Vol. 9 Issue 1 Jan 2026

Multimedia Research

Received 11 December, Revised 22 January, Accepted 23 January

Application of Gray-Level Co-Occurrence Matrix (GLCM) Feature Selection Method in Precancerous Cervical Lesion Prediction of Women Living with HIV in Nigeria

Emmanuel Ogala
Computer Science Department
Joseph Sarwuan Tarka University, Makurdi, Benue State, Nigeria.

Aboho Moses Demenongo
Computer Science Department
Joseph Sarwuan Tarka University, Makurdi, Benue State, Nigeria.

Iortyom Solomon
Computer Science Department
Joseph Sarwuan Tarka University, Makurdi, Benue State, Nigeria.

Abstract: Women living with HIV face a heightened risk of developing cervical cancer, particularly in low-resource settings such as Nigeria. Early detection of precancerous cervical lesions is critical to improving outcomes, yet conventional screening methods like VIA/VILI are highly dependent on clinician expertise and subject to variability. These limitations often result in inconsistent diagnoses and missed lesions, especially in settings with limited specialist availability. This study proposes an automated lesion detection system using Gray-Level Co-Occurrence Matrix (GLCM) feature extraction combined with machine learning classifiers to improve diagnostic accuracy. A dataset of 2,516 VIA/VILI cervical images from HIV-positive women was pre-processed using non-local means denoising and grayscale conversion, and key texture features including contrast, energy, entropy, and homogeneity were extracted. Four classifiers, such as Random Forest, Decision Tree, Gradient Boosting, and Logistic Regression, were trained and evaluated using an 80:20 train-test split, with performance assessed through accuracy, AUC, precision, recall, and F1-score. The Random Forest model exhibited superior predictive performance, achieving 97% accuracy and an AUC of 0.99, surpassing existing clinical and image-based approaches. The Decision Tree demonstrated strong performance, while Gradient Boosting showed moderate effectiveness, and Logistic Regression performed least effectively. Comparative analysis with prior studies confirms the superiority of ensemble-based, feature-engineered models for automated cervical lesion detection. The findings highlight the potential of integrating such models into mobile screening platforms like AVIVA to enhance real-time diagnosis, reduce missed cases, and support cervical cancer prevention in low-resource settings. Future work will focus on external validation and scalability to ensure robustness and broad clinical applicability.

Keywords: Cervical Cancer Screening, Precancerous Lesions, HIV-Positive Women, Machine Learning, Low-Resource Settings

Nomenclature

Abbreviation	Definition
HIV	Human Immunodeficiency Virus
VIA	Visual Inspection with Acetic Acid
VILI	Visual Inspection with Lugol's Iodine
HPV	Human Papillomavirus
LMICs	Low- and Middle-Income Countries
GLCM	Gray-Level Co-Occurrence Matrix
NLM	Non-Local Means (denoising algorithm)
RGB	Red, Green, Blue
AUC	Area Under the Curve (from ROC analysis)
ROC	Receiver Operating Characteristic curve
RF	Random Forest
DT	Decision Tree
GB	Gradient Boosting
LR	Logistic Regression
ART	Antiretroviral Therapy

1. Introduction

Women living with HIV face a substantially higher risk of developing cervical cancer compared to women without HIV infection [15]. Globally, approximately 6% of women diagnosed with cervical cancer are HIV-

https://doi.org/10.46253/j.mr.v9i1.a5

positive, and nearly 5% of all cervical cancer cases are associated with HIV [4]. Cervical cancer remains the fourth most prevalent cancer among women worldwide, with an estimated 660,000 new cases reported in 2022, of which approximately 94% of the 350,000 deaths occurred in LMICs [15]. Regions such as sub-Saharan Africa, Central America, and South-East Asia experience the highest incidence and mortality rates due to disparities in vaccination, screening, and treatment coverage, compounded by social and economic factors, including gender inequities and poverty. These intersecting risk factors disproportionately affect vulnerable populations, limiting access to timely and effective cervical cancer prevention services.

In Nigeria, cervical cancer is the second leading cause of cancer-related morbidity and mortality among women, with 12,100 new cases and 8,000 deaths reported in 2020 [11]. Approximately 60.9 million Nigerian women aged 15 and older are at risk of developing cervical cancer if cost-effective prevention and early detection measures are not widely adopted. The global 90-70-90 targets to eliminate cervical cancer provide a roadmap for LMICs, aiming for 90% of girls to be fully vaccinated against HPV by age 15, 70% of women to undergo high-performance screening by ages 35 and 45, and 90% of those diagnosed to receive appropriate treatment by 2030 [11]. Achieving these targets in Nigeria requires innovative, context-appropriate screening strategies that can operate effectively within constrained healthcare systems.

