applies PCA and t-distributed Stochastic Neighbor Embedding for dimensionality reduction. A fully connected neural network then performs the final classification.

Tan et al. [39] evaluated thirteen pre-trained CNN models and found that DenseNet-201 achieved the best performance. Their results demonstrated that these models could accurately classify cervical cancer subtypes without requiring manual segmentation or handcrafted feature extraction, highlighting the potential of DL to automate and streamline cervical cancer detection.

Singh et al. [40] extracted deep features using eight pre-trained CNN models. For hybrid feature combinations, PCA was employed to reduce dimensionality. Several classifiers—including SVM, KNN, Multi-Layer Perceptron (MLP), LR, RF, and Extreme Gradient Boosting (XGBoost)—as well as their ensemble variants were evaluated. The best performance was achieved by combining features extracted from ResNet101, DenseNet121, and DenseNet169 with a soft-voting ensemble classifier.

Kaur et al. [41] evaluated sixteen pre-trained CNN architectures for cervical cancer classification, including VGG16, VGG19, several ResNet variants, DenseNet models, MobileNet, XceptionNet, InceptionV3, and InceptionResNetV2. Their results indicated that VGG16 and ResNet50 achieved the highest classification accuracy.

Bilal et al. [42] proposed an ensemble-based DL framework that integrates three pre-trained CNNs—DenseNet169, MobileNetV2, and DenseNet201—as base learners. A grid search strategy is employed to optimally weight each base model, enabling the ensemble to improve overall classification accuracy.

Khowaja et al. [43] proposed a hybrid framework that integrates a transformer-based model with attention-enhanced CNNs to capture both global and local features. A ViT models the global spatial context, while a token-to-token module extracts fine-grained cellular details. Attention-enhanced ResNet101 and DenseNet169 further refine feature importance. The outputs are then fused using weighted voting, resulting in improved classification accuracy and reduced risk of misdiagnosis.

Wubineh et al. [44] proposed a DL framework for cervical cell segmentation and classification, employing an SE-DeepLabV3+ model with a dynamic atrous rate to enhance segmentation. For classification, the framework uses pre-trained ResNet50V2, VGG19, Xception, and VGG16 as frozen feature extractors. The globally pooled features from these models are adaptively weighted, concatenated, and refined through fully connected layers to improve robustness and classification accuracy.

Sharma et al. [45] proposed a hybrid DL approach for cervical cancer classification that combines the strengths of multiple pre-trained models, specifically ResNet50 and VGG19, with machine learning classifiers. Their method also incorporates fuzzy min–max neural networks and ensemble-based data augmentation to improve both accuracy and robustness.

3. Proposed Method

Using transfer learning, the proposed method extracts discriminative features from three pre-trained CNNs fine-tuned on preprocessed Pap smear images. These images are

first resized and augmented, then passed through the CNN models to generate feature vectors. The extracted features are concatenated to form a unified representation, which serves as input for various classifiers to categorize the images. **Error! Reference source not found.** illustrates the architecture of the proposed method, while the following sections provide a detailed explanation of the methodology.

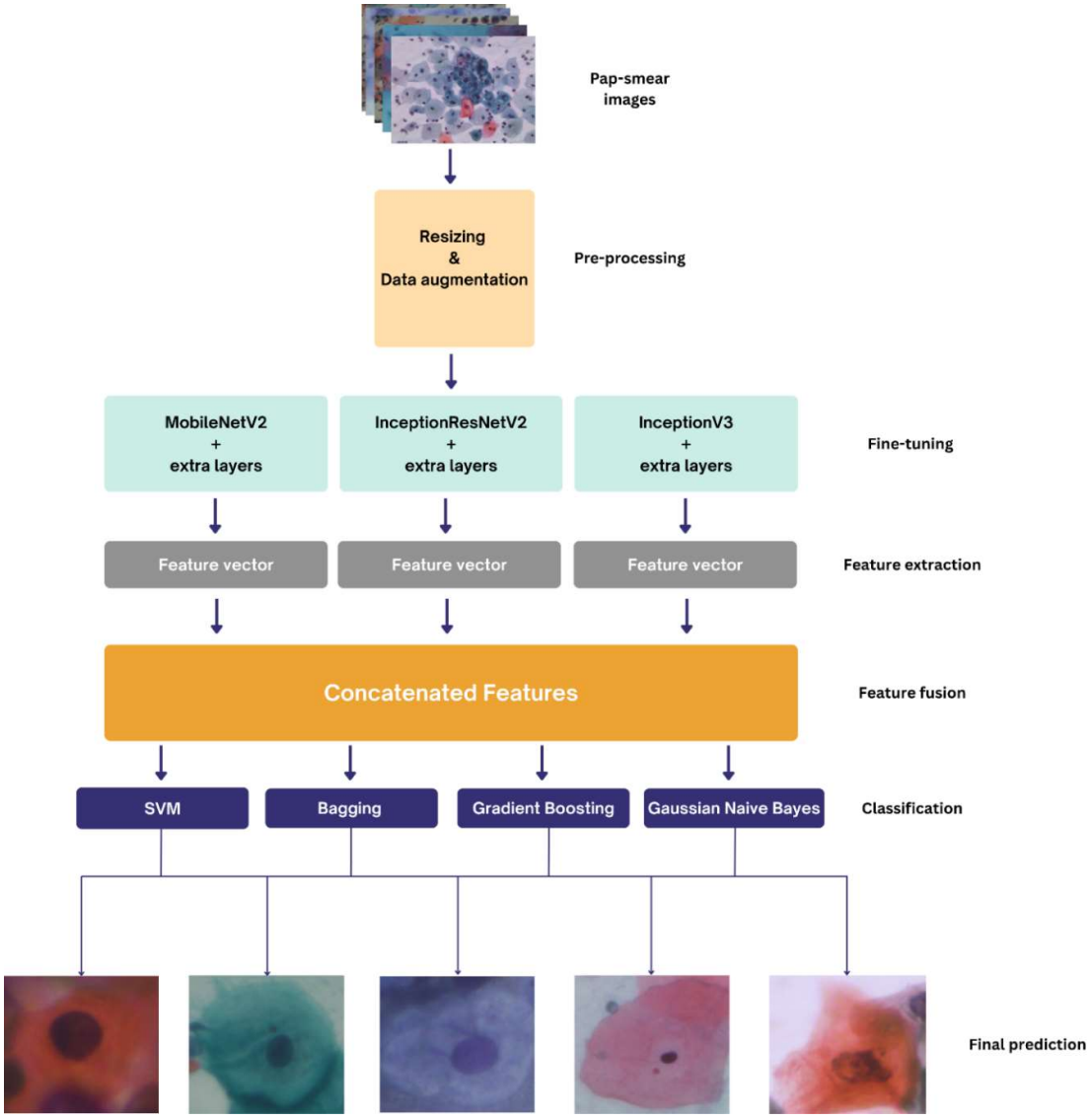

Figure 1. The overall schema of the proposed approach for cervical cancer detection

3.1. Data Pre-processing

Image preprocessing is an essential stage in digital image analysis, employing different techniques to improve input images and boost the performance of downstream tasks. This section explains the preprocessing methods applied in this study.

Image resizing involves adjusting the dimensions of images, which reduces the number of pixels and decreases the computational cost and training time for neural networks. This study uses three cervical Pap smear image datasets: SIPaKMeD, Mendeley LBC, and Herlev, all provided in BMP format. The image sizes in the SIPaKMeD dataset vary widely, ranging from 71 by 59 pixels to 490 by 474 pixels. Mendeley LBC images have a fixed size of 2048 by 1536 pixels, while Herlev images range between 43 by 77 pixels and 768 by 300 pixels. To maintain consistency and compatibility with the CNN models, all images were resized to 256 by 256 pixels using the OpenCV “resize” function.

