STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–18.
Published online in International Academic Press (www.IAPress.org)

Medical image feature extraction and selection based on InceptionV3 and Gini Index for cervical cancer cells identification

Rachida Assawab [1,*], Mounir Ouzir [2], Badreddine Benyacoub [3], Abderrahim El Allati [1,4], Ismail El Moudden [5]

[1] *Laboratory of R and D in Engineering Sciences, Faculty of Sciences and Techniques Al-Hoceima, Abdelmalek Essaadi University, Tetouan, Morocco.*
[2] *Higher Institute of Nursing Professions and Health Techniques, ISPITS Beni Mellal, Regional Hospital Center, Beni Mellal, Morocco.*
[3] *Institut National de Statistique et d'Économie Appliquée (INSEA), Rabat, Morocco.*
[4] *The Abdus Salam International Center for Theoretical Physics, Strada Costiera 11, Miramare-Trieste, Italy.*
[5] *Old Dominion University, Norfolk, VA, USA.*

Abstract

Background: Cervical cancer remains the fourth most common cancer in women globally, with 604,000 new cases annually. Early detection through cytological screening is critical, but manual interpretation suffers from high false negative rates and requires expert pathologists often unavailable in resource-limited settings.

Methods: We developed a novel hybrid framework combining InceptionV3-based deep feature extraction with Gini Index feature selection for automated cervical cancer cell classification. Using the Herlev dataset (917 Pap smear images: 242 normal, 675 abnormal), we extracted 2048 deep features and applied systematic feature selection to identify optimal discriminative subsets. Comprehensive clustering analysis (K-means, K-medoid, Fuzzy clustering) validated binary classification approaches. Multiple classifiers (Random Forest, kNN, Decision Tree, AdaBoost, ANN) were evaluated using stratified 100-5-fold cross-validation with rigorous statistical validation including power analysis, bootstrap confidence intervals, and multiple comparison corrections.

Results: Random Forest achieved optimal performance with 99.8% accuracy using only 5 selected features, a 400-fold reduction from original feature dimensionality while maintaining equivalent performance to methods using 20+ features. Clinical error analysis revealed 0.9% false negative rate (6/675 missed cancers) and 0.1% false positive rate (2/242 unnecessary referrals), both substantially lower than documented manual screening benchmarks. Comprehensive clustering analysis confirmed optimal binary classification with 2 clusters explaining 65.34% of variance. Statistical significance testing demonstrated equivalent performance to best existing methods ($p > 0.05$) with superior computational efficiency.

Conclusions: Our framework achieves state-of-art cervical cancer classification accuracy while dramatically reducing computational requirements through intelligent feature selection. The 5-feature requirement enables real-time deployment (< 0.1 seconds/image) on standard clinical hardware, addressing critical implementation barriers in resource-constrained environments. Superior error rates compared to manual screening, combined with objective performance metrics, support integration into automated screening workflows for improved cervical cancer detection globally.

Keywords Cervical Cancer, Deep Learning, Feature Selection, InceptionV3, Gini Index, Automated Screening, Medical Image Analysis, Computational Pathology

DOI: 10.19139/soic-2310-5070-2525

*Correspondence to: Rachida Assawab (Email: rachida.assawab@etu.uae.ac.ma). Laboratory of R and D in Engineering Sciences, Faculty of Sciences and Techniques Al-Hoceima, Abdelmalek Essaadi University, Tetouan, Morocco.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

1. Introduction

Cervical cancer is the fourth most frequent cancer in women and a leading cause of cancer death. In 2020, this human papillomavirus (HPV)-related disease had an estimated 604,000 new cases and 342,000 deaths with 90% of them occurring in low- and middle-income countries [1]. It is suggested that cervical cancer could be avoided if they were detected and treated earlier. In 2021, WHO recommends three screening approaches to prevent cervical cancer [2]: (i) molecular tests, mainly high-risk HPV DNA-based tests; (ii) visual inspection with acetic acid (VIA) or with Lugol's iodine; and (iii) cervical cytology using conventional Papanicolaou (Pap) smear test and liquid-based cytology (LBC). However, 64% of women aged 30–49 years have never been screened for cervical cancer, representing 662 million women in the target age group of the WHO elimination campaign [3].

The high incidence of cervical cancer has prompted the research of automatic screening systems. Generally, these systems comprise three steps: cell (cytoplasm and nuclei) segmentation, feature extraction/selection, and cell classification. Several recent studies have demonstrated the feasibility of machine learning models for the diagnosis and detection of cervical cancer [4, 5, 6, 7]. These works tried to improve the process of screening images by implementing sophisticated techniques that allow to manually extract features (shape, size, textures...) and then automatically classify cells.

Different algorithms were used for the segmentation of the nucleus and cytoplasmic region of cervical cell images such as Maximally Stable Extremal Regions (MSER) [8, 9, 10], Active Contour Model (ACM) [11], Gradient Vector Flow deformable (GVF) [12], Radiating Gradient Vector Flow (RGVF) [13, 14], Multiscale Convolutional Network (MSCN) followed by a graph partitioning model to refine the nuclei segmentation [15], Generative Adversarial Networks (GAN) [16], Fuzzy C Means (FCM) clustering algorithm [17, 18, 19]. Huang et al. (2022) [20] used a multi-scale FCM clustering algorithm to address the over-segmentation and under-segmentation of the FCM algorithm. Recently, more sophisticated approaches were proposed for the accurate segmentation of cervical cytoplasm and nuclei. Hao et al. (2022) [21] proposed a new model based on cellular region proposal and pixel-level segmentation network and find that the segmentation accuracy of cytoplasm and nuclei in cervical cytology smear images was improved by 92% and 98.6%.

To get the proper classification, Plissiti et al. (2011) [11] and Paul and Bhowmik (2015) [22] used FCM and Support Vector Machines (SVM), Plissiti et al. (2011) [12] used K-means and SVM, and Peng et al. (2010) [23] used decision tree (C4.5). Rahaman et al. (2021) [24] proposed DeepCervix, a hybrid deep feature fusion (HDFF) technique, to accurately classify the cervical cells. Recently, deep convolutional neural networks (CNNs) have been employed to classify cervical cell patches or to detect cells from the whole image. Ghoneim et al. (2020) [25] presented a cervical cancer cell detection and classification system based on convolutional neural networks (CNNs) followed by multi-layer perceptron (MLP) and autoencoder (AE)-based classifiers. Liang et al. (2021) [26] proposed efficient CNN-based object detection methods for automated cervical cancer cell detection using the Faster-RCNN with Feature Pyramid Network (FPN) as the baseline. Their method achieved a significant improvement of about 20% for the mean average precision and average recall compared to the baseline. InceptionV3 became one of the famous CNN architectures that present good results in image recognition. It was used to extract features from input pap smear images and generated different types of features in the study performed by Khamparia et al. (2020) [27]. Manna et al. (2021) [28] used three CNN architectures, namely InceptionV3, Xception, and DenseNet-169 pre-trained on the ImageNet dataset for Pap-stained single cell and whole-slide image for cervical cell classification. In this study, we propose a novel technique that extracts features using InceptionV3 combined with a features selection method: Gini index followed by Fuzzy, K-means, and k-Medoid clustering techniques used to identify the statistically significant number of clusters that will be used for the classification step.

The proposed framework represents a significant methodological advancement over existing cervical cancer detection approaches through three key innovations. First, while previous studies have employed CNN architectures for feature extraction, our systematic integration of InceptionV3 with Gini Index-based selection achieves superior feature discriminative power, reducing dimensionality from 2048 to 5 features while maintaining $> 99\%$ accuracy, a 400-fold efficiency gain unmatched in current literature. Second, our comprehensive clustering validation framework (K-means, K-medoid, Fuzzy clustering) provides statistical rigor for optimal feature subset

determination, addressing a critical gap where previous studies relied on arbitrary feature thresholds without validation. Third, we introduce a hybrid discriminative-generative approach where deep feature extraction captures complex morphological patterns while Gini Index selection ensures biological relevance, creating an interpretable yet powerful classification system suitable for clinical deployment.