Although several screening methods exist—such as conventional cytology (PAP smear), liquid-based cytology, HPV testing, and VIA/VILI—the Nigerian Federal Ministry of Health recommends VIA/VILI combined with cryotherapy for precancerous lesions as the primary screening and treatment approach [10]. VIA/VILI allows a same-visit "see-and-treat" intervention, but its effectiveness is highly dependent on the skill and experience of healthcare providers, introducing variability and potential diagnostic errors [3]. This limitation underscores the need for automated or assisted approaches that can enhance accuracy and consistency in low-resource settings. Reducing reliance on subjective visual interpretation is therefore critical to improving screening reliability and coverage.

To address this challenge, this study proposes a machine learning–based framework for detecting precancerous cervical lesions in HIV-positive women in Nigeria using VIA/VILI images. The approach utilizes GLCM features combined with traditional classifiers, including Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting, to automatically detect and classify precancerous lesions. Once trained, these models can be deployed across healthcare facilities or remote areas, providing consistent, reliable, and scalable cervical cancer screening, independent of individual healthcare provider proficiency. Such automation has the potential to standardize screening outcomes and support task-shifting in under-resourced clinical environments.

The novelty and contributions of this study are summarized as follows:

- Development of a GLCM-based feature extraction framework tailored for precancerous cervical lesion detection in HIV-positive women.
- Comparative evaluation of multiple machine learning classifiers to identify the most effective model for VIA/VILI image classification.
- Emphasis on computationally efficient approaches suitable for low-resource healthcare settings in Nigeria.
- Provision of a scalable automated screening solution to support early detection and intervention in cervical cancer programs.
- Demonstration of the model's potential to enhance accuracy, reduce human error, and improve consistency in cervical cancer screening.

Collectively, these contributions aim to bridge existing gaps between clinical needs and deployable artificial intelligence solutions.

The remainder of the paper is organized as follows: Section 2 reviews related work and identifies research gaps. Section 3 presents the methodology, including dataset description, feature extraction, and classification strategies. Section 4 details the experimental results and analysis. Section 5 discusses findings, clinical relevance, limitations, and implications. Finally, Section 6 concludes the study and outlines directions for future research.

2. Related Works

AI and ML have increasingly been applied to improve cervical cancer screening, particularly in low-resource settings where access to advanced diagnostic tools is limited. Existing research can be broadly categorized into three methodological approaches: clinical risk–based prediction, handcrafted image feature–based classification, and deep learning–based cervigram analysis. Each approach offers specific advantages, but important limitations remain regarding accuracy, generalizability, and real-world

applicability. Understanding the strengths and weaknesses of these approaches is essential for identifying suitable solutions for resource-constrained healthcare environments.

Clinical risk models utilize demographic and medical variables to identify women at elevated risk of cervical precancer. For example, Namalinzi *et al*. (2024) [9] used features such as ART duration, WHO clinical stage, tuberculosis preventive therapy, viral load suppression, and family planning status in a Random Forest classifier, achieving an AUC of 0.901. While effective for risk stratification, these models depend heavily on complete and high-quality clinical records, which are often inconsistent in Nigerian health facilities. Furthermore, clinical variables do not capture lesion morphology, color, or texture, limiting their ability to replace image-based diagnostic evaluation. Consequently, such models are useful for identifying at-risk populations but insufficient for automated VIA/VILI-based screening. This limitation highlights the necessity of incorporating visual information directly into automated screening systems.

Handcrafted feature methods extract color, texture, and shape descriptors from cervical images to support automated lesion detection. Asiedu *et al*. (2019) [2] demonstrated that an SVM classifier using VIA/VILI images could outperform expert physicians, highlighting the value of manually engineered features. However, these approaches are highly sensitive to variations in illumination, camera type, and image quality, which are common in real-world Nigerian screening programs where smartphones with varying specifications are used. As a result, traditional ML models using handcrafted features may exhibit limited robustness and generalizability. Ensuring resilience to image acquisition variability remains a major challenge for handcrafted feature-based systems.

Deep learning models, such as region-based CNNs [8] and Mask-RCNN architectures [1], automatically learn spatial and semantic features from large image datasets. These models have shown superior performance in lesion localization, segmentation, and classification. However, they require substantial computational resources and large annotated datasets, which are scarce in low-resource environments. Deploying such models on mobile devices is challenging due to model size, inference time, and potential performance drops when applied to out-of-domain images. These constraints limit the practical adoption of deep learning approaches in decentralized screening programs.

Smartphone-based VIA systems focus on practicality and point-of-care feasibility. Viñals *et al*. (2021) [14] introduced a video-based VIA assessment that leverages temporal changes in cervical whitening, improving detection over static-image methods, but requiring video capture capabilities that may not be widely available. Harsono *et al*. (2022) [6] developed a lightweight Android application with high specificity, though its small sample size limits confidence in generalizability. A systematic review by Viñals *et al*. (2023) [13] further highlights wide variations in AI-VIA model performance across studies, due to methodological inconsistencies, limited datasets, and lack of external validation. These findings emphasize the importance of balancing innovation with feasibility and reproducibility.