Image augmentation increases the effective size of the training dataset by introducing random modifications to the original images. This technique helps the model generalize better by exposing it to a wider variety of input scenarios during training. In this paper, augmentation was performed using the "ImageDataGenerator" class from the Keras API, which applies random transformations on-the-fly during each training epoch. Table 1 summarizes the augmentation techniques and their corresponding parameters.

Table 1. Data augmentation techniques

Technique	Range
Rotation	45 (degrees)
Horizontal Shift	0.2
Vertical Shift	0.2

3.2. *Fine-tuning of pre-trained CNNs*

In this paper, the InceptionV3, InceptionResNetV2, and MobileNetV2 architectures are fine-tuned on Pap smear datasets. To enable this adaptation, additional layers are appended to each base model. These layers include a convolutional layer with 128 filters of size 3×3, followed by a 2×2 max-pooling layer, a flatten layer, a fully connected layer containing 100 neurons, and a final output layer with neurons corresponding to the number of classes. The output layer employs the SoftMax activation function for multi-class classification. Figure 2 illustrates these

additional layers, including the classification layer with five neurons corresponding to the SIPaKMeD dataset. The following subsection provides detailed descriptions of the selected base architectures.

InceptionV3, proposed by Szegedy et al. [46], is a deep CNN recognized for its high accuracy and computational efficiency in image classification. It extends the original Inception design using inception modules, which combine convolutional filters of multiple sizes, enabling the extraction of diverse spatial features from images. The architecture also incorporates strategies such as grid reduction and factorized convolutions to minimize parameter count and computational load compared to earlier Inception versions. Its robust performance has led to widespread use in computer vision tasks, including classification, object detection, and segmentation.

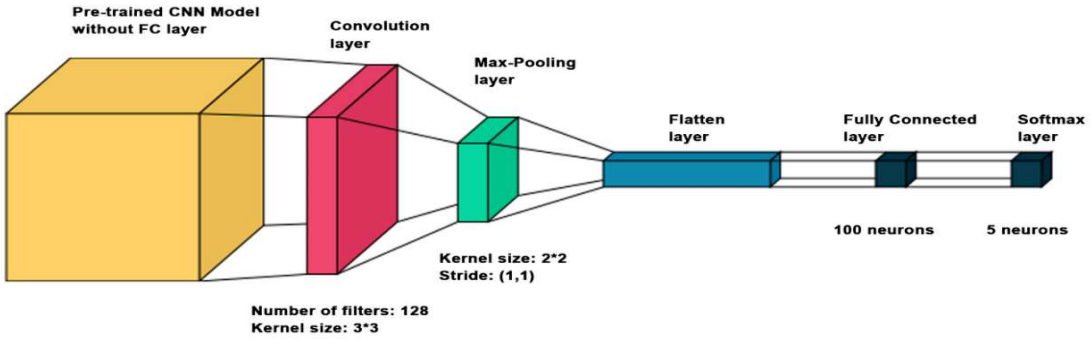

Figure 2. Additional layers integrated into the pre-trained CNN architectures

InceptionResNetV2, proposed by Szegedy et al. [11], is a CNN that integrates the advantages of Inception modules with residual connections. Like InceptionV3, it uses inception modules that combine convolutions of multiple sizes to capture varied spatial features in an image. Residual connections, inspired by ResNet architectures, help the network learn effectively from gradients in deep layers, mitigating the vanishing gradient problem. By merging these two concepts, InceptionResNetV2 offers improved accuracy compared to InceptionV3 while remaining computationally efficient. Its balanced performance makes it suitable for applications with both high accuracy and resource constraints, including image classification, object detection, and image segmentation.

MobileNetV2, proposed by Sandler et al. [47], is a lightweight CNN architecture designed for mobile and embedded devices with limited resources. Unlike traditional CNNs, which can be computationally expensive, MobileNetV2 prioritizes efficiency. It achieves this through innovative techniques like inverted residual blocks with bottleneck layers. These blocks handle information efficiently, lowering computational demands while preserving strong accuracy in image classification. This makes MobileNetV2 a valuable tool for deploying DL

models on devices where processing power and battery life are at a premium. Additionally, its versatility allows it to be adapted for various tasks beyond image classification, making it a powerful and efficient choice for computer vision applications on mobile platforms.

3.3. *Feature Extraction*

Transfer learning leverages pre-trained CNNs that have been trained on large datasets. These networks excel at extracting comprehensive and deep features from the data. These extracted features enable one to apply the network quickly and accurately to various problems. One of the advantages of employing CNNs is their ability to capture conceptual and abstract features from diverse data, which can significantly assist new models in addressing specific problems. To harness these valuable features, one can trim the layers of the original model and utilize the resulting features as input for a new model or new output layers. In the context of the paper, features are specifically extracted from the fully connected layer immediately preceding the classification layer.

3.4. *Feature Fusion*

Feature fusion is a widely used strategy in DL and medical image analysis to combine complementary information extracted from multiple models and improve classification performance [48]. Among various fusion strategies, feature-level concatenation offers a simple yet effective mechanism for preserving discriminative information from each feature source without introducing additional optimization complexity [49].

In this study, feature vectors extracted from three fine-tuned CNN architectures—InceptionV3, InceptionResNetV2, and MobileNetV2—are concatenated to form a unified and extended representation. Unlike conventional fusion approaches that merely increase feature dimensionality, the proposed fusion strategy is designed to exploit architectural diversity and complementary feature characteristics.

Specifically, InceptionV3 captures multi-scale spatial patterns through parallel convolutional filters, enabling effective representation of global and local cellular structures. InceptionResNetV2 enhances deep discriminative representations via residual learning, improving gradient flow and feature abstraction. MobileNetV2 contributes lightweight yet fine-grained local features optimized for computational efficiency. By extracting features from the fully connected layers of these architecturally diverse networks and concatenating them, the proposed fusion framework integrates heterogeneous and complementary information rather than redundant features.

Unlike prior fusion-based frameworks that rely on dimensionality reduction (e.g., PCA), fuzzy fusion rules, or heuristic weighting schemes, the proposed method preserves the full discriminative capacity of fused deep features and aligns the fusion strategy with a variance-reducing ensemble classifier. This design choice enables robust representation learning while avoiding handcrafted fusion rules or additional optimization stages, making the approach particularly suitable for medical image classification tasks where subtle morphological variations must be captured under limited data conditions.

3.5. *Classification*

Following feature fusion, the resulting high-dimensional feature vectors are classified using several machine learning algorithms, including SVM, Gaussian Naive Bayes, Gradient Boosting, RF, and Bagging with RF as the base estimator. This comprehensive evaluation allows systematic analysis of how different classification strategies interact with fused deep features.

Bagging with RF base learners is selected as the final classification strategy due to its robustness to high-dimensional feature spaces and its ability to reduce variance through bootstrap aggregation. While RF itself benefits from ensemble learning via feature randomness, the Bagging framework further enhances stability by training multiple RF models on bootstrapped subsets of the data and aggregating their predictions. This variance-reduction property is particularly important when dealing with concatenated deep features extracted from multiple CNNs and relatively small medical datasets.

In contrast to end-to-end ensemble DL approaches, which often suffer from overfitting and unstable training on limited medical data, the bagging ensemble of RF learners provides a computationally efficient and reliable alternative for classifying fused deep features. This approach avoids additional parameter tuning at the deep network level while maintaining strong generalization performance. Experimental results (Section 4.6) demonstrate that the Bagging with RF learners consistently outperforms both single classifiers and alternative ensemble methods,

justifying its selection as the final decision-making component of the proposed framework.

4. Experimental results

4.1. Training Environment and Resources

Model development and training in this study were conducted using Google Colab, leveraging an NVIDIA A100 GPU with 50 GB of memory, alongside a system equipped with 64 GB of RAM to facilitate efficient training and testing of the proposed network.

4.2. Dataset Description

To evaluate the proposed method, we employed three publicly available datasets containing images of cervical cells: SIPaKMeD [14], Mendeley LBC [15], and Herlev [16] datasets.