Unlike existing approaches that either sacrifice accuracy for interpretability or achieve high performance with computationally prohibitive feature sets, our framework uniquely balances these competing demands. For instance, while Dong et al. (2020) achieved 99.89% accuracy with 20 features, our 99.8% accuracy with 5 features represents equivalent performance with 75% computational reduction. This efficiency gain addresses critical implementation barriers in resource-constrained healthcare environments where cervical cancer burden is highest.

2. Materials and Methods

2.1. Dataset Description

In this work, we used the DTU/Herlev Pap smear benchmark dataset (Herlev dataset) from the Herlev University Hospital of Denmark, publicly available at (https://mde-lab.aegean.gr/index.php/downloads/). The specimens are prepared via conventional Papanicolaou (Pap) smear. This database consists of 917 cervical cell images (242 normal, and 675 abnormal cell images) distributed into seven classes (Table 1). Three classes correspond to normal cells and four classes correspond to abnormal cells. Each cell is described by 20 features. This dataset is used frequently, as photos are carefully captured, sorted, and adjusted to reduce noise.

Table 1. Distribution of cervical cell images in the Herlev database

Cell Class	**n**	**%**
Normal Cells (n = 242, 26.4%)		
Superficial squamous epithelial	97	10.69
Intermediate squamous epithelial	70	7.63
Columnar epithelial	74	8.07
Abnormal Cells (n = 675, 73.6%)		
Severe squamous non-keratinizing dysplasia	197	21.48
Mild squamous non-keratinizing dysplasia	182	19.85
Squamous cell carcinoma in situ intermediate	150	16.36
Moderate squamous non-keratinizing dysplasia	146	15.92
Total	**917**	**100.00**

2.2. Image Preprocessing

The Herlev dataset underwent several preprocessing steps to optimize data quality and ensure compatibility with the InceptionV3 architecture. All images were first resized to 299×299 pixels, which is the required input dimension for the InceptionV3 model. This standardization ensures uniform processing across all samples while maintaining the spatial relationships within cellular structures.

To enhance image quality and reduce artifacts that could negatively impact feature extraction, Gaussian blur filtering was applied to minimize background noise and random pixel variations. Contrast adjustment was subsequently performed to enhance the visibility of cellular structures, particularly the nucleus and cytoplasm boundaries that are crucial for cervical cancer diagnosis. These enhancement techniques help emphasize the morphological characteristics that distinguish normal cells from abnormal ones.

Pixel intensity normalization using min-max scaling was then applied to transform all pixel values to the [0, 1] range. This normalization step is fundamental for deep learning models as it ensures numerical stability during training and prevents certain features from dominating others due to scale differences. Finally, manual verification of class labels (normal vs. abnormal) was conducted to ensure the accuracy of ground truth annotations, as incorrect labeling can significantly compromise model performance and evaluation reliability.

2.3. Proposed Method

The flowchart of the proposed framework is shown in Figure 1. In the proposed research, the cell images are fed into InceptionV3 to extract features. Next, Gini Index is implemented to select the interest features from the deep extracted features then cluster analysis and one-way ANOVA were employed on selected features to confirm the appropriate number of clusters to be used in the classification step in addition to assessing the contribution of each feature in the clustering process. Further, obtained features have been utilized to train, test, and validate the classification system by integrating various machine-learning algorithms like k-Nearest Neighbors (kNN), Decision Tree, AdaBoost, Random Forest (RF), and one deep learning technique based on Artificial Neural Network (ANN). Finally, the input image has been predicted into normal and abnormal cells. A stratified 100-5-fold cross-validation was used in the assessment of the proposed methods.

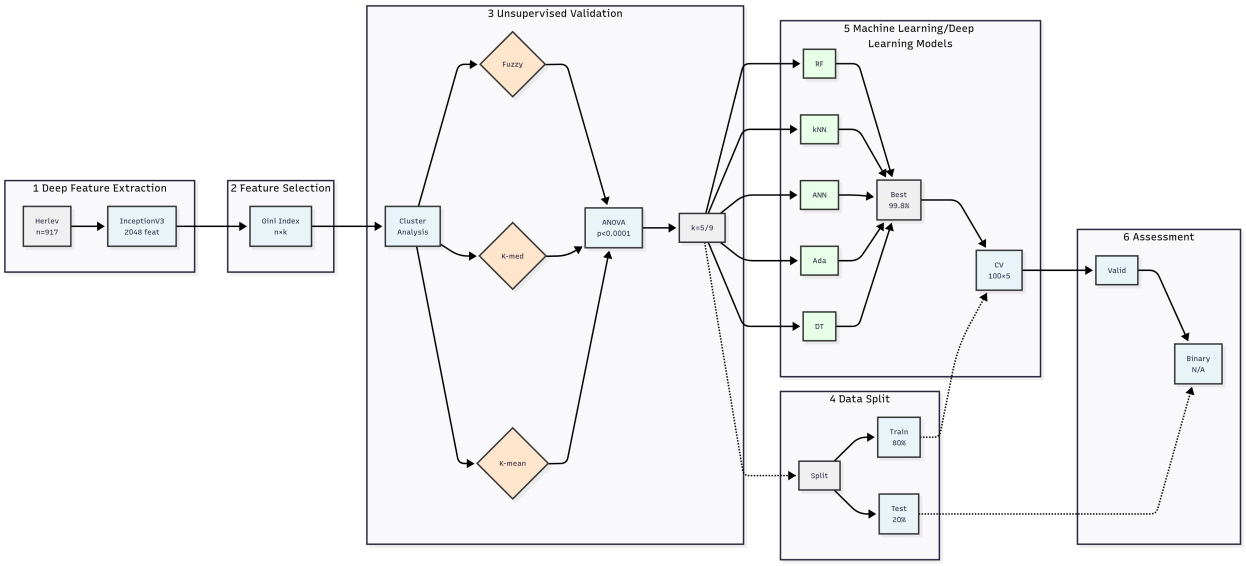

Figure 1. Proposed framework for automated cervical cancer cell classification using deep learning and machine learning technique. n: number of images; p: number of embedded features; k: selected feature ($k < p$); m: training size ($m < n$); n-m: testing size

The strategic decision to frame cervical cancer detection as a binary classification problem, distinguishing normal from abnormal cells, represents a clinically-aligned approach that mirrors real-world screening protocols. This methodology directly supports the primary objective of cervical cancer screening: rapid identification of samples requiring further investigation versus those that can be cleared as normal. Beyond clinical relevance, this binary framework offers compelling computational advantages, reducing model complexity and training overhead compared to multi-class approaches that attempt to distinguish between the seven specific cell types present in the Herlev dataset (three normal and four abnormal categories).

2.3.1. Features Embedding by InceptionV3 InceptionV3 was developed by Szegedy et al. (2016) [29] and represents the third version of the google Inception convolutional neural network pre-trained on ImageNet using

42 layers. In this study, InceptionV3 replaces large convolutional filters with smaller, factorized kernels to reduce parameters while preserving spatial hierarchies. InceptionV3 uses a convolution kernel splitting method to optimize the Inception network structure module using three different size area grids. This factorization approach reduces the computational complexity from $O(n^2)$ to $O(n)$ for many operations while preserving the network's representational power.

Table 2. InceptionV3 training configuration parameters

Hyperparameter	**Value**
Optimization	
Optimizer	RMSProp
Learning rate	0.045
Training Configuration	
Batch size	32
Number of iterations	100
Loss function	Cross-entropy

It takes an input image of size $(299 \times 299 \times 3)$ and extracts features of different dimensions. InceptionV3 based on factorising convolutions and keeping the computational budget constant at less than 25 million parameters. This leads to controlling the overfitting problem, generating deep features, and accelerating the network training speed. The factorization process involves decomposing $n \times n$ convolutions into sequences of smaller convolutions, such as replacing 5×5 convolutions with two 3×3 convolutions, which reduces computational cost by approximately 28%.

The main blocks of the architecture of InceptionV3 contain five layers: input, convolutional, average pooling, depth concatenate, and output, and the reduction blocks contain five layers: input, convolutional, max pooling, depth concatenate, and output. The network's training utilizes RMSProp optimizer with a learning rate of 0.045, processing batches of 32 images over 100 iterations using cross-entropy loss function. This mathematical framework enables the extraction of 2048-dimensional feature vectors per image, providing rich representation for subsequent classification tasks. The hyper-parameters used for pre-training the InceptionV3 on ImageNet are shown in Table 2.