2.1 Comparative Insights

Taken together, existing methods reveal a trade-off between diagnostic accuracy, computational efficiency, and real-world feasibility:

Table 1. *Comparison of AI Methods: Advantages, Limitations, and Low-Resource Applicability*

Method	Advantages	Limitations	Applicability in Low-Resource Settings
Clinical risk models	Lightweight, interpretable	Cannot perform image-based diagnosis	Useful for risk stratification but not automated screening
Handcrafted image features	Computationally efficient, interpretable	Sensitive to lighting, camera, and image quality	Moderate applicability; may require standardization
Deep learning models	High accuracy, automatic feature learning	Requires large datasets, high computational cost	Limited deployment on mobile devices; scalability issues

Despite progress, a clear methodological gap exists: there is a need for efficient, robust, and scalable texture-based models capable of capturing structural characteristics of precancerous lesions while remaining computationally lightweight. GLCM features are particularly promising in this regard, yet remain underexplored for automated VIA/VILI image analysis in HIV-positive women in Nigeria. This study addresses this gap by combining GLCM feature extraction with traditional machine learning classifiers, aiming to provide a practical, high-performing solution suitable for low-resource healthcare settings. By focusing on texture-driven representations, the proposed approach seeks to achieve an optimal balance between accuracy, interpretability, and deploy ability.

3. Methodology

The proposed model for this research comprises five main phases: Data Collection, Pre-Processing, Feature Extraction, Classification, and Performance Evaluation, as illustrated in Fig. 1. This structured pipeline ensures systematic transformation of raw cervical images into reliable diagnostic predictions.

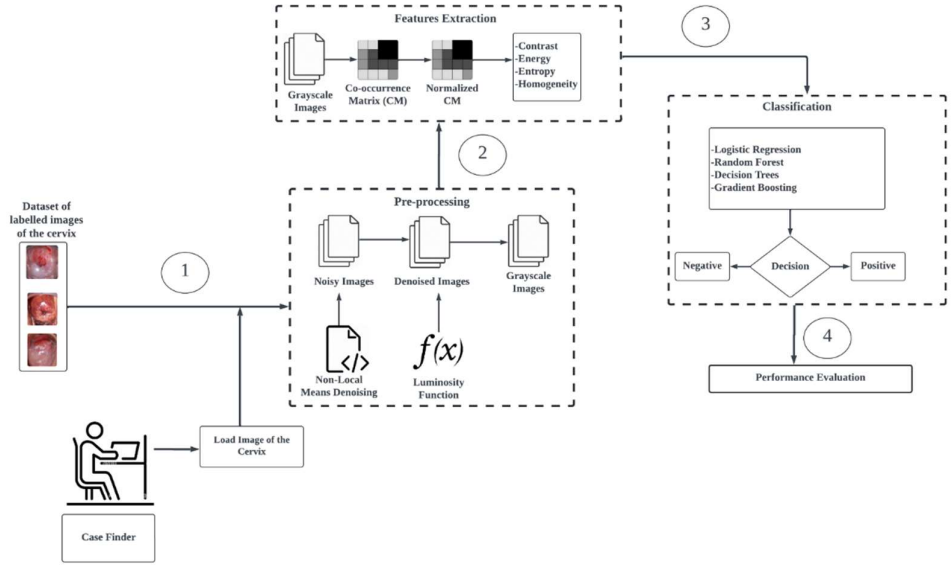

Fig. 1. *The proposed architecture of the model*

3.1 Data Collection

Data for this study were obtained using the APIN VIA Innovative Visual Application (AVIVA), a mobile application developed and maintained by APIN Public Health Initiatives. AVIVA is specifically designed to support cervical cancer screening by enabling real-time review and diagnosis of VIA/VILI-stained cervix images. The app functions by capturing images of VIA/VILI-stained cervixes and transmitting them instantly to expert reviewers, who are trained gynecologists. These reviewers analyze the images and provide diagnostic insights along with treatment recommendations. APIN Public Health Initiatives is a reputable non-governmental organization that provides prevention, care, and treatment services for diseases such as HIV/AIDS, Tuberculosis, and Malaria. It also offers technical support to governmental agencies to enhance the delivery of Reproductive Health, Maternal, Newborn, and Child Health (RMNCH), and Laboratory services. The use of AVIVA ensures that image annotations are clinically validated by domain experts.

A dataset consisting of 2516 images captured from the cervixes of women living with HIV was obtained through the AVIVA mobile application. The dataset consists of 1516 lesions identified as VIA/VILI positive, 1000 lesions identified as VIA/VILI negative. Each image was accompanied by a label indicating whether the cervix had a detected precancerous lesion or not. This balanced dataset provides sufficient variability for training and evaluating machine learning models.

3.2 Pre-Processing

Pre-processing aimed to improve image quality and standardize inputs for feature extraction. Two key steps were applied: These steps were designed to reduce noise-related distortions while preserving diagnostically relevant features.