The SIPaKMeD dataset contains 4,049 labeled cervical cell images. Expert cytopathologists classified the cells into five groups according to their morphology and appearance: two types of normal cells (superficial-intermediate and parabasal), two types of abnormal cells (koilocytotic and dyskeratotic), and one type of benign cells (metaplastic). The images were captured using an optical microscope equipped with a Charge-Coupled Device (CCD) camera [14]. In this paper, the SIPaKMeD dataset is used with its original five-class labeling scheme. The distribution of cells across these categories is shown in Table 2.

The Mendeley LBC dataset consists of 963 liquid-based cytology images, divided into four categories based on the Bethesda System (TBS): Negative for Intraepithelial Lesion or Malignancy (NILM), Low-grade Intraepithelial Lesions (LSIL), High-grade Intraepithelial Lesions (HSIL), and Squamous Cell Carcinoma (SCC), reflecting different stages of cervical cancer. These images, showcasing pre-cancerous and cancerous lesions, were captured at 40× magnification using a Leica ICC50 HD microscope. They were collected from 460 patients receiving care at the Obstetrics and Gynecology (O&G) department of a public hospital, with informed consent [15]. The distribution of cells across the categories is shown in Table 3.

The Herlev dataset contains 917 samples divided unevenly into seven classes, which can further be grouped into normal and abnormal categories (Table 4). In this paper, only the binary classification of normal versus abnormal is considered. Each image is represented by 20 features extracted from single-cell images captured at Herlev University Hospital using a digital microscope and camera [16].

Table 2. Description of the SIPaKMeD dataset

Index	Category	Cell type	Number of cells
4	Normal	Superficial-Intermediate	813
3	Normal	Parabasal	787
2	Benign	Metaplastic	793
1	Abnormal	Koilocytotic	825
0	Abnormal	Dyskeratotic	813

Table 3. Description of the Mendeley LBC dataset

Index	Cell type	Number of cells
3	NILM	613
2	LSIL	113
1	HSIL	163
0	SCC	74

Table 4. Description of the Herlev dataset

Index	Category	Cell type	Number of cells
1	Normal	Superficial squamous epithelial	74
1	Normal	Intermediate squamous epithelial	70
1	Normal	Columnar epithelial	98
0	Abnormal	Mild squamous non-keratinizing dysplasia	182
0	Abnormal	Moderate squamous non-keratinizing dysplasia	146
0	Abnormal	Severe squamous non-keratinizing dysplasia	197
0	Abnormal	Squamous cell carcinoma in situ intermediate	150

4.3. Evaluation Methodology

This study employed 5-fold cross-validation to evaluate the proposed method on the SIPaKMeD dataset, providing a comprehensive assessment by exposing the model to diverse subsets of the data and yielding a more balanced measure of performance [50]. For the Mendeley LBC and Herlev datasets, the method was trained on 80% of the images, with the remaining 20% used for validation.

4.4. Evaluation Metrics

The performance of the proposed model was assessed using the following metrics:

Accuracy: This metric represents the proportion of correctly predicted samples relative to the total number of samples, providing an overall indication of the model's effectiveness [51]. While accuracy is widely used, particularly for balanced datasets, it can be misleading when class distributions are imbalanced. In such cases, additional metrics like precision, recall, and $F1$-score are necessary for a more comprehensive evaluation. Accuracy is computed as shown in (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

In this formula, TP (True Positive) denotes the count of samples correctly identified as part of the positive class, while TN (True Negative) represents samples correctly classified as part of the negative class. FP (False Positive) refers to samples incorrectly labeled as positive when they are actually negative, and FN (False Negative) corresponds to samples wrongly classified as negative when they truly belong to the positive class.

Precision: This metric quantifies the fraction of samples predicted as positive that truly belong to the positive class [51]. Essentially, it reflects the accuracy of positive predictions. Precision is particularly relevant in contexts where false positives have significant consequences. The calculation is shown in (2).

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall: This metric represents the fraction of actual positive samples that the model correctly identifies as positive [51]. In other words, it measures how many true positive instances are successfully detected. Recall is particularly critical in scenarios where failing to identify positive cases (false negatives) can have serious consequences. Its computation is given in (3).

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$F1$-score: This metric combines precision and recall to provide a single measure of a model's performance in correctly identifying positive samples [51]. It is especially valuable when dealing with imbalanced datasets, where focusing solely on precision or recall may not give a complete picture. The $F1$-score is computed using (4).

$$F1 - score = \frac{2 * (Precision * Recall)}{Precision + Recall} \quad (4)$$

4.5. *Hyperparameter Configuration*

To identify the optimal hyperparameters for fine-tuning the CNN models and extracting meaningful features, a comprehensive grid search was performed using the SIPaKMeD dataset. Each CNN architecture was extended by incorporating several custom layers to better adapt the models to the dataset's characteristics. The grid search evaluated learning rates of 1e-3, 1e-4, 1e-5, and 1e-6, along with batch sizes of 8, 16, and 32. For each combination, 80% of the images were randomly selected for training, while the remaining 20% were used for validation. The results showed that a learning rate of 1e-4 and a batch size of 16 yielded the best validation performance. Model training was conducted using the Adam optimizer and the categorical cross-entropy loss function over 70 epochs. Following training, feature vectors were extracted from the fully connected layer immediately preceding the classification layer.

To classify the fused features, several classifiers were evaluated, including SVM (with linear, polynomial, and Radial Basis Function (RBF) kernels), RF, Bagging with RF as base learners, Gaussian Naive Bayes, and Gradient Boosting. The SVM used a polynomial kernel with a cost parameter $C = 1$ and a polynomial degree of 4. The Gradient Boosting classifier was configured with 100 decision trees, a learning rate of 0.1, and a maximum tree depth of 3. The RF classifier comprised 100 decision trees, while the Bagging ensemble consisted of 10 RF models, each containing 100 decision trees.

4.6. *Results*

Table 5. Classification results using fused features with various classifiers on the SIPaKMeD dataset

, Table 6. Classification results using fused features with various classifiers on the Mendeley LBC dataset, and Table 7. Classification results using fused features with various classifiers on the Herlev dataset summarize the classification performance of fused deep features across the SIPaKMeD, Mendeley LBC, and Herlev datasets, respectively, using various classifiers. The evaluated methods include SVM with linear, polynomial, and RBF kernels, Gaussian Naive Bayes, Gradient Boosting, Random Forest (RF), and the Bagging ensemble with RF as base learners (Bagging-RF). Among these, Gradient Boosting, RF, and Bagging-RF are ensemble approaches.

On the SIPaKMeD dataset (Table 5), Bagging-RF achieves the highest performance, with 97.25% accuracy, 97.26% precision, 97.28% recall, and an $F1$-score of 97.26%. For the Mendeley LBC dataset (Table 6), Bagging-RF attains performance that is superior or comparable to other classifiers, achieving 99.47% accuracy, 99.10% precision, 98.33% recall, and 98.68% $F1$-score. On the Herlev dataset (Table 7), Bagging-RF maintains strong performance, with 96.72% accuracy, 96.96% precision, 93.68% recall, and 95.20% $F1$-score. While Gradient Boosting performs competitively and slightly surpasses Bagging-RF in precision, it is outperformed in terms of accuracy, recall, and $F1$-score.