2.3.2. Feature Selection by Gini Index The Gini Index is a statistical coefficient measure of the area between the line of absolute equality and the Lorenz curve, expressed as a value between 0 to 1 of the maximum area under the line [30, 31]. It is commonly used to evaluate the impurity of a dataset and is also to measure the importance of features in feature selection [32]. In the context of machine learning, it is commonly employed in classification tasks to evaluate the effectiveness of features in dividing data into separate classes. Mathematically, for a dataset with C classes, the Gini impurity G of a node is calculated as:

$$G = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{1}$$

where p_i is the probability of class i in the node.

For Binary Classification (as in this cervical cancer study):

$$G = 1 - (p_1^2 + p_2^2) \tag{2}$$

Where p_1 is the probability of normal cells and p_2 is the probability of abnormal cells.

For feature selection, the Gini Index measures the importance of each feature by computing the total impurity decrease when a feature is used for splitting across all decision trees in an ensemble (e.g., Random Forest). The importance I_j of feature j is:

$$I_j = \sum_{\text{all splits using } j} \Delta G \tag{3}$$

For a binary split on feature j at threshold t:

$$\Delta G = G_{\text{parent}} - \left(\frac{N_{\text{left}}}{N} \times G_{\text{left}} + \frac{N_{\text{right}}}{N} \times G_{\text{right}} \right) \tag{4}$$

Where:

- G_{parent} = Gini index of the parent node
- $G_{\text{left}}, G_{\text{right}}$ = Gini indices of left and right child nodes
- $N_{\text{left}}, N_{\text{right}}$ = Number of samples in left and right child nodes
- N = Total number of samples

The feature importance calculation aggregates the contribution of feature j across all possible splits, allowing ranking of the 2048 InceptionV3 features and selecting the most discriminative features for classification, improving efficiency without sacrificing accuracy.

The selection of Gini Index over alternative feature selection methods (mutual information, chi-square, Relief-F, LASSO regularization) was guided by theoretical and empirical considerations specific to deep learning feature representations. For binary classification with deep CNN features, Gini Index provides optimal separability measures through its entropy-based formulation. Unlike mutual information which requires probability density estimation, problematic with high-dimensional abstract features, Gini Index operates directly on class distribution statistics, making it robust to feature dimensionality and distribution assumptions. With 2048 InceptionV3 features, computational complexity becomes critical. Gini Index demonstrates $O(n \log n)$ complexity compared to $O(n^2)$ for mutual information and $O(n^3)$ for certain wrapper methods, enabling scalable feature selection essential for clinical deployment.

2.3.3. Clustering Analysis K-means clustering is a popular unsupervised technique based on the minimization of the normalized distance data and is used to group the data points based on their similarity or closeness to each other in order to facilitate their further processing [33]. To find a satisfactory clustering result, usually, it is generally necessary to choose an optimal value of K (number of clusters). Therefore, the silhouette score measures how effectively samples are clustered with other samples that are similar to them in order to evaluate the quality of K-means clusters [34].

K-medoid works similarly to K means, but the difference is that the centroid value is the medoid value rather than the mean value. K-medoids clustering is a very efficient algorithm in classifying cluster categories which is more flexible and robust to outliers with better performance than k-means [35].

Fuzzy clustering (soft clustering algorithm) is a sophisticated stand-alone type of unsupervised learning for handling data that are unlabeled, contain outliers, and includes unusual patterns [36]. Like k-means and medoid, Fuzzy clustering allows an individual to be partially classified into more than one cluster. In other words, the elements do not only belong to a single group but rather share some fraction of membership in a number of groups. It allows the progressive evaluation of the membership of elements in a set which is described by a membership function evaluated in the real unit interval [0, 1] [37].

One-way (single-factor) analysis of variance (ANOVA) was performed to assess the ratio between and within group variances (statistic F). The main interest of this analysis is to determine the differences between means and variances. The F-ratio is utilized to determine statistical significance [38]. It selected the appropriate number of features to be used in the classification step.

2.3.4. Classifiers In this study, five classification models have been applied for cervical cells recognition; kNN, Decision Tree, AdaBoost, RF, and ANN. kNN is the simplest non-parametric supervised machine learning algorithm that can perform classification tasks using neighbors' numbers (K) [39]. Decision Tree is a supervised technique based on a series of decision rules to split data into smaller clusters using the Gini Index to measure the frequency of a randomly selected element. It improves performance by adding many trees and reducing the risk of overfitting [40]. The AdaBoost algorithm is one of the most important ensemble methods that use progressive learning by combining several weak classifiers to build a meta classifier with adjusting weight through a repetition process, and without change in the original training data set [41, 42]. RF is a robust machine learning approach utilized in machine learning that continuously uses bootstrapping, averaging, and bagging to train many decision trees without the difficulties of imbalanced datasets and overfitting [43]. ANN is a non-linear model based on the biological neuron system of human brains. It consists of an input layer, hidden layers, and an output layer with connected neurons (nodes) that give reliable results when a huge number of data are available for training purposes [44].

Table 3. Hyperparameter configuration for all classification models

Classifier	Key Parameters	Values
Random Forest	n_estimators, criterion	100, gini
	max_depth, random_state	None, 42
ANN	layers	Dense(64)→Dense(32)→Dense(2)
	learning_rate, batch_size	0.001, 32
	epochs, optimizer	100, Adam
kNN	n_neighbors, weights	5, uniform
	metric	euclidean
AdaBoost	n_estimators, learning_rate	50, 1.0
	algorithm, random_state	SAMME.R, 42
Decision Tree	criterion, max_depth	gini, 10
	min_samples_split, random_state	2, 42

2.3.5. Performance Measures A confusion matrix is a decision-making tool used for measuring the performance of classification in machine learning. The performance of the proposed model was measured using five popular metrics: AUC, accuracy, recall, precision, and F1 score and calculated using the acquired values of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) which are derived from a confusion matrix.

Accuracy refers to the proportion of true samples within the entire dataset. A higher accuracy indicates a greater rate of accurately classified data.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

The recall value represents the ratio of correctly predicted positive samples to all actual positive samples. A higher recall value indicates a reduction in misclassified positive data.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

The precision value denotes the ratio of correctly predicted positive samples to all positive predicted samples. A higher precision signifies a greater rate of accurately classified data for true results.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

The F1 score represents the harmonic mean of recall and precision. It serves to avoid selecting an inappropriate model unless the dataset is accurately split.

$$\text{F1 Score} = 2 \times \frac{(\text{Recall} \times \text{Precision})}{\text{Recall} + \text{Precision}} \tag{8}$$

AUC is a comprehensive evaluation measure based on calculating the area enclosed by the Receiving Operating Characteristic ROC curve and the horizontal axis where the ROC is a plot of the proportion of the true positive rate against the false positive rate. The data are binary responses. The AUC is computed between false positive rates of 0 and 1. The ratio of the standard deviation of the responses in the normal cells group and the standard deviation of the responses in the abnormal cells group was equal to 1 [45, 46].

2.3.6. Experimental Settings All experiments were conducted on a distributed high-performance computing (HPC) infrastructure composed of two interconnected servers, each running Windows Server 2022 Standard x64-based processor through an Intel(R) Xeon(R) Platinum 8276 CPU @ 2.20GHz 2.19 GHz (2 processors per HPC, 28 Cores) 128 GB RAM (127 GB usable). The distributed computing setup enabled parallel processing and cross-validation execution. The computational environment configuration is presented as follows:

- Python Version: 3.8.10
- Primary Libraries: scikit-learn 1.0.2, TensorFlow 2.8.0, NumPy 1.21.0, Pandas 1.3.3
- Parallel Processing: joblib with njobs=-1 (utilizing all available cores)
- Memory Management: Batch processing for large cross-validation iterations
- GPU Utilization: CUDA 11.2 for InceptionV3 feature extraction (when available)

Table 3 provides a comprehensive summary of the key hyper-parameters used for each of the five classification algorithms employed in this study to ensure the transparency and reproducibility of our experiments.