3.2.1 Non-Local Means (NLM) Denoising

The proposed model employed the non-local means (NLM) denoising algorithm and Luminosity function for this purpose. NLM denoising was used to remove noise while preserving image details such as edges and boundaries, textures, contrast, luminance, and color fidelity. The NLM denoising equation computes the denoised pixel value by averaging the noisy pixel values in the image, with weights determined by the similarity between patches centered around different pixels. Equation 1 provides a mathematical representation of the Non-Local Means (NLM) denoising algorithm. This approach is particularly suitable for medical images where fine structural details must be retained.

Let f represent the noisy image and u represent the denoised image. For each pixel i in the image, the denoised pixel value u_i is calculated as follows:

$$u_i = \frac{\sum_i w(i,j) \cdot f_j}{\sum_i w(i,j)} \quad (1)$$

Where u_i is the denoised value of the pixel at position i. f_j is the noisy value of the pixel at position j. $w(i, j)$ is the weight assigned to the pixel value at position j when denoising pixel i. This weight is determined based on the similarity between the patch centered around pixel i and the patch centered around pixel j. The weights are typically computed using a Gaussian-like function of the form presented in equation 2.

$$w(i,j) = \frac{1}{Z(i)} \exp\left(-\frac{||P_i - P_j||^2}{h^2}\right) \quad (2)$$

Where P_i and P_j are patches centered around pixels i and j, respectively. $||P_i - P_j||^2$ represents the squared Euclidean distance between the patches. h is a parameter that controls the amount of smoothing applied during denoising. $Z(i)$ is a normalization factor to ensure that the weights sum up to 1. The summation in the numerator and denominator is performed over all pixels j in the image.

The luminosity function was used to convert colour images of the cervix from RGB (Red, Green, Blue) colour space to grayscale. The equation is derived from the CIE (Commission Internationale de l'Eclairage) standard luminosity function, which is based on the average sensitivity of the human eye to different wavelengths of light, and the coefficients 0.299, 0.587, and 0.114 represent the relative luminance of each color channel, adjusted to match human perception. Equation 3 presents the formula used to convert the colour images to grayscale. Grayscale conversion simplifies texture analysis by reducing color redundancy.

$$Y = 0.299R + 0.587G + 0.114B \quad (3)$$

Where R, G, and B are the red, green, and blue colour channels of the RGB image, respectively. Y represents the resulting grayscale intensity.

3.2.2 Feature Extraction

The co-occurrence matrix was used to harvest relevant features such as contrast, energy, entropy, and homogeneity from the grayscale images. Texture-based features are well-suited for capturing structural irregularities associated with precancerous lesions. To compute the co-occurrence matrix of the grayscale images, the direction and offset between pixels are set to horizontal and 1, respectively, and for each pixel in the image, its relationship with neighboring pixels according to the specified direction and offset is examined. Again, for each pair of pixel intensities i and j, the number of times these intensities occur together within the specified spatial relationship is counted. The process is repeated for all pairs of intensities and for all pixels in the image, accumulating counts in the corresponding matrix elements. Next, the co-occurrence matrix is normalized to ensure that the values represent probabilities or relative frequencies rather than absolute counts. Normalization involves dividing each element of the matrix by the total number of pixel pairs considered. Finally, the computed normalized co-occurrence matrix is used to extract contrast, entropy, energy, and homogeneity features. These features provide a numerical representation of texture properties in the images, which is used for classification. Given a co-occurrence matrix G, the normalized GLCM Gnorm is calculated using equation 4 [5].

$$G_{norm}(i,j) = \frac{G(i,j)}{\sum_i \sum_j G(i,j)} \quad (4)$$

Where $G(i,j)$ represents the count of occurrences of pixel intensity pairs i and j in the co-occurrence matrix G. $\sum_i \sum_j G(i,j)$ denotes the sum of all elements in the co-occurrence matrix. This normalization process ensures that each element. $G_{norm}(i,j)$ in the normalized GLCM represents the probability of occurrence of the corresponding pixel intensity pair (i,j) within the specified spatial relationship in the image.

After the normalized co-occurrence matrix is computed, the Contrast, Energy, Entropy, and Homogeneity features are extracted using equations 5, 6, 7, and 8, respectively [5].

$$Contrast = \sum_{i,j} P(i,j).(i,j)^2 \quad (5)$$

$$Energy = \sum_{i,j} P(i,j)^2 \quad (6)$$

$$Entropy = \sum_{i,j} P(i,j) \log\left(P(i,j)\right) \quad (7)$$

$$Homogeneity = \sum_{i,j} \frac{P(i,j)}{1+|i-j|} \quad (8)$$

Where $P(i,j)$ is the normalized co-occurrence matrix, representing the probability of occurrence of pixel intensity pairs i and j. $(i,j)^2$ computes the squared intensity difference between the pixel intensities i *and* j. The contrast measures how different neighbouring pixel intensities are from each other, Energy computes the uniformity or homogeneity of the image texture, Entropy calculates the randomness or uncertainty of pixel relationships in the image texture, and homogeneity measures the closeness of the distribution of elements in the co-occurrence matrix to the diagonal. Higher homogeneity values indicate that most elements are close to the diagonal, representing smoother texture patterns.