Table 5. Classification results using fused features with various classifiers on the SIPaKMeD dataset

Classifier	Accuracy	Precision	Recall	$F1$-score
SVM-linear	97.15 ± 0.46	97.18 ± 0.49	97.19 ± 0.41	97.17 ± 0.44
SVM-RBF	97.01 ± 0.44	97.02 ± 0.46	97.04 ± 0.39	97.02 ± 0.42
SVM-Polynomial	97.06 ± 0.46	97.09 ± 0.44	97.07 ± 0.44	97.07 ± 0.44
Gaussian Naive Bayes	96.02 ± 0.95	96.17 ± 0.91	96.04 ± 0.93	96.07 ± 0.93
Gradient Boosting	96.71 ± 0.47	96.72 ± 0.48	96.73 ± 0.42	96.72 ± 0.46
RF	97.05 ± 0.45	97.08 ± 0.46	97.10 ± 0.41	97.09 ± 0.43
Bagging-RF	97.25 ± 0.42	97.26 ± 0.44	97.28 ± 0.37	97.26 ± 0.4

Table 6. Classification results using fused features with various classifiers on the Mendeley LBC dataset

Classifier	Accuracy	Precision	Recall	$F1$-score
SVM-linear	99.33	99.02	98.22	98.54
SVM-RBF	99.45	99.05	98.31	98.66
SVM-Polynomial	99.46	99.08	98.33	98.68
Gaussian Naive Bayes	98.95	98.27	98.13	98.14
Gradient Boosting	98.95	98.27	96.66	97.32
RF	99.42	99.07	98.30	98.65
Bagging-RF	99.47	99.10	98.33	98.68

Table 7. Classification results using fused features with various classifiers on the Herlev dataset

Classifier	Accuracy	Precision	Recall	$F1$-score
SVM-linear	95.08	96.92	93.47	94.43
SVM-RBF	95.08	96.67	93.52	94.19
SVM-Polynomial	95.08	96.98	93.36	94.45
Gaussian Naive Bayes	95.62	96.42	93.01	94.76
Gradient Boosting	96.02	97.04	93.57	95.09
RF	96.17	96.25	93.35	94.98
Bagging-RF	96.72	96.96	93.68	95.20

These results effectively serve as an ablation study to evaluate the impact of classification strategy when fused deep features are used with and without an ensemble framework. Although several classifiers achieve competitive performance, the Bagging-RF ensemble consistently provides the most stable and accurate results, demonstrating its suitability for high-dimensional fused features. The comparison between RF and Bagging-RF indicates that while RF alone benefits from the fused features, Bagging further reduces variance and enhances stability, which is crucial for medical imaging datasets that are relatively small and often imbalanced.

Although advanced end-to-end ensemble deep learning approaches exist, they are prone to overfitting and unstable training on limited medical datasets. In contrast, RF provides inherent robustness to noise and high-dimensional input, and the Bagging strategy further improves generalization by aggregating multiple learners trained on bootstrap samples. These characteristics justify the selection of Bagging-RF as the final decision-making component of the proposed framework.

In addition to its robustness, the use of Random Forest within the bagging framework offers a degree of transparency compared to fully end-to-end deep neural classifiers. Unlike black-box models that directly output predictions from deep representations, Random Forest is composed of decision trees that perform feature-wise splits, enabling inspection of feature importance measures and decision structure. In practice, standard impurity-based feature importance scores can be extracted from the trained RF model, allowing identification of the most influential deep feature dimensions. Although the fused deep features are high-dimensional and not directly interpretable at the

pixel level, the tree-based architecture allows qualitative assessment of which feature subsets contribute more prominently to classification decisions. This provides limited but meaningful insight into model behavior while maintaining strong generalization performance.

Furthermore, **Error! Reference source not found.** presents the confusion matrices resulting from the classification of fused features using the Bagging-RF on the SIPaKMeD, Mendeley LBC, and Herlev datasets. The indices in each matrix correspond to the class labels defined in shown in Table 2.

, **Error! Reference source not found.**, and **Error! Reference source not found.**. A visual inspection of the main diagonal elements reveals that the proposed method achieves strong classification performance, with most instances correctly identified across all classes. In the SIPaKMeD dataset, the highest number of errors occurs for true class label 1, with four instances misclassified as class 0, five as class 2, one as class 3, and two as class 4. These results indicate that the model achieves high precision and recall across all classes, demonstrating robust class-wise performance. For the Mendeley LBC dataset, the classifier demonstrates near-perfect performance: no misclassifications occurred for class labels 0, 1, and 3, and only a single instance from class 2 was misclassified. In the Herlev dataset, which involves binary classification, the model shows excellent class separation, with only six misclassified samples out of 183. This reflects a high level of confidence in the model's binary classification capability.

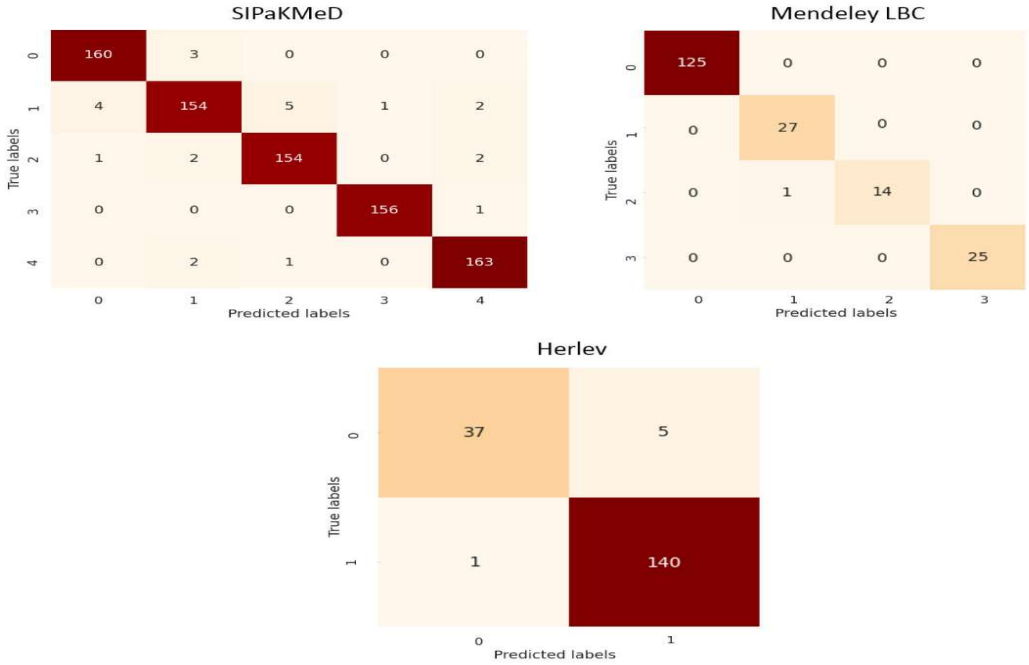

Figure 3. Confusion matrices resulting from the proposed method with the Bagging classifier on the SIPaKMeD, Mendeley LBC, and Herlev datasets. The indices in the matrices correspond to the classes listed in Tables 2–4

The classification performance of the fused deep features using the Bagging-RF classifier, referred to as the proposed method, is compared with that of three individual CNN backbones—InceptionV3, Inception-ResNetV2, and MobileNetV2. The results for the SIPaKMeD, Mendeley LBC, and Herlev datasets are reported in Tables 8, 9, and 10, respectively.

On the SIPaKMeD dataset (Table 8), the proposed method achieves the best performance across all evaluation metrics, attaining an accuracy of 97.25% and an $F1$-score of 97.26%. Among the individual CNNs, Inception-ResNetV2 performs strongly (Accuracy: 96.81%, $F1$-score: 96.81%), followed by InceptionV3 (Accuracy: 94.07%, $F1$-score: 94.07%), while MobileNetV2 yields comparatively lower results (Accuracy: 90.49%, $F1$-score: 90.45%). This trend suggests that lightweight architectures such as MobileNetV2 may have limited capacity to capture the complex morphological and textural variations inherent in

cervical cytology images. In contrast, deeper architectures provide richer hierarchical representations, yet still remain inferior to the fused multi-CNN representation. These results highlight the benefit of combining complementary features extracted by multiple CNNs and leveraging Bagging-RF to enhance robustness and discrimination.