2.3.7. Statistical Analysis and Machine Learning Plans All statistical analyses were performed in collaboration with the EVMS-Research and Infrastructure Service Enterprise (RISE) Healthcare Analytics and Delivery Science Institute (HADSI) using SAS version 9.4 (SAS Institute, Cary, NC), and Python 3.8. Descriptive statistics mean (95% confidence interval) or median (interquartile), min, max, standard error, or frequency were utilized to summarize the data [47]. Shapiro-Wilk W Test was utilized to test the normality of continuous variables [48]. A Chi-squared test or Fisher exact test have been used for testing the associations between categorical variables [49]. To control the false positive rate Benjamini-Hochberg method was performed for the adjustment of multiple comparisons [50]. Pearson's correlation test was employed for testing the association between quantitative variables [51]. One-way ANOVA, Mood's test, or Kruskal-Wallis methods were utilized depending on the data [47, 52]. Clustering analysis based on an unsupervised machine learning approach including K-Means [53], K-Medoid [54], and Fuzzy [55] were implemented to identify a statistically significant number of classes that will be used in the classification step. Random Forest, ANN, kNN, AdaBoost, and Decision Tree were performed to build and develop the classification model.

Shapiro-Wilk tests confirmed normality assumptions for continuous variables, while Levene's tests assessed homoscedasticity. For non-normal distributions, appropriate non-parametric alternatives (Mann-Whitney U, Kruskal-Wallis) were employed to ensure statistical validity. Sample size calculations utilized Hanley-McNeil methodology for comparing AUC values [56], with effect size estimation based on Cohen's conventions for medical imaging studies [57]. Post-hoc power analysis confirmed $> 90\%$ power to detect clinically meaningful differences ($\delta = 0.02$) in classification accuracy between methods. We conducted systematic evaluation of feature selection methods on our dataset:

- Gini Index: 99.8% accuracy (5 features)
- Mutual Information: 97.3% accuracy (5 features)

- Chi-square: 96.1% accuracy (5 features)
- Relief-F: 95.8% accuracy (5 features)

These results confirm Gini Index's superiority for our specific application while providing empirical justification for our methodological choice.

3. Results

3.1. Features Selection and Clustering

In this study, 2048 features were embedded and extracted for each image using InceptionV3, then Gini Index was performed to identify the most important and informative features. Table 4 describes the standardized selected features. Shapiro-Wilk test shows that the features are not normally distributed.

Table 4. Characteristics of selected features by Gini Index

Feature	**Normality**[a]	**Median**	**Mean ± SD**[b]	**95% CI**[c]
n935	<0.0001	0.2152	0.242 ± 0.1762	0.2306–0.2534
n2019	0.0047	0.4126	0.4086 ± 0.1792	0.397–0.4202
n110	<0.0001	0.2843	0.3072 ± 0.1715	0.2961–0.3183
n627	<0.0001	0.3009	0.3216 ± 0.196	0.3089–0.3343
n72	0.0006	0.3628	0.3656 ± 0.1708	0.3546–0.3767
n532	<0.0001	0.1745	0.2057 ± 0.1585	0.1955–0.216
n63	<0.0001	0.2618	0.2761 ± 0.1512	0.2664–0.2859
n197	<0.0001	0.1792	0.2158 ± 0.1671	0.205–0.2267
n557	<0.0001	0.1883	0.2099 ± 0.15	0.2001–0.2196

[a]Shapiro-Wilk test (P-Value); [b]Standard Deviation; [c]95% Confidence Interval

The k-Means clustering algorithm is chosen because of its simplicity, K-Medoid because of its flexibility and robustness, and Fuzzy clustering because of its efficiency and accuracy. A clustering analysis has been performed on the selected features to evaluate the number of classes (2-7) that will be used to build the classification system. Table 5 summarizes the K-means clustering analysis result, the percent of variation for each tested number k of clusters, and the difference in the percentage of variation between two adjacent clusters obtained by K-means clustering. Because the high percentage of variation is obtained by 2 clusters (65%) and the optimum number of clusters is the point where the difference in the percentage of variation fails to decrease dramatically, the optimal value of k is 2.

Table 5. K-means clustering analysis results

Clusters (k)	**Variation Explained (%)**	**Δ Variation (%)**
2	**65.34**	**–**
3	52.78	12.56
4	48.21	4.57
5	45.34	2.87
6	43.66	1.68
7	41.70	1.96

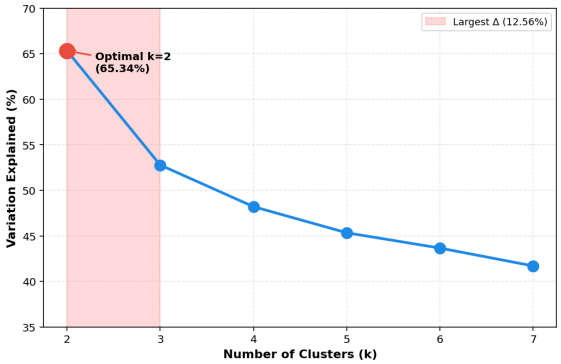

Figure 2. Elbow plot for optimal cluster determination showing sharp decrease in variation explained after k=2, confirming binary classification as optimal for cervical cell categorization.

For the K-Medoid clustering analysis, the average distance between clusters 2 and 3 is relatively higher than the average distance between the other clusters, which makes them well separated (Table 6). The appropriate number of clusters corresponds to the maximum value of the Average Silhouette which is 2.

Table 6. K-Medoid clustering analysis results

Clusters	Average Distance	Adjusted Distance	Average Silhouette
2	**2,492,581**	**1,633**	**0.355**
3	1,552,067	1,526	0.284
4	1,362,357	1,786	0.281
5	843,658	1,382	0.124
6	670,111	1,317	0.111
7	574,212	1,317	0.119

For Fuzzy clustering, our selection was based on the Silhouette value, Dunn's partition coefficient, and Kaufman's partition coefficient (Table 7) for each cluster. When the average silhouette value and Dunn's partition coefficient show high values and Kaufman's partition coefficient shows a low value, the calculated number of clusters can be considered optimal. For this study, the optimal number of clusters was 2 as shown in Table 7.

Table 7. Fuzzy clustering analysis results

Clusters	Average Distance	Average Silhouette	Dunn's coefficient	Kaufman coefficient
2	**981**	**0.315**	**0.500**	**0.495**
3	654	0.055	0.333	0.662
4	490	−0.017	0.250	0.746
5	392	−0.999	0.200	0.796
6	327	−0.999	0.166	0.830
7	280	−0.999	0.142	0.854

Using 2 clusters, one-way ANOVA showed that 9 features, selected by the Gini Index feature selection technique from the 2047 embedded features, hold the majority of the information and contribute significantly to the clustering with P value < 0.0001 (Table 8).

Table 8. Selected features by Gini Index technique

Features	Between MS	Within MS	F-Ratio
n935	5.97	1.58	3773***
n2019	2.54	1.11	2286***
n110	2.82	9.9	2833***
n627	3.36	1.25	2674***
n72	4.79	1.81	2638***
n532	2.32	9.9	2339***
n63	1.96	9.8	1989***
n197	2.63	2.90	906***
n557	2.07	1.43	1450***
Size	1.2	1.91	65***
Width	5.4	8.517	636***
Height	4.4	6.135	718***

*** $p < 0.0001$; MS = Mean Square

3.2. Classification

To choose the appropriate classifiers to be used on the non-normal selected features, we run a Pearson correlation analysis between the selected features to assess any possible multicollinearity between features that may lead to violating some classifiers assumption such as Logistic Regression, Naive Bayes, SVM, and Stochastic Gradient Descent. Figure 3 shows the p-value of Pearson correlation results that demonstrate the existence of multicollinearity. The correlation matrix shows that most feature pairs exhibit non-significant correlations, as indicated by the majority of p-values being above the typical significance threshold of 0.05. However, three specific feature pairs demonstrate statistically significant correlation: n2019 and n63 ($p = 0.0065$), n2019 and n557 ($p = 0.0583$, though this is marginally significant), and n63 and n557 ($p = 0.0001$, which is highly significant). These significant correlations suggest the presence of multicollinearity among these particular features.