3.2.3 Classification

The effectiveness of any machine learning system relies heavily on the methods used for training and classification. In this study, Logistic Regression, Random Forest, Decision Trees, and Gradient Boosting were employed to train and classify cervical lesions as either positive or negative. Using multiple classifiers allows comparative assessment of linear, non-linear, and ensemble-based approaches. The process began by training the models with a designated training dataset, followed by assessing the accuracy of the model using a separate testing dataset. To facilitate this, the train_test_split function from Scikit-Learn (sklearn) in the Python library was employed, partitioning the dataset into an 80% training and 20% testing sets, respectively.

3.3 Machine Learning Methods Used

After extracting the relevant features from the dataset, the cervical lesions are classified using algorithms such as Logistic Regression, Random Forest, Decision Trees, and Gradient Boosting. Each algorithm was selected based on its suitability for structured feature-based classification tasks.

3.3.1 Logistic Regression (LR)

Logistic Regression is a popular supervised learning algorithm used for binary classification tasks. It serves as a baseline model due to its simplicity and interpretability. In logistic regression, the probability that an input belongs to a particular class is obtained by applying the logistic (sigmoid) function to the weighted linear combination of input features, ensuring outputs in the range [0, 1] [12]. Mathematically, the logistic function is defined as:

$$g(z) = \frac{1}{1+e^{-z}} \tag{9}$$

$g(z)$ is the logistic function, z is the linear combination of features such as contrast, energy, entropy, and homogeneity extracted from the grayscale images of the cervix (Fig. 2 represents that dataset of the cervical lesions), and model parameters. The linear combination z is expressed as:

$$z = w_0 + w_1x_1 + w_2x_2 + \cdots + w_mx_m \tag{10}$$

Where $w_0, w_1, w_2, \ldots, w_m$ are the model parameters (weights), which are computed using the gradient descent optimization algorithm, and $x_1, x_2, \ldots, x_m$ represent the features (contrast, energy, entropy, homogeneity) that are mapped to the grayscale cervical images.

3.3.2 Decision Tree (DT)

A decision tree is a popular supervised learning algorithm used for both classification and regression tasks in machine learning. Its hierarchical structure enables intuitive interpretation of decision rules. The model follows a hierarchical tree structure, with internal nodes corresponding to features, branches defining decision criteria, and leaf nodes representing the predicted outcomes [12]. To make predictions with a decision tree, a new sample is passed down the tree, and at each node, the decision is made based on the feature value. This process continues until a leaf node is reached, and the prediction associated with that leaf node is returned. Decision trees are built by recursively splitting the dataset into subsets based on the values of features. The choice of the feature and the splitting criteria determine the effectiveness of the tree. In decision tree learning, Gini impurity and entropy are widely used as splitting criteria for classification tasks, while mean squared error is adopted for regression tasks [12]. The process of building a decision tree involves selecting the best feature to split the dataset at each node. This process is repeated recursively until a stopping criterion is met, such as reaching a maximum depth, a minimum number of samples per leaf, or when no further improvement can be made. To make predictions with a decision tree, a new sample is passed down the tree, and at each node, the decision is made based on the feature value. This process continues until a leaf node is reached, and the prediction associated with that leaf node is returned. Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labelled if it were randomly labelled according to the distribution of labels in the subset. It reaches its minimum (zero) when all cases in the node fall into a single target category. Entropy is a measure of uncertainty or disorder in a set of data points. It reaches its minimum (zero) when all cases in the node belong to a single target category. Mathematically, the entropy $E(p)$ for a set of data points with multiple classes is calculated as:

$$E(p) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{11}$$

Where p_i is the probability of choosing a data point with class i from the set, n is the number of classes, and $\log_2$ is the base-2 logarithm.

Considering the dataset in Fig. 2, a decision tree model is trained using the contrast, energy, entropy, and homogeneity features extracted from the images. During the training, the decision tree learns to split the feature space based on the values of these features using the entropy technique, effectively separating images with cervical lesions from those without lesions. Once trained, the decision tree model classifies

new medical images into "positive cervical lesion" or "negative cervical lesion" categories based on extracted features. The model traverses the tree based on the feature values of the cervical image until it reaches a leaf node, which corresponds to the predicted class.