On the Mendeley LBC dataset (Table 9), it achieves an accuracy of 99.47% and an $F1$-score of 98.68%, clearly outperforming all baseline models. InceptionV3 shows competitive performance (Accuracy: 97.91%, $F1$-score: 94.95%), whereas MobileNetV2 exhibits a substantial performance degradation ($F1$-score: 19.71), suggesting reduced generalization stability under this dataset configuration.

The training and validation loss curves shown in Figure 8 indicate that, under the corresponding experimental setting, MobileNetV2 may exhibit a tendency toward overfitting on the Mendeley LBC dataset, as reflected by the divergence between training and validation losses during later epochs. Although this observation is derived from the learning dynamics analysis rather than directly from the tabulated results, it suggests reduced robustness of lightweight architectures under certain data conditions. Several factors may contribute to this behavior. First, domain shift between training and testing distributions — including variations in staining characteristics, imaging protocols, and cellular morphology — may disproportionately affect models with reduced representational capacity. Second, MobileNetV2's depthwise separable convolutions, although computationally efficient, may be more sensitive to fine-grained structural variations and resolution changes that are critical in cytology analysis. Third, potential class imbalance may further exacerbate performance instability, particularly for models with limited capacity to model minority-class features effectively.

On the Herlev dataset (Table 10), Inception-ResNetV2 slightly outperforms the proposed method in some metrics (Accuracy: 97.26% vs. 96.72%, $F1$-score: 96.03% vs. 95.20%). Nevertheless, the proposed approach maintains consistently high precision and recall and substantially outperforms InceptionV3 and MobileNetV2, particularly in terms of recall. Notably, MobileNetV2 again demonstrates unstable behavior (Recall: 0.50%), indicating difficulty in capturing discriminative features for certain classes within this dataset. In contrast to the tendencies observed for Mendeley, the loss curves for Herlev (Figure 9) do not show a pronounced divergence between training and validation losses, suggesting that the performance limitation here is more likely related to representational capacity rather than overfitting. Although a single deep CNN may occasionally achieve marginally higher performance on a specific dataset, the proposed multi-CNN fusion with Bagging-RF provides more stable and balanced results across varying data distributions.

In summary, while individual CNN backbones can perform competitively under certain conditions, lightweight architectures such as MobileNetV2 appear more vulnerable to distributional variations, resolution sensitivity, and class imbalance effects in cervical cytology datasets. In contrast, the proposed framework consistently delivers high and well-balanced classification performance across all three datasets, indicating improved stability and strong empirical robustness for cervical cell classification.

Table 8. Results achieved using the proposed method and different CNNs on the SIPaKMeD dataset

Model	Accuracy	Precision	Recall	$F1$-score
Inception-ResNetV2	96.81 ± 0.56	96.83 ± 0.6	96.85 ± 0.48	96.81 ± 0.54
InceptionV3	94.07 ± 1.64	94.28 ± 1.64	94.12 ± 1.58	94.07 ± 1.63
MobileNet-V2	90.49 ± 1.42	90.75 ± 1.4	90.49 ± 1.48	90.45 ± 1.43
Proposed method	97.25 ± 0.42	97.26 ± 0.44	97.28 ± 0.37	97.26 ± 0.4

Table 9. Results achieved using the proposed method and different CNNs on the Mendeley LBC dataset

Model	Accuracy	Precision	Recall	$F1$-score
Inception-ResNetV2	91.66	87.30	85.18	81.04
InceptionV3	97.91	94.58	95.55	94.95
MobileNetV2	65.10	16.27	2.50	19.71
Proposed method	99.47	99.10	98.33	98.68

Table 10. Results achieved using the proposed method and different CNNs on the Herlev dataset

Model	Accuracy	Precision	Recall	$F1$-score
Inception-ResNetV2	97.26	97.32	94.88	96.03
InceptionV3	92.89	91.51	87.86	89.50
MobileNetV2	77.04	38.52	0.50	43.51

Proposed method	96.72	96.96	93.68	95.20

Figures 4, 5, and 6 compare the classification accuracy of the proposed method with existing approaches, including several hybrid methods described in Section 2, on the SIPaKMeD, Mendeley LBC, and Herlev datasets, respectively. On the SIPaKMeD dataset (Figure 4), the proposed method achieves the highest accuracy of 97.25%, outperforming all competing approaches. Among existing methods, the hybrid multi-CNN ensemble framework proposed by Singh et al. [40] achieves 96.67% accuracy by extracting features from multiple pre-trained CNNs, applying PCA for dimensionality reduction, and employing an ensemble of classifiers for final decision-making. Another hybrid approach introduced by Pramanik et al. [3] reports an accuracy of 96.47% by fusing features from three pre-trained CNNs using a fuzzy logic–based ensemble strategy. Similarly, the hybrid method proposed by Hemalatha et al. [25] combines features extracted from DenseNet201 and a ViT, followed by fuzzy feature selection, achieving an accuracy of 96.13%. Moreover, the hybrid ensemble approach introduced by Manna et al. [36] attains 95.43% accuracy by integrating the outputs of InceptionV3, Xception, and DenseNet169 through a fuzzy rank-based fusion strategy applied to classifier decision scores. Non-hybrid methods, including those by Yaman and Tuncer [32], Win et al. [35], and Haryanto et al. [34], achieve lower accuracies of 94.97%, 94.09%, and 87.32%, respectively.

As illustrated in Figure 5, the proposed method achieves an accuracy of 99.47% on the Mendeley LBC dataset, outperforming most existing approaches. Its performance is comparable to the method proposed by Yaman and Tuncer [32], which employs CNN-based feature extraction followed by Neighborhood Component Analysis for feature selection and classification using an SVM. Although effective, this approach relies on a single CNN backbone and does not incorporate multi-model fusion. Among hybrid approaches, the decision-level ensemble framework proposed by Manna et al. [36] also demonstrates competitive performance, achieving an accuracy of 99.23% by integrating the outputs of three pre-trained CNN models through a fuzzy rank-based fusion strategy applied to classifier decision scores. Similarly, the hybrid framework proposed by Khowaja et al. [43] integrates ViT with attention-enhanced CNNs and employs weighted voting at the decision level, reporting an accuracy of 98.44%. In addition, the hybrid multi-CNN ensemble approach introduced by Bilal et al. [42] combines DenseNet169, MobileNetV2, and DenseNet201 using grid-search-optimized weighting and achieves an accuracy of 97.94%. Furthermore, the proposed method surpasses the non-hybrid approach of Chauhan and Singh [31], which achieves 96.89% accuracy, by a margin exceeding 2.5%. The weakest performance is observed in the method proposed by Macancela et al. [33], which reports 91% accuracy—more than 8% lower than that of the proposed framework.

As illustrated in Figure 6, the proposed method achieves the highest accuracy of 96.72% on the Herlev dataset, outperforming all other compared approaches. Among existing methods, the hybrid approach proposed by Sharma et al. [45], which extracts features from ResNet50 and VGG19, ranks second with an accuracy of 95.43%. The hybrid framework proposed by Nguyen et al. [17] follows as the third-best method, achieving an accuracy of 92.63% by extracting features from three pre-trained CNN models, concatenating them, and classifying the fused representation through fully connected layers. Another hybrid multi-CNN feature fusion approach introduced by Wubineh et al. [44] employs four pre-trained CNNs as feature extractors. Their globally pooled features are adaptively weighted, concatenated, and refined through fully connected layers, achieving an accuracy of 90%. Non-hybrid methods proposed by Fekri Ershad [30], Liu et al. [29], and Bhatt et al. [20] achieve accuracies of 90.16%, 88.03%, and 78.14%, respectively.

Overall, according to the comparative results shown in Figures 4–6, although hybrid methods demonstrate strong performance, they consistently fall short of the accuracy achieved by the proposed approach. The superior performance of the proposed framework arises from the effective feature-level fusion of complementary representations extracted from three diverse pre-trained CNN architectures, combined with the Bagging-RF ensemble, which together yield a robust and highly discriminative model without relying on complex fuzzy inference or extensive feature selection.