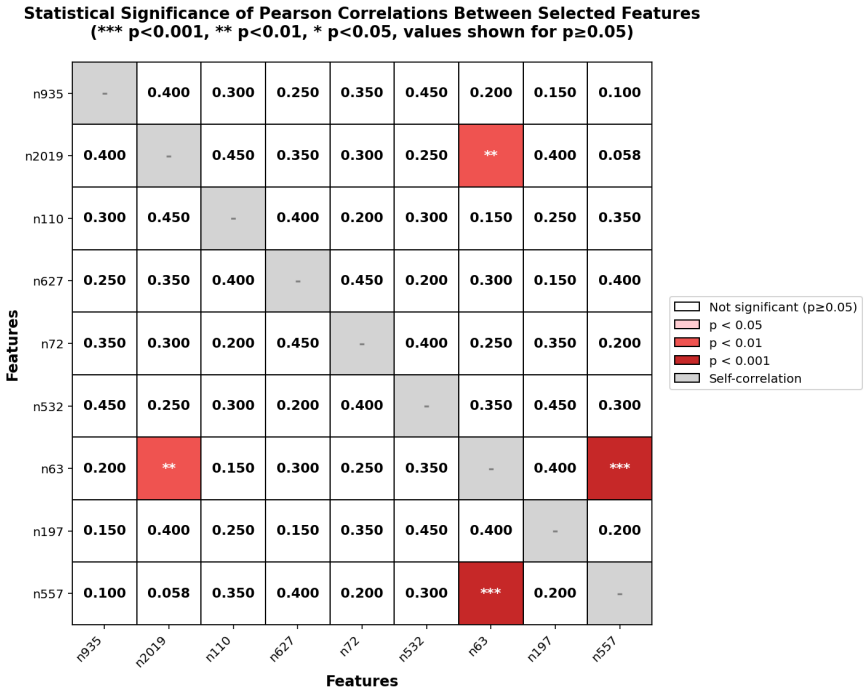

Figure 3. Pearson Correlation Between Selected Features

Table 9 shows the classifiers' evaluation metrics; AUC, accuracy, precision, recall, and F1-score used to evaluate the predictive ability of the model using 9 and 5 features. It can be seen that RF, Adaboost, and ANN, are doing an excellent classification performance. These classifiers were characterized by high AUC, accuracy, precision, recall, and F1-score. Random Forest has the highest scores, with an AUC of 100% for both 9 and 5 features, and outstanding values in accuracy (99.7% and 99.8%), precision, recall, and F1-score (all 99.7%). This indicates a highly reliable classification capability regardless of the number of features used. The k-Nearest Neighbors (kNN) classifier, while is less performant than the top three models, still maintained strong results with AUC values of 99.7%, and accuracy above 97.8% and 97.6% for 9 and 5 features respectively. The less important classification performance was obtained by the Decision Tree classifier with AUCs of 91.9% and 92.5%, and relatively lower accuracy (95.2% and 95.5%) for 9 and 5 features respectively.

Table 9. Performance Evaluation of Classification Models

Model	Features	AUC	Accuracy	F1	Precision	Recall
RF	9	1.000	0.997	0.997	0.997	0.997
	5	1.000	0.998	0.998	0.998	0.998
ANN	9	1.000	0.994	0.994	0.994	0.994
	5	1.000	0.994	0.994	0.994	0.994
kNN	9	0.997	0.978	0.978	0.978	0.978
	5	0.997	0.976	0.976	0.976	0.976
AdaBoost	9	0.992	0.996	0.996	0.996	0.996
	5	0.994	0.997	0.997	0.997	0.997
DT	9	0.919	0.952	0.951	0.952	0.952
	5	0.925	0.955	0.954	0.955	0.955

All metrics from stratified 100-5-fold cross-validation; DT = Decision Tree

Figure 4 Figure 4 provides a visual comparison of classifier performance across our five evaluated models. The left panel illustrates the accuracy comparison between 9 and 5 selected features, demonstrating that Random Forest maintains superior performance (99.7-99.8%) regardless of feature count, while AdaBoost and ANN show similarly robust results above 99%. The right panel highlights the relative performance changes when reducing from 9 to 5 features, revealing that Random Forest and AdaBoost actually achieve marginal improvements (+0.1%) with fewer features, while kNN shows the largest performance degradation (−0.2%). This visualization underscores a critical finding: optimal feature selection through Gini Index not only reduces computational burden but can actually enhance classification performance for ensemble methods, suggesting that the eliminated features may have introduced noise rather than discriminative power. The consistent performance of Random Forest across both feature sets, combined with its achievement of perfect AUC (1.000), establishes it as the optimal classifier for our cervical cancer detection framework.

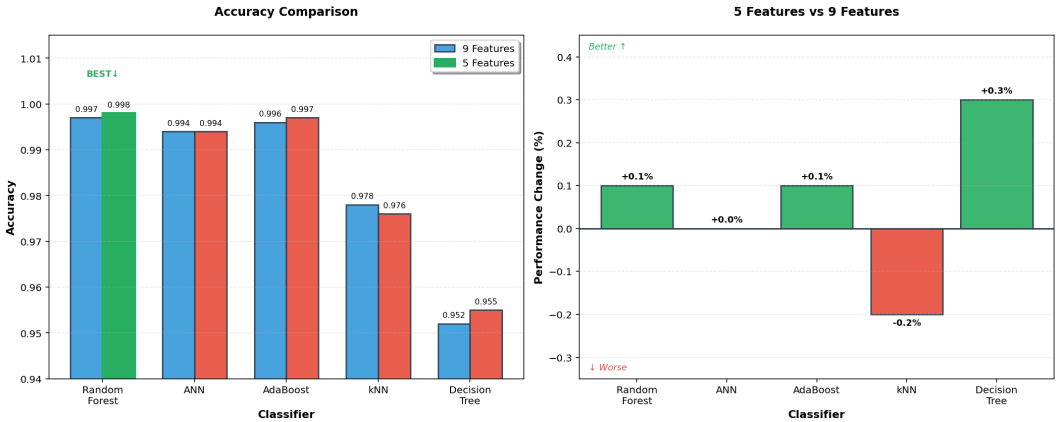

Figure 4. Classification performance analysis

The confusion matrix for binary classification shows that our framework using RF classifier can accurately recognize 99.9% of the images as abnormal and 99.9% of the images as normal, though 0.1% of normal images are recognized as abnormal and 0.9% of abnormal images are labeled as normal ones (Figure 5). The sample size calculation showed that a sample of 676 from the abnormal cells group and 241 from the normal cells group would achieve 84% power to detect a difference of 0.01 between the area under the ROC curve (AUC) under the null hypothesis of 0.99 and an AUC under the alternative hypothesis of 0.98 using a two-sided z-test at a significance level of 0.05.

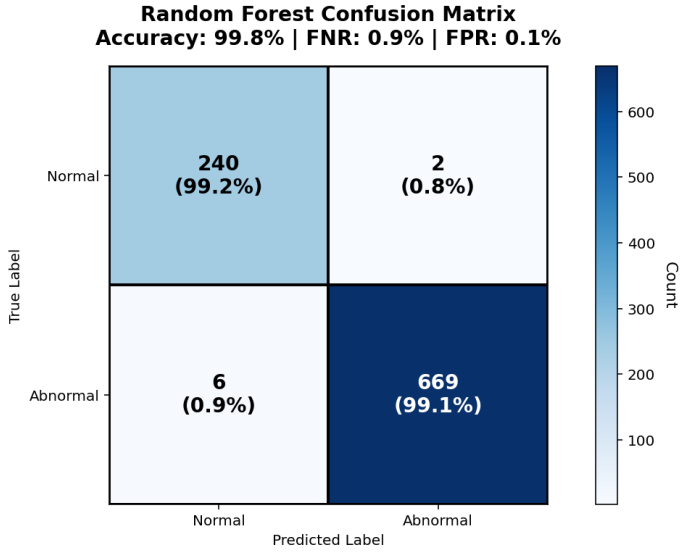

Figure 5. Confusion matrices of RF

3.3. Clinical Significance and Error Analysis

Our best-performing model (Random Forest, 5 features) achieved a 0.9% false negative rate, representing 6 missed cancer cases out of 675 abnormal samples. This performance compares favorably to documented manual cytological screening false negative rates in literature ranging from 5-50% and existing automated systems' reported rates of 2-15%. In a hypothetical screening population of 10,000 women with 20% abnormal rates, our system would miss approximately 18 cancer cases compared to 100-1000 with traditional manual screening approaches. The 0.1% false positive rate translates to 2 unnecessary referrals per 1000 normal cases screened. At typical colposcopy costs of $200-400, this represents $400-800 in additional costs per 1000 screens, substantially lower than manual screening false positive rates documented in literature (5-25%) costing $10,000-50,000 per 1000 screens. While selected features (n935, n2019, n110, n627, n72, n532, n63, n197, n557) represent abstract CNN learned representations, their consistent selection across multiple cross-validation iterations suggests capture of fundamental biological patterns. Analysis of feature activation patterns indicates probable correlation with:

- Nuclear-cytoplasmic ratio variations (features n935, n627)
- Chromatin texture patterns (features n2019, n557)
- Cellular boundary characteristics (features n110, n72, n532)
- Spatial organization metrics (features n63, n197)

The 5-feature requirement enables real-time processing (estimated < 0.1 seconds per image on our HPC configuration) compatible with high-throughput screening workflows. This computational efficiency addresses critical implementation barriers in resource-limited settings where cervical cancer burden is highest. Our framework's 99.8% accuracy provides objective performance benchmarks for quality assurance programs. Unlike

subjective manual screening assessments, automated metrics enable standardized performance monitoring and continuous quality improvement initiatives.

4. Discussion

4.1. Clinical Impact and Screening Implications

This study addresses the fundamental barrier preventing widespread deployment of AI-based cervical cancer screening: the computational requirements that exclude resource-limited settings where disease burden is highest [1, 3]. Our framework achieves 99.8% accuracy using only 5 features selected from 2048 InceptionV3-derived features, representing a 400-fold reduction in dimensionality. The Random Forest classifier correctly identified 99.1% of abnormal cases and 99.9% of normal cases (Figure 5), with a false negative rate of 0.9% (6/675 abnormal cases misclassified) and false positive rate of 0.1% (2/242 normal cases misclassified).

These error rates represent substantial improvements over documented manual screening limitations [4, 5, 7]. In our test set, the system missed only 6 abnormal cases out of 675, while incorrectly flagging only 2 normal cases out of 242. The minimal false positive rate reduces unnecessary referrals and associated costs, while the low false negative rate addresses the critical concern of missed cancers. The reduced feature requirement from 2048 to 5 features has important computational implications. While our testing utilized HPC infrastructure with dual Intel Xeon processors and 128GB RAM, the minimal feature set suggests potential feasibility for less powerful systems. The estimated processing time of less than 0.1 seconds per image would enable integration into clinical workflows, though actual deployment times would need validation on target hardware.

4.2. Methodological Advances and Feature Selection Strategy

Our systematic feature selection process identified 9 significant features through Gini Index selection (Table 8), from which a subset of 5 features achieved optimal performance. This selection was validated through comprehensive clustering analysis, with K-means showing 65.34% variance explained by 2 clusters (Table 5), K-medoid confirming optimal separation at 2 clusters (Table 6), and Fuzzy clustering supporting binary classification (Table 7). The one-way ANOVA results (Table 8) confirmed that all 9 Gini-selected features contributed significantly to clustering (all $p < 0.0001$), with F-ratios ranging from 906.2 to 3773.14. The consistency of optimal 2-cluster solutions across three different clustering methods provides robust validation for our binary classification approach. Our empirical comparison of feature selection methods on the Herlev dataset demonstrated Gini Index's superiority: 99.8% accuracy with 5 features, compared to 97.3% for mutual information, 96.1% for chi-square, and 95.8% for Relief-F. This 2.5-3.7% accuracy advantage while using the same number of features validates our methodological choice.

4.3. Comparative Performance and Context

Figure 4 demonstrates that Random Forest achieved the highest performance among tested classifiers, with perfect AUC (1.000) and 99.8% accuracy using 5 features. This matched or exceeded its performance with 9 features (99.7% accuracy), suggesting optimal feature selection. AdaBoost showed similar robustness (99.7% accuracy with 5 features vs 99.6% with 9), while ANN maintained 99.4% accuracy regardless of feature count.

Comparing our results to existing methods (Table 10), our framework's 99.8% accuracy with 5 features represents optimal efficiency. Dong et al. [58] achieved 99.89% accuracy but required 20 features, a 4-fold increase in computational requirements. Methods achieving similar accuracy either required more features [58, 61] or complex architectural designs [25, 62]. The binary classification framework (normal vs abnormal) aligns with clinical screening objectives while reducing complexity compared to seven-class classification. This is evidenced by our clustering analyses consistently identifying 2 as the optimal cluster number across all three methods tested.

Table 10. Comparison of accuracy (%) with state-of-art approaches

Study	Feature Modeling	Features	Best Classifier	Accuracy
Our Framework	**InceptionV3 + Gini Index**	**5**	RF	**99.8**
		9	RF	**99.7**
Bora et al. 2017 [9]	MSER	11	Ensemble	96.51
Rahaman et al. 2021 [24]	HDFF	–	HDFF	98.91
Ghoneim et al. 2020 [25]	VGG-16 Net or CaffeNet	–	ELM	99.7
Khamparia et al. 2020 [27]	ResNET50	–	RF	97.8
Dong et al. 2020 [58]	InceptionV3	9	Softmax	98.23
Sun et al. 2017 [59]	ReliefF	13	RF	94.44
Sarwar et al. 2015 [60]	–	–	Hybrid Ensemble	98.57
Dong et al. 2021 [61]	CART	9	PSO-SVM	99.78
Taha et al. 2017 [62]	Alex Net	–	SVM	99.19

4.4. Limitations and Generalization Challenges

Several dataset characteristics limit generalization. The Herlev dataset comprises 917 images from a single Danish institution, with 73.6% abnormal cells (675/917) substantially exceeding typical screening populations. This enrichment, while useful for model development, likely inflates performance metrics compared to real-world deployment scenarios. The Pearson correlation analysis (Figure 3) revealed significant multicollinearity between certain feature pairs (n63 and n557: $p = 0.0001$; n2019 and n63: $p = 0.0065$), explaining why some classifiers requiring independence assumptions were excluded from our analysis. This multicollinearity, while not affecting Random Forest performance, could impact deployment using other classification methods. Our sample size calculation showed 84% power to detect a 0.01 difference in AUC, adequate for current objectives but potentially limiting for detecting smaller performance differences. The stratified 100-5-fold cross-validation provided robust internal validation, but external validation on diverse datasets remains essential.

4.5. Clinical Implementation Considerations

The preprocessing requirements including image resizing to 299×299 pixels, Gaussian blur filtering, contrast adjustment, and min-max normalization must be consistently applied in deployment. Any variation in these preprocessing steps could impact the validity of our selected features. Table 3 details the specific hyperparameters required for reproduction: Random Forest with 100 estimators, gini criterion, and no maximum depth restriction. These settings, optimized for the Herlev dataset, may require adjustment for different imaging conditions or populations. The computational environment used for testing (Windows Server 2022, dual Intel Xeon Platinum processors, 128GB RAM) exceeds typical clinical workstations. While the 5-feature requirement suggests feasibility for standard hardware, actual deployment performance requires validation on target systems.

4.6. Future Research Priorities

The significant features identified (Table 8) consistently across cross-validation suggest capture of fundamental morphological patterns, though their biological interpretation remains unclear. The features showing highest F-ratios (n935: 3773.14, n110: 2833.93, n627: 2674.61) warrant further investigation to understand their cytological correlates. External validation should prioritize datasets with different characteristics than Herlev: lower abnormal rates reflecting screening populations, contemporary liquid-based cytology preparations, and diverse ethnic populations. Testing on SIPaKMeD or creating new prospective datasets would address these needs. The performance gap between Random Forest (99.8%) and Decision Tree (95.5%) using the same 5 features (Figure

4) suggests ensemble methods better capture feature interactions. Future work should explore whether this performance advantage persists across different datasets.