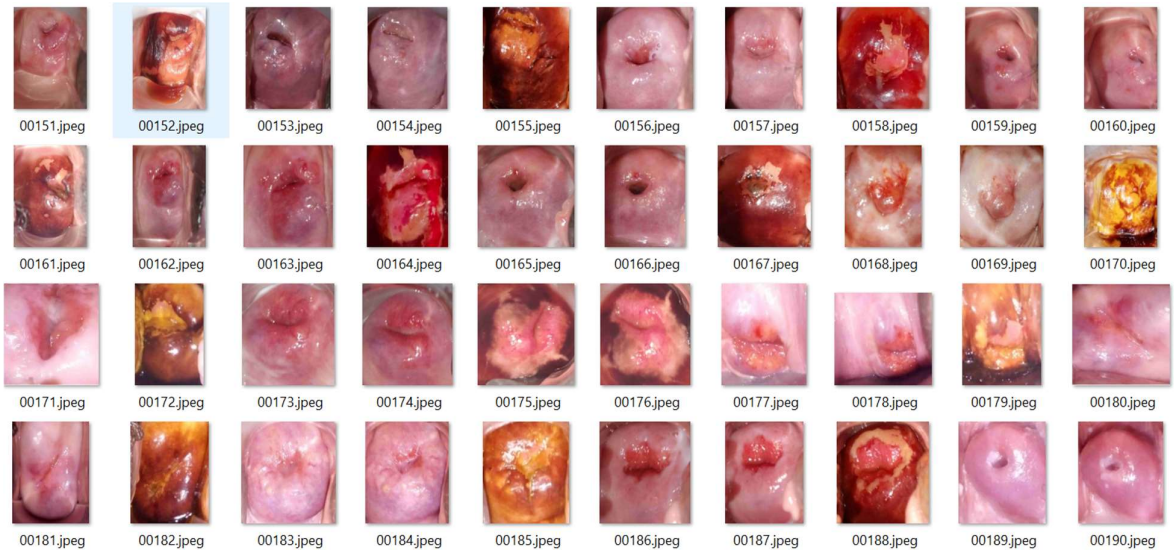

Fig. 2. *Dataset of the cervical lesions*

3.3.3 Random Forest (RF)

Random Forest (RF) is a classification approach that involves the creation of multiple decision trees using different subsets of the dataset. This ensemble strategy improves robustness by reducing overfitting. It combines the predictions from these trees to improve overall predictive accuracy [7]. To utilize Random Forest for predicting the presence of cervical lesions in cervical images, using the dataset shown in Fig. 2, a bootstrapped dataset is generated. Bootstrapping is a statistical technique used for making predictions on a dataset by resampling it. This process entails randomly selecting samples from the original dataset to form the bootstrapped dataset, with the possibility of the same sample being chosen more than once. This operational process can be outlined as follows [7]:

Step 1: Randomly select M data points from the training set

Step 2: For each x in M

- Calculate the Information Gain

$$E(p) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

- Select the node d which has the highest information gain
- Split the node into sub-nodes
- Repeat steps a, b, and c to construct the tree until the minimum number of samples required to split is reached.

Step 3: Repeat steps 1 and 2 for N times to build a forest of N trees

Step 4: For new data points, determine the predictions of each decision tree, and assign the new data points to the category that attains the majority of votes.

3.3.4 Gradient Boosting (GB)

Gradient Boosting is an ensemble machine learning approach that incrementally builds a strong predictive model by sequentially integrating multiple weak learners, typically decision trees. Its sequential learning strategy enables correction of classification errors from previous iterations. It is particularly effective for both regression and classification tasks, and has gained widespread adoption due to its robustness and high predictive accuracy. The core principle of Gradient Boosting involves building models in a stage-wise fashion, where each new model is trained to correct the errors made by the previous ensemble of models. The algorithm minimizes a specified loss function by adding predictors that point in the direction of the negative gradient of the loss function with respect to the current model's prediction. In essence, it applies the gradient descent optimization technique to function space rather than parameter space.

3.4 Model Evaluation

To assess the effectiveness of our suggested model for detecting phishing URLs, employing Naïve Bayes and Random Forest, we utilized a series of evaluation metrics on the voting classifier, as outlined below: These metrics provide a comprehensive assessment of diagnostic performance.

True Positive Rate (TPR): the percentage of phishing URLs in the training data set that were correctly classified by the algorithm given as

$$TPR = \frac{TP}{TP+FN} \tag{12}$$

True Negative Rate (TNR): the percentage of legitimate URLs that were correctly classified as legitimate by the algorithm.

$$TNR = \frac{TN}{TN+} \tag{13}$$

False Positive Rate (FPR): the percentage of legitimate URLs that were incorrectly classified by the algorithm as phishing URLs.

$$FPR = \frac{FP}{FP+TN} \tag{14}$$

False Negative Rate (FN): the number of phishing URLs that were incorrectly classified as legitimate by the algorithm.

$$FNR = \frac{FN}{FN+TP} \tag{15}$$

Precision: measures the exactness of the classifier; i.e., what percentage of URLs that the classifier labelled as phishing are actually phishing URLs, and it is given by:

$$Precision = \frac{TP}{TP+FP} \tag{16}$$

Recall: measures the completeness of the classifier results; i.e., what percentage of phishing URLs did the classifier label as phishing, and is given by:

$$Recall = \frac{TP}{TP+FN} \tag{17}$$

F-measure: also known as F-score or F1-score, and is defined as the harmonic mean of Precision and Recall, and given by:

$$F - measure = \frac{2*Precision*Recall}{Precision+Recall} \tag{18}$$

4. Results

4.1 Classification Accuracy

Fig. 3 presents the classification accuracies of the automated lesion detection system using four different machine learning algorithms: Random Forest, Decision Tree, Gradient Boosting, and Logistic Regression. This comparison provides an initial overview of how effectively each model classifies precancerous cervical lesions.