Figure 7 illustrates the training and validation loss curves for each model on the fifth fold of the SIPaKMeD dataset, while Figures 8 and 9 present the corresponding loss variations on the Mendeley LBC and Herlev datasets, respectively. Overall, InceptionV3 and InceptionResNetV2 exhibit stable convergence with a small gap between training and validation losses, indicating strong

generalization capability. As shown in Figure 7, MobileNetV2 initially experiences convergence instability during the early epochs on the SIPaKMeD dataset, possibly due to entrapment in a local minimum; however, with continued training and learning rate adjustments, it converges to an optimal solution without overfitting. Moreover, Figure 8 shows that MobileNetV2 tends to overfit on the Mendeley LBC dataset, as evidenced by the growing gap between training and validation losses.

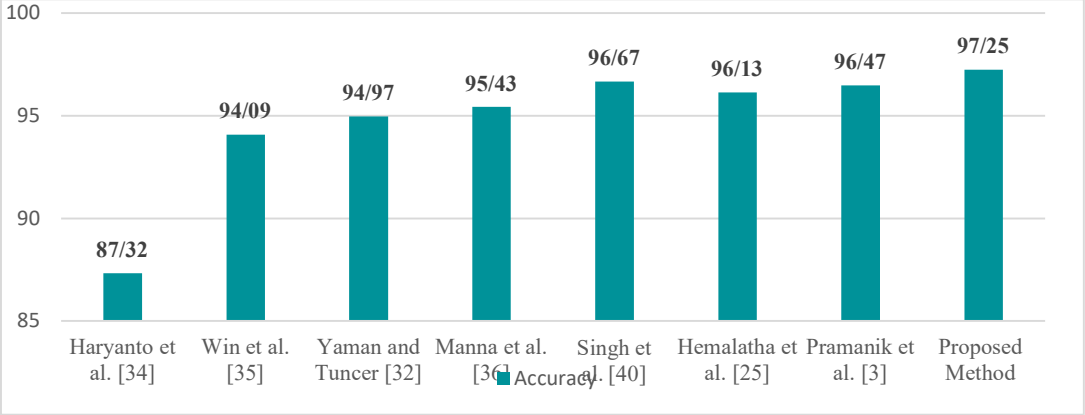

Figure 4. Accuracy comparison with some state-of-the-art approaches on the SIPaKMeD dataset

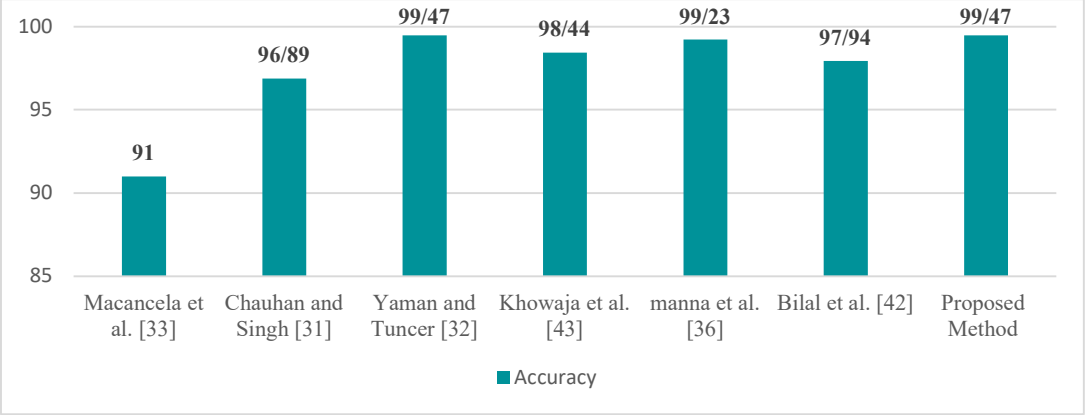

Figure 5. Accuracy comparison with some state-of-the-art approaches on the mendeley lbc dataset

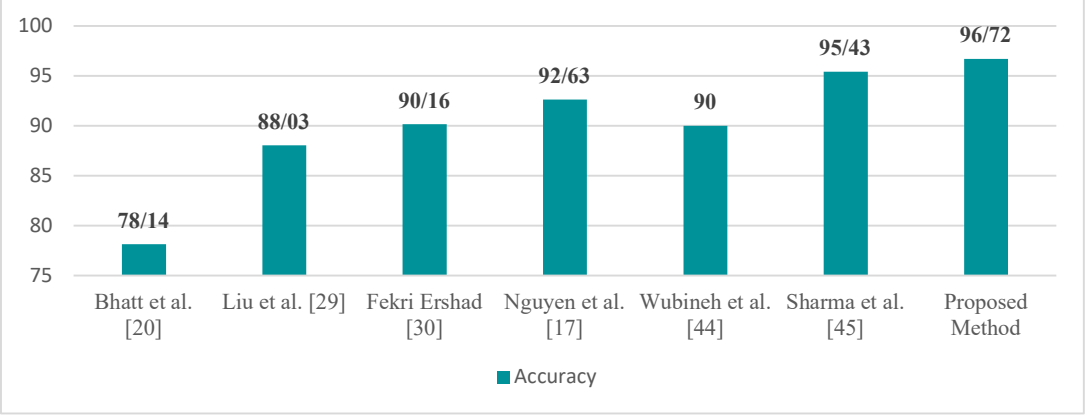

Figure 6. Accuracy comparison with some state-of-the-art approaches on the Herlev dataset

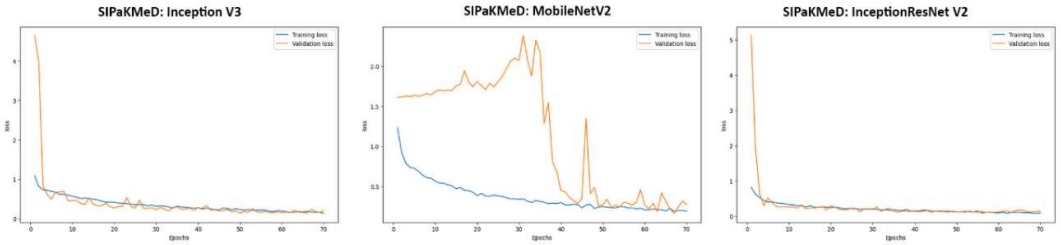

Figure 7. Loss variation during the convergence process corresponding to the fifth fold on the SIPaKMeD dataset

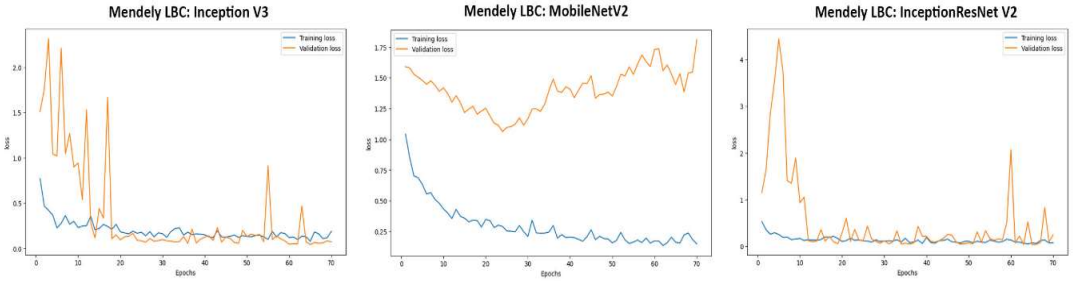

Figure 8. Loss variation during the convergence process on the Mendeley LBC dataset