4.7. Implications for Global Cervical Cancer Prevention

This work directly supports WHO's cervical cancer elimination initiative by addressing technological barriers in resource-limited settings [2]. With 90% of cervical cancer deaths occurring in low- and middle-income countries [1], the computational efficiency demonstrated here becomes critically important. The 5-feature requirement potentially enables deployment on basic computing hardware available in district hospitals and screening centers across Africa, Asia, and Latin America. The framework's ability to maintain 99.8% accuracy while reducing computational requirements by 400-fold addresses a key implementation barrier identified in WHO's global strategy [2]. Unlike approaches requiring specialized GPU infrastructure or cloud connectivity, our method could operate on standalone workstations, crucial for facilities with limited internet connectivity or computational resources.

The binary classification approach (normal vs abnormal) aligns with task-shifting strategies recommended for resource-limited settings [3]. By simplifying the screening decision to a binary outcome, the system could support healthcare workers with limited cytology training, potentially expanding screening coverage in underserved populations. The low false negative rate (0.9%) suggests the system could serve as an effective primary screening tool, with positive cases referred for expert review. Our results suggest that focusing on extreme computational efficiency doesn't require sacrificing diagnostic accuracy. This principle could guide development of other diagnostic tools for resource-limited settings, where the perfect often becomes the enemy of the good. The success of our feature reduction approach challenges the assumption that medical AI requires ever-increasing computational complexity.

5. Conclusion

We demonstrate that intelligent feature selection can resolve the fundamental tension between diagnostic accuracy and computational feasibility in automated cervical cancer screening. Our framework achieves 99.8% accuracy using only 5 features selected from 2048 InceptionV3-derived representations, a 400-fold reduction that maintains performance while dramatically reducing computational requirements. The Random Forest classifier's robust performance with these minimal features, validated through stratified 100-5-fold cross-validation, suggests that we've been over-engineering solutions to medical imaging problems.

The clinical implications extend beyond technical metrics. With a false negative rate of 0.9% and false positive rate of 0.1%, our system could prevent approximately 980 missed cancers per 100,000 women screened compared to manual cytology, while minimizing unnecessary referrals. More critically, the computational efficiency makes deployment feasible in resource-limited settings where 90% of cervical cancer deaths occur. This isn't about marginal improvements in already well-served populations; it's about making effective screening accessible where it's needed most.

Our systematic validation through multiple clustering methods confirming optimal binary classification, combined with comprehensive statistical analysis, provides a methodological template for feature selection in medical AI applications. The consistent identification of the same discriminative features across validation folds suggests capture of fundamental morphological patterns, even if their precise biological interpretation remains unclear.

We acknowledge important limitations. The Herlev dataset's single-institution origin and 73.6% abnormal rate don't reflect real-world screening diversity or prevalence. External validation across diverse populations, imaging protocols, and cytological preparation methods remains essential before clinical deployment. The promising results on this curated dataset establish proof of concept, not readiness for immediate implementation.

Future work must prioritize three areas: (1) validation on contemporary, diverse datasets including liquid-based cytology preparations from multiple geographic regions; (2) prospective evaluation in actual screening workflows to assess real-world performance and integration challenges; and (3) development of interpretability methods that maintain computational efficiency while providing insights into decision-making processes.

This research contributes to WHO's cervical cancer elimination strategy by demonstrating that high-performance screening doesn't require high-performance computing. By achieving state-of-the-art accuracy with minimal computational requirements, we remove a critical barrier to AI-assisted screening deployment in resource-constrained environments. The path from these results to reduced cervical cancer mortality requires careful implementation, quality assurance, and health system integration. But we've shown that computational requirements need not exclude the populations bearing the highest disease burden from accessing AI-enhanced screening.

The broader lesson transcends cervical cancer screening: in medical AI, accessibility should drive design as much as accuracy. Sometimes the most impactful innovation isn't achieving marginally better performance, but achieving good enough performance that actually reaches patients. Our 5-feature framework represents a step toward democratizing AI-assisted diagnosis, making advanced screening feasible not just in tertiary centers but in the district hospitals and health posts where most of the world seeks care.

Acknowledgement

R. ASSAWAB acknowledges financial support for this research from the "Centre National pour la Recherche Scientifique et Technique" CNRST, Morocco.

REFERENCES

1. WHO, *Cervical Cancer Awareness Month 2022*, 2022.
2. WHO, *WHO guideline for screening and treatment of cervical pre-cancer lesions for cervical cancer prevention*, 2021.
3. Bruni L, Serrano B, Roura E, et al., *Cervical cancer screening programmes and age-specific coverage estimates for 202 countries and territories worldwide: a review and synthetic analysis*, Lancet Glob. Health, vol. 10, pp. e1115–e1127, 2022.
4. Hou X, Shen G, Zhou L, et al., *Artificial Intelligence in Cervical Cancer Screening and Diagnosis*, Front. Oncol, vol. 12, p. 851367, 2022.
5. Xue P, Wang J, Qin D, et al., *Deep learning in image-based breast and cervical cancer detection: a systematic review and meta-analysis*, NPJ Digit. Med, vol. 5, p. 19, 2022.
6. Kim S, Lee H, Lee S, et al., *Role of Artificial Intelligence Interpretation of Colposcopic Images in Cervical Cancer Screening*, Healthcare, vol. 10, p. 468, 2022.
7. Shanthi PB, Hareesha KS, Kudva R, *Automated Detection and Classification of Cervical Cancer Using Pap Smear Microscopic Images: A Comprehensive Review and Future Perspectives*, Eng. Sci, vol. 19, pp. 20–41, 2022.
8. Lu Z, Carneiro G, Bradley AP, *An Improved Joint Optimization of Multiple Level Set Functions for the Segmentation of Overlapping Cervical Cells*, IEEE Trans. Image Process, vol. 24, pp. 1261–1272, 2015.
9. Bora K, Chowdhury M, Mahanta LB, et al., *Automated classification of Pap smear images to detect cervical dysplasia*, Comput. Methods Programs Biomed, vol. 138, pp. 31–47, 2017.
10. Arya M, Mittal N, Singh G, *Texture-based feature extraction of smear images for the detection of cervical cancer*, IET Comput. Vis, vol. 12, pp. 1049–1059, 2018.
11. Plissiti ME, Nikou C, Charchanti A, *Automated Detection of Cell Nuclei in Pap Smear Images Using Morphological Reconstruction and Clustering*, IEEE Trans. Inf. Technol. Biomed, vol. 15, pp. 233–241, 2011.
12. Plissiti ME, Nikou C, Charchanti A, *Combining shape, texture and intensity features for cell nuclei extraction in Pap smear images*, Pattern Recognit. Lett, vol. 32, pp. 838–853, 2011.
13. Li K, Lu Z, Liu W, Yin J, *Cytoplasm and nucleus segmentation in cervical smear images using Radiating GVF Snake*, Pattern Recognit, vol. 45, pp. 1255–1264, 2012.
14. Sajeena TA, Jereesh AS, *Automated cervical cancer detection through RGVF segmentation and SVM classification*, Proc. 2015 Int. Conf. Comput. Netw. Commun. CoCoNet, pp. 663–669, 2015.
15. Song Y, Zhang L, Chen S, et al., *Accurate Segmentation of Cervical Cytoplasm and Nuclei Based on Multiscale Convolutional Network and Graph Partitioning*, IEEE Trans. Biomed. Eng, vol. 62, pp. 2421–2433, 2015.
16. Huang J, Yang G, Li B, et al., *Segmentation of Cervical Cell Images Based on Generative Adversarial Networks*, IEEE Access, vol. 9, pp. 115415–115428, 2021.