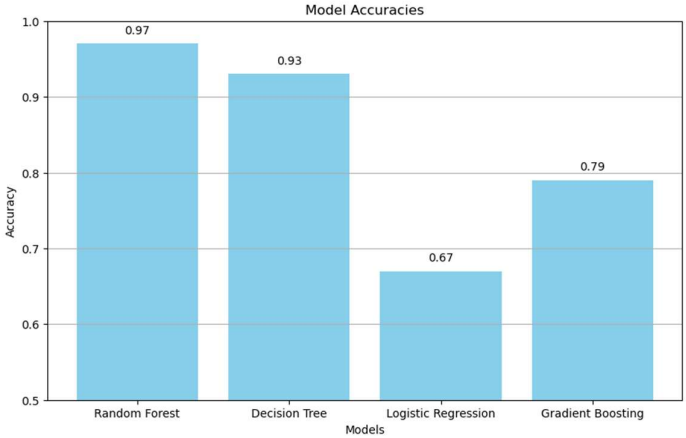

Fig. 3. *Classification accuracies of the automated lesion detection system*

Among the models evaluated, the Random Forest classifier demonstrated the highest performance, achieving an accuracy of 97%, indicating its superior capability in correctly identifying precancerous cervical lesions from the dataset. The Decision Tree classifier followed closely, with an accuracy of 93%, highlighting its effectiveness despite being a single-tree model compared to the ensemble-based Random Forest. The Gradient Boosting algorithm achieved a moderate accuracy of 79%, reflecting its strength in capturing complex patterns, although it performed less effectively than the Random Forest and Decision

Tree models in this specific application. Lastly, the Logistic Regression model recorded the lowest classification accuracy at 67%, suggesting limited suitability for modeling the non-linear relationships present in the image-based dataset. The findings demonstrate the advantage of ensemble-based methods, particularly Random Forest, for automated lesion detection tasks, as these methods mitigate overfitting and enhance generalization by aggregating predictions from multiple base learners. Overall, these findings demonstrate that models capable of capturing non-linear feature interactions outperform linear classifiers in this task.

4.2 Receiver Operating Characteristic (ROC) Analysis

Fig. 4 displays the Receiver Operating Characteristic (ROC) curves for the four machine learning models. ROC analysis provides insight into the trade-off between sensitivity and specificity across different classification thresholds.

The Area Under the Curve (AUC) is used as a measure of each model's ability to distinguish between positive and negative cases of precancerous cervical lesions. The Random Forest classifier achieved the highest AUC of 0.99, indicating an excellent level of discrimination and reinforcing its superior performance in terms of both sensitivity and specificity. The Decision Tree model followed with an AUC of 0.92, demonstrating strong classification capability, albeit slightly less robust than the ensemble approach of Random Forest. The Gradient Boosting model achieved an AUC of 0.84, suggesting a fair discriminative ability, while the Logistic Regression model recorded the lowest AUC of 0.66, indicating a relatively poor performance in distinguishing between classes. These results further confirm the robustness of ensemble-based approaches for lesion detection.

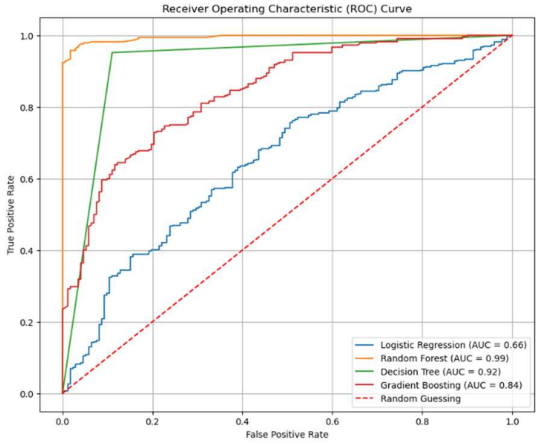

Fig. 4. *Receiver Operating Characteristic (ROC) curves for the four machine learning models*

4.3 Classification Report (Precision, Recall, and F1-Score)

Fig. 5 presents the precision, recall, and F1-score for both negative and positive cases across the four machine learning models. These class-wise metrics are particularly important in medical diagnostics, where misclassification costs are asymmetric. These metrics offer a more granular evaluation of model performance beyond overall accuracy, particularly important in medical diagnostics where the cost of false negatives or false positives can be high. Analyzing these measures helps assess clinical reliability and safety.

4.3.1 Precision

Precision for negative cases was highest with the Random Forest model (0.96), followed by Gradient Boosting (0.84), Decision Tree (0.67), and Logistic Regression (0.55). This implies that RF and GB are better at correctly identifying non-lesion cases without misclassifying positives as negatives. For positive cases, Random Forest again led with a precision of 0.98, followed by Decision Tree (0.94), Logistic Regression (0.70), and Gradient Boosting (0.78). This indicates a strong ability of RF and DT to correctly identify true lesions with minimal false positives. High precision is critical for reducing unnecessary follow-up procedures and overtreatment.