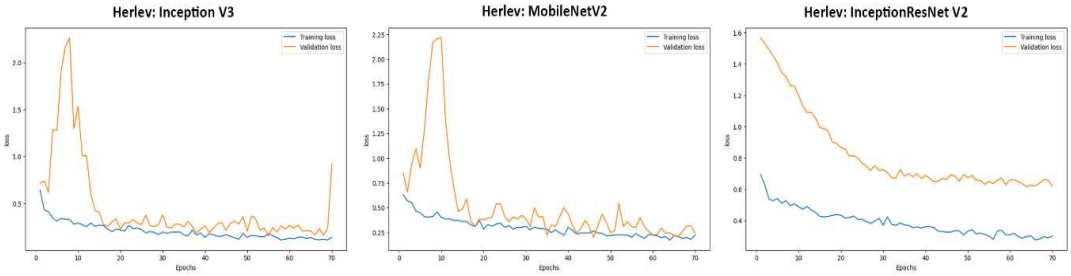

Figure 9. Loss variation during the convergence process on the Herlev dataset

5. Conclusion and Future Suggestions

This study presented a robust cervical cancer classification framework based on multi-CNN feature fusion and ensemble learning. Pre-trained convolutional neural networks—MobileNetV2, InceptionResNetV2, and InceptionV3—were employed as feature extractors using transfer learning, enabling the model to capture complementary representations of cervical cytology images. The extracted deep features were concatenated into a unified representation and evaluated using multiple classification strategies, including SVM, RF, Gradient Boosting, Gaussian Naive Bayes, and Bagging with RF as base learners. Comprehensive experiments conducted on three benchmark Pap smear datasets—SIPaKMeD, Mendeley LBC, and Herlev—demonstrated that the proposed multi-CNN fusion combined with Bagging–RF consistently achieves superior and more stable performance compared to individual CNN models and alternative classifiers. Ablation analyses confirmed that both the feature fusion strategy and the bagging-based ensemble contribute significantly to performance gains by improving generalization and reducing variance, particularly in small

and imbalanced medical datasets. These findings highlight the suitability of the proposed framework as a reliable decision-support tool for automated cervical cancer screening.

Despite the strong performance achieved, several directions remain for future research. Evaluating the proposed framework on larger and more diverse real-world clinical datasets would further validate its robustness and generalizability. Incorporating explainable artificial intelligence techniques, such as attention visualization or feature attribution methods, could enhance model interpretability and increase clinical trust. In addition, integrating multimodal information—such as patient metadata or cytological context—may further improve diagnostic accuracy. Future work may also explore lightweight model optimization for real-time deployment on resource-constrained devices, as well as semi-supervised or self-supervised learning strategies to reduce reliance on extensive labeled data. These extensions would support broader clinical adoption and strengthen the practical impact of the proposed approach.

Authors' Contributions

All authors equally contributed to this study.

Declaration

The authors affirm that all scientific content and conclusions presented in this article are their own. A language model (ChatGPT) was used exclusively to refine the English grammar and improve the clarity of writing. The model was not used for generating ideas, interpreting results, or performing data analysis.

Transparency Statement

The SIPaKMeD dataset used in this paper is available at the TMIYP (Department of Computer and Informatics Engineering Polytechnic School - University of Ioannina) and can be accessed at https://www.cse.uoi.gr/~marina/sipakmed.html. The Herlev dataset used in this paper is available at MDE-Lab (Management and Decision Engineering Laboratory - University of the Aegean) and can be accessed at https://mde-lab.aegean.gr/index.php/downloads/. The Mendely LBC dataset used in this paper is available at Mendeley Data repositories and can be accessed at https://data.mendeley.com/datasets/zddtpgzv63/4.

Acknowledgments

We would like to express our gratitude to all individuals helped us to do the project.

Declaration of Interest

The authors declare that they have no conflict of interest. The authors also declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

According to the authors, this article has no financial support.

Ethical Considerations

Not applicable.

References

[1] C. A. Burmeister and et al., "Cervical cancer therapies: Current challenges and future perspectives," *Tumour Virus Res,* vol. 13, p. 200238, 2022, doi: 10.1016/j.tvr.2022.200238.

[2] M. Karimi-Zarchi, L. Zanbagh, A. Shafii, S. Taghipour-Zahir, S. Teimoori, and P. Yazdian-Anari, "Comparison of pap smear and colposcopy in screening for cervical cancer in patients with secondary immunodeficiency," *Electron Physician,* vol. 7, no. 7, p. 1542, 2015, doi: 10.19082/1542.

[3] R. Pramanik, M. Biswas, S. Sen, L. A. de Souza Júnior, J. P. Papa, and R. Sarkar, "A fuzzy distance-based ensemble of deep models for cervical cancer detection," *Comput Methods Programs Biomed,* vol. 219, p. 106776, 2022, doi: 10.1016/j.cmpb.2022.106776.

[4] M. M. Ali and et al., "Machine learning-based statistical analysis for early stage detection of cervical cancer," *Comput Biol Med,* vol. 139, p. 104985, 2021, doi: 10.1016/j.compbiomed.2021.104985.

[5] M. Kaushik and et al., "Cytokine gene variants and socio-demographic characteristics as predictors of cervical cancer: A machine learning approach," *Comput Biol Med,* vol. 134, p. 104559, 2021, doi: 10.1016/j.compbiomed.2021.104559.

[6] M. Lalasa and J. Thomas, "A Review of Deep Learning Methods in Cervical Cancer Detection," 2022: Springer, pp. 624-633, doi: 10.1007/978-3-031-27524-1_60.