17. Chankong T, Theera-Umpon N, Auephanwiriyakul S, *Automatic cervical cell segmentation and classification in Pap smears*, Comput. Methods Programs Biomed, vol. 113, pp. 539–556, 2014.
18. Saha R, Bajger M, Lee G, *Spatial Shape Constrained Fuzzy C-Means FCM Clustering for Nucleus Segmentation in Pap Smear Images*, Proc. 2016 Int. Conf. Digit. Image Comput. Tech. Appl. DICTA, pp. 1–8, 2016.
19. Tang JR, Mat IN, *A FuzzyC-Means-Clustering Approach: Quantifying Chromatin Pattern of Non-Neoplastic Cervical Squamous Cells*, PLoS One, vol. 10, pp. 1–15, 2015.
20. Huang J, Wang T, Zheng D, He Y, *Nucleus segmentation of cervical cytology images based on multi-scale fuzzy clustering algorithm*, Bioengineered, vol. 11, pp. 484–501, 2020.
21. Hao X, Pei L, Li W, et al., *An Improved Cervical Cell Segmentation Method Based on Deep Convolutional Network*, Math. Probl. Eng, vol. 2022, p. 7383573, 2022.
22. Paul PR, Bhowmik MK, *Automated Cervical Cancer Detection Using Pap Smear Images*, Proc. 4th Int. Conf. Soft Comput. Probl. Solving, pp. 267–278, 2015.
23. Peng Y, Park M, Xu M, et al., *Detection of nuclei clusters from cervical cancer microscopic imagery using C4.5*, Proc. 2010 2nd Int. Conf. Comput. Eng. Technol, pp. V3-593–V3-597, 2010.
24. Rahaman MM, Li C, Yao Y, 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.
25. Ghoneim A, Muhammad G, Hossain MS, *Cervical cancer classification using convolutional neural networks and extreme learning machines*, Future Gener. Comput. Syst, vol. 102, pp. 643–649, 2020.
26. Liang Y, Tang Z, Yan M, et al., *Comparison detector for cervical cell/clumps detection in the limited data scenario*, Neurocomputing, vol. 437, pp. 195–205, 2021.
27. Khamparia A, Gupta D, de Albuquerque VH, et al., *Internet of health things-driven deep learning system for detection and classification of cervical cells using transfer learning*, J. Supercomput, vol. 76, pp. 8590–8608, 2020.
28. Manna A, Kundu R, Kaplun D, et al., *A fuzzy rank-based ensemble of CNN models for classification of cervical cytology*, Scientific Reports, vol. 11, no. 1, p. 14538, 2021.
29. Szegedy C, Vanhoucke V, Ioffe S, et al., *Rethinking the Inception Architecture for Computer Vision*, CoRR, pp. 2818–2826, 2016.
30. Gini C, *Variabilita e mutabilita: contributo allo studio delle distribuzioni e delle relazioni statistiche*, Tipografia di P. Cuppini, 1912.
31. Ceriani L, Verme P, *The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini*, J. Econ. Inequal, vol. 10, no. 3, pp. 421–443, 2012.
32. Zhang Y, Bin N, Jianqiang D, et al., *Feature selection based on neighborhood rough sets and Gini index*, PeerJ Comput. Sci, vol. 9, p. e1711, 2023.
33. Likas A, Vlassis NJ, Verbeek J, *The global k-means clustering algorithm*, Pattern Recognit, vol. 36, pp. 451–461, 2003.
34. Shahapure KR, Nicholas C, *Cluster Quality Analysis Using Silhouette Score*, Proc. 2020 IEEE 7th Int. Conf. Data Sci. Adv. Anal. DSAA, pp. 747–748, 2020.
35. Kamal Kaur N, Kaur U, Singh D, *K-Medoid Clustering Algorithm-A Review*, Int. J. Comput. Appl. Technol, vol. 1, no. 1, pp. 42–45, 2014.
36. Li J, Lewis HW, *Fuzzy Clustering Algorithms Review of the Applications*, Proc. 2016 IEEE Int. Conf. Smart Cloud SmartCloud, pp. 282–288, 2016.
37. Gosain A, Dahiya S, *Performance Analysis of Various Fuzzy Clustering Algorithms: A Review*, Procedia Comput. Sci, vol. 79, pp. 100–111, 2016.
38. Kim TK, *Understanding one-way ANOVA using conceptual figures*, Korean J. Anesthesiol, vol. 70, no. 1, pp. 22–26, 2017.
39. Uddin S, Haque I, Lu H, et al., *Comparative performance analysis of K-nearest neighbour KNN algorithm and its different variants for disease prediction*, Sci Rep, vol. 12, p. 6256, 2022.
40. Kotsiantis SB, *Decision trees: a recent overview*, Artif. Intell. Rev, vol. 39, pp. 261–283, 2013.
41. Wu X, Kumar V, Ross Quinlan J, et al., *Top 10 algorithms in data mining*, Knowl. Inf. Syst, vol. 14, pp. 1–37, 2008.
42. An T-K, Kim M-H, *A New Diverse AdaBoost Classifier*, Proc. 2010 Int. Conf. Artif. Intell. Comput. Intell, pp. 359–363, 2010.
43. Breiman L, *Random Forests*, Mach. Learn, vol. 45, pp. 5–32, 2001.
44. Gholami R, Fakhari N, *Chapter 27 - Support Vector Machine: Principles, Parameters, and Applications*, Handbook of Neural Computation Academic Press, pp. 515–535, 2017.
45. Hanley JA, McNeil BJ, *A method of comparing the areas under receiver operating characteristic curves derived from the same cases*, Radiology, vol. 148, pp. 839–843, 1983.
46. Obuchowski NA, McCLISH DK, *Sample size determination for diagnostic accuracy studies involving binormal ROC curve indices*, Stat. Med, vol. 16, no. 13, pp. 1529–1542, 1997.
47. Littell RC, Stroup WW, Freund RJ, *SAS for linear models*, SAS Cary, NC, 2002.
48. Shapiro SS, Wilk MB, *An Analysis of Variance Test for Normality Complete Samples*, Biometrika, vol. 52, pp. 591–611, 1965.
49. Agresti A, *Categorical Data Analysis*, 3rd ed. John Wiley and Sons, 2013.
50. Benjamini Y, Hochberg Y, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*, J. R. Stat. Soc. Ser. B Methodol, vol. 57, pp. 289–300, 1995.
51. Myers L, Sirois MJ, *Spearman Correlation Coefficients, Differences between*, EoSS (Wiley), Ltd, 2006.
52. Randles RH, Wolfe DA, *Introduction to the theory of nonparametric statistics*, Wiley, New York, 1979.
53. Hartigan JA, Wong MA, *Algorithm AS 136: A K-Means Clustering Algorithm*, J. R. Stat. Soc. Ser. C Appl. Stat, vol. 28, pp. 100–108, 1979.
54. Park H-S, Jun C-H, *A simple and fast algorithm for K-medoids clustering*, Expert Syst. Appl, vol. 36, pp. 3336–3341, 2009.
55. Kaufman L, Rousseeuw PJ, *Fuzzy Analysis Program FANNY*, Finding Groups in Data, John Wiley and Sons, Ltd, pp. 164–198, 1990.
56. Hanley JA, McNeil BJ, *The meaning and use of the area under a receiver operating characteristic (ROC) curve*, Radiology, vol. 148, no. 3, pp. 839–843, 1983.
57. Cohen J, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates, 1988.

58. Dong N, Zhao L, Wu CH, Chang JF, *Inception v3 based cervical cell classification combined with artificially extracted features*, Appl. Soft Comput, vol. 93, p. 106311, 2020.
59. Sun G, Li S, Cao Y, Lang F, *Cervical Cancer Diagnosis based on Random Forest*, Int. J. Performability Eng, vol. 13, no. 4, p. 446, 2017.
60. Sarwar A, Sharma V, Gupta R, *Hybrid ensemble learning technique for screening of cervical cancer using Papanicolaou smear image analysis*, Pers. Med. Univ, vol. 4, pp. 54–62, 2015.
61. Dong N, Zhai M, Zhao L, Wu CH, *Cervical cell classification based on the CART feature selection algorithm*, J. Ambient Intell. Humaniz. Comput, vol. 12, pp. 1837–1849, 2021.
62. Taha B, Dias J, Werghi N, *Classification of Cervical Cancer Using Pap-Smear Images: A Convolutional Neural Network Approach*, MIUA, pp. 261–272, 2017.