[7] R. Lozano, "Comparison of computer-assisted and manual screening of cervical cytology," *Gynecol Oncol,* vol. 104, no. 1, pp. 134-138, 2007, doi: 10.1016/j.ygyno.2006.07.025.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," 2012, vol. 25. [Online]. Available: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[9] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Trans Knowl Data Eng,* vol. 22, no. 10, pp. 1345-1359, 2010, doi: 10.1109/TKDE.2009.191.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.
[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," 2016, pp. 2818-2826, doi: 10.1109/CVPR.2016.308.
[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: https://arxiv.org/abs/1409.1556
[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2017, pp. 4700-4708, doi: 10.1109/CVPR.2017.243.
[14] M. Plissiti, P. Dimitrakopoulos, G. Sfikas, C. Nikou, O. Krikoni, and A. Charchanti, "Sipakmed: A New Dataset for Feature and Image Based Classification of Normal and Pathological Cervical Cells in Pap Smear Images," 2018, doi: 10.1109/ICIP.2018.8451588.
[15] E. Hussain, L. B. Mahanta, H. Borah, and C. R. Das, "Liquid based-cytology Pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions," *Data Brief,* vol. 30, p. 105589, 2020, doi: 10.1016/j.dib.2020.105589.
[16] J. Jantzen, J. Norup, G. Dounias, and B. Bjerregaard, "Pap-smear Benchmark Data For Pattern Classification," 2005. [Online]. Available: https://www.researchgate.net/profile/Jan-Jantzen/publication/282157686_The_Pap_Smear_Benchmark/links/582ae2fe08ae102f071ff4bb/The-Pap-Smear-Benchmark.pdf.
[17] L. D. Nguyen, D. Lin, Z. Lin, and J. Cao, "Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation," 2018: IEEE, pp. 1-5, doi: 10.1109/ISCAS.2018.8351550.
[18] H. Lin, Y. Hu, S. Chen, J. Yao, and L. Zhang, "Fine-Grained Classification of Cervical Cells Using Morphological and Appearance Based Convolutional Neural Networks," *IEEE Access,* vol. 7, pp. 71541-71549, 2019, doi: 10.1109/ACCESS.2019.2919390.
[19] A. Dongyao Jia, B. Zhengyi Li, and C. Chuanwang Zhang, "Detection of cervical cancer cells based on strong feature CNN-SVM network," *Neurocomputing,* vol. 411, pp. 112-127, 2020, doi: 10.1016/j.neucom.2020.06.006.
[20] A. R. Bhatt, A. Ganatra, and K. Kotecha, "Cervical cancer detection in pap smear whole slide images using convnet with transfer learning and progressive resizing," *PeerJ Comput Sci,* vol. 7, p. e348, 2021, doi: 10.7717/peerj-cs.348.
[21] A. Khamparia, D. Gupta, V. H. C. Albuquerque, A. Kumar, and R. Jhaveri, "Internet of health things-driven deep learning system for detection and classification of cervical cells using transfer learning," *J Supercomput,* vol. 76, 2020, doi: 10.1007/s11227-020-03159-4.
[22] M. M. Rahaman and et al., "DeepCervix: A deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques," *Comput Biol Med,* vol. 136, p. 104649, 2021, doi: 10.1016/j.compbiomed.2021.104649.
[23] C. Zhang, D. Jia, Z. Li, and N. Wu, "Auxiliary classification of cervical cells based on multi-domain hybrid deep learning framework," *Biomed Signal Process Control,* vol. 77, p. 103739, 2022, doi: 10.1016/j.bspc.2022.103739.
[24] W. Liu and et al., "CVM-Cervix: A hybrid cervical Pap-smear image classification framework using CNN, visual transformer and multilayer perceptron," *Pattern Recognit,* vol. 130, p. 108829, 2022, doi: 10.1016/j.patcog.2022.108829.
[25] K. Hemalatha, V. Vetriselvi, and A. Aruna Gladys, "CervixFuzzyFusion for cervical cancer cell image classification," *Biomed Signal Process Control,* vol. 85, p. 104920, 2023, doi: 10.1016/j.bspc.2023.104920.
[26] O. Attallah, "CerCan·Net: Cervical cancer classification model via multi-layer feature ensembles of lightweight CNNs and transfer learning," *Expert Syst Appl,* vol. 229, p. 120624, 2023, doi: 10.1016/j.eswa.2023.120624.
[27] B. Deo, M. Pal, P. Panigarhi, and A. Pradhan, "CerviFormer: A Pap-smear based cervical cancer classification method using cross attention and latent transformer," 2023, doi: 10.1002/ima.23043.
[28] L. Qian, Q. Huang, Y. Chen, and J. Chen, "A Voting-Stacking Ensemble of Inception Networks for Cervical Cytology Classification," 2023.
[29] D. Liu, S. Wang, D. Huang, G. Deng, F. Zeng, and H. Chen, "Medical image classification using spatial adjacent histogram based on adaptive local binary patterns," *Comput Biol Med,* vol. 72, 2016, doi: 10.1016/j.compbiomed.2016.03.010.
[30] S. Fekri Ershad, "Pap smear classification using combination of global significant value, texture statistical features and time series features," *Multimed Tools Appl,* vol. 78, 2019, doi: 10.1007/s11042-019-07937-y.
[31] N. K. Chauhan and K. Singh, "Impact of Variation in Number of Channels in CNN Classification model for Cervical Cancer Detection," 2021, pp. 1-6, doi: 10.1109/ICRITO51393.2021.9596366.
[32] O. Yaman and T. Tuncer, "Exemplar pyramid deep feature extraction based cervical cancer image classification model using pap-smear images," *Biomed Signal Process Control,* vol. 73, p. 103428, 2022, doi: 10.1016/j.bspc.2021.103428.
[33] C. Macancela, M. E. Morocho-Cayamcela, and O. Chang, "Deep Reinforcement Learning for Efficient Digital Pap Smear Analysis," *Computation,* vol. 11, no. 12, 2023, doi: 10.3390/computation11120252.
[34] T. Haryanto, I. S. Sitanggang, M. A. Agmalaro, and R. Rulaningtyas, "The Utilization of Padding Scheme on Convolutional Neural Network for Cervical Cell Images Classification," 2020, pp. 34-38, doi: 10.1109/CENIM51130.2020.9297895.
[35] K. Win, Y. Kitjaidure, K. Hamamoto, and T. Aung, "Computer-Assisted Screening for Cervical Cancer Using Digital Image Processing of Pap Smear Images," *Applied Sciences,* vol. 10, p. 1800, 2020, doi: 10.3390/app10051800.
[36] A. Manna, R. Kundu, D. Kaplun, A. Sinitca, and R. Sarkar, "A fuzzy rank-based ensemble of CNN models for classification of cervical cytology," *Sci Rep,* vol. 11, no. 1, p. 14538, 2021, doi: 10.1038/s41598-021-93783-8.
[37] M. Raza and et al., "Advanced Feature Extraction for Cervical Cancer Image Classification: Integrating Neural Feature Extraction and AutoInt Models," *Sensors,* vol. 25, p. 2826, 2025, doi: 10.3390/s25092826.
[38] A. Sharma and R. Parvathi, "Enhancing Cervical Cancer Classification: Through a Hybrid Deep Learning Approach Integrating DenseNet201 and InceptionV3," *IEEE Access,* vol. PP, p. 1, 2025, doi: 10.1109/ACCESS.2025.3527677.
[39] S. L. Tan, G. Selvachandran, W. Ding, R. Paramesran, and K. Kotecha, "Cervical Cancer Classification From Pap Smear Images Using Deep Convolutional Neural Network Models," *Interdiscip Sci,* vol. 16, pp. 1-23, 2023, doi: 10.1007/s12539-023-00589-5.
[40] R. Singh, H. Kaur, and J. Malhotra, "Cervical cancer detection through Pap smear images using hybrid deep feature extraction and ensemble machine learning,"

Multiscale and Multidisciplinary Modeling, Experiments and Design, vol. 8, 2025, doi: 10.1007/s41939-025-00834-y.

[41] H. Kaur, R. Sharma, and J. Kaur, "Comparison of deep transfer learning models for classification of cervical cancer from pap smear images," *Sci Rep,* vol. 15, 2025, doi: 10.1038/s41598-024-74531-0.

[42] O. Bilal, A. Hekmat, and S. U. R. Khan, "Automated cervical cancer cell diagnosis via grid search-optimized multi-CNN ensemble networks," *Network Modeling Analysis in Health Informatics and Bioinformatics,* vol. 14, no. 1, p. 67, 2025, doi: 10.1007/s13721-025-00563-9.

[43] A. Khowaja, B. Zou, and X. Kui, "Enhancing cervical cancer diagnosis: Integrated attention-transformer system with weakly supervised learning," *Image Vis Comput,* vol. 149, p. 105193, 2024, doi: 10.1016/j.imavis.2024.105193.

[44] B. Z. Wubineh, A. Rusiecki, and K. Halawa, "SE-DeepLabV3+: Cervical Cell Segmentation and Classification Using a Novel SE-Based DeepLabV3+ and Ensemble Method," *IEEE Access,* 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/11072403/.

[45] A. K. Sharma, A. Nandal, A. Dhaka, A. Alhudhaif, K. Polat, and A. Sharma, "Diagnosis of cervical cancer using CNN deep learning model with transfer learning approaches," *Biomed Signal Process Control,* vol. 105, p. 107639, 2025, doi: 10.1016/j.bspc.2025.107639.

[46] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," 2017, doi: 10.1609/aaai.v31i1.11231.

[47] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," 2018, pp. 4510-4520, doi: 10.1109/CVPR.2018.00474.

[48] S. Akçay, M. E. Kundegorski, M. Devereux, and T. P. Breckon, "Transfer learning using convolutional neural networks for object classification within X-ray baggage security imagery," 2016: IEEE, pp. 1057-1061, doi: 10.1109/ICIP.2016.7532519.

[49] L. D. Nguyen, R. Gao, D. Lin, and Z. Lin, "Biomedical image classification based on a feature concatenation and ensemble of deep CNNs," *J Ambient Intell Humaniz Comput,* pp. 1-13, 2019, doi: 10.1007/s12652-019-01276-4.

[50] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013.

[51] H. Dalianis, "Evaluation Metrics and Evaluation

Clinical Text Mining: Secondary Use of Electronic Patient Records." Cham: Springer International Publishing, 2018, pp. 45-53.