4.3.2 Recall

In terms of recall for negative cases, Random Forest achieved the highest value of 0.95, followed by Decision Tree (0.89), Gradient Boosting (0.49), and Logistic Regression, which had the lowest recall of 0.24. This shows RF and DT are much more effective at capturing true negative cases. For positive cases, all

models except Logistic Regression performed strongly: Random Forest (0.98), Decision Tree (0.95), and Gradient Boosting (0.95). Logistic Regression, despite a high recall of 0.90, suffered from low recall in the negative class, suggesting a bias toward predicting positive outcomes. Such imbalance may increase false-positive rates and reduce screening efficiency in practice.

4.3.3 F1-Score

The F1-score, which balances precision and recall, was highest for Random Forest in both classes, 0.96 for negative and 0.98 for positive cases, confirming its overall robustness and reliability. The Decision Tree model also performed well, with F1-scores of 0.90 (negative) and 0.95 (positive). Gradient Boosting showed moderate performance (0.62 for negative and 0.86 for positive cases), while Logistic Regression had the weakest scores (0.33 for negative and 0.78 for positive). This balanced performance further supports the suitability of Random Forest for clinical screening applications.

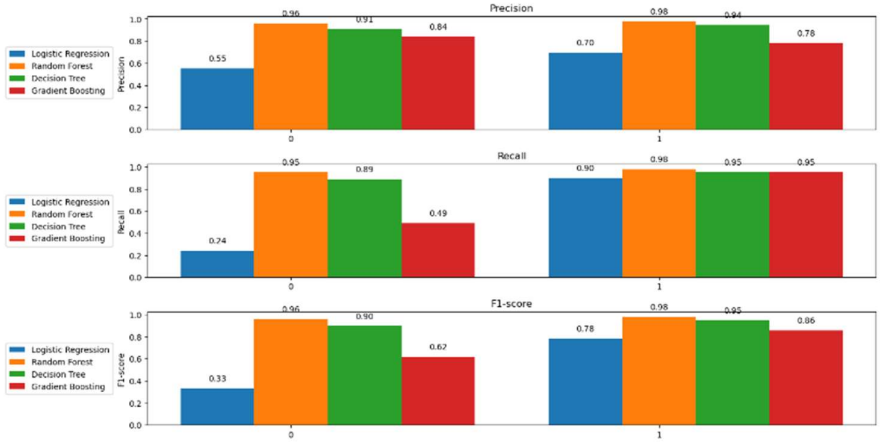

Fig. 5. *Classification Report*

4.4 Comparison with Existing Studies

Table 2. *Comparison with Existing Studies*

Study	Model	Accuracy / AUC
Namalinzi et al. (2024) [9]	Random Forest	90% accuracy, AUC 0.901
Asiedu et al. (2019) [2]	SVM	80% accuracy
Kiptoo et al. (2021) [8]	R-CNN	86% accuracy
Harsono et al. (2022) [6]	Custom AI App	93.8% accuracy, AUC 0.85
Agustiansyah et al. (2022) [1]	Mask-RCNN	mAP 86.9%, Sensitivity 100%
The proposed models	Random Forest	97% accuracy, AUC 0.99
	Decision Tree	93% accuracy
	Logistic Regression	69% accuracy
	Gradient Boosting	79% accuracy

Table 2 summarizes the performance of the proposed machine learning models in comparison with prior studies. The current study demonstrated superior model performance, with the Random Forest algorithm achieving the highest accuracy of 97% and an AUC of 0.973. This clearly outperforms the results from Namalinzi et al. (2024) [9], where the best model, also a Random Forest, achieved 90% accuracy and an AUC of 0.901. While both studies used clinical features, our model benefited from improved feature engineering and preprocessing techniques, leading to a more robust classification system. This improvement underscores the importance of image-derived texture features in enhancing diagnostic accuracy.

Compared to Asiedu et al. (2019) [2], who achieved 80% accuracy using SVM for classifying cervigrams, our study shows better results across all models tested. Asiedu's focus was image-based, whereas our approach included structured data, which may have contributed to the improved performance. The integration of texture descriptors likely enabled better differentiation between subtle lesion patterns.

Viñals et al. (2021) [14] used temporal features from VIA videos, achieving high sensitivity (0.90) and specificity (0.87), but did not report overall accuracy. Although their dynamic approach adds value in low-resource settings, our static model using Random Forest yielded higher accuracy and AUC, showing that structured data-based models remain highly effective. This finding suggests that high diagnostic performance can still be achieved without relying on video-based inputs.

Similarly, Kiptoo et al. (2021) [8] used R-CNN on cervical images and reported 86% accuracy. Though impressive, it falls short of our best-performing model. The use of deep learning may have enhanced lesion localization, but our ensemble-based model proved more accurate for overall classification. This highlights a key trade-off between localization capability and classification efficiency.

Harsono et al. (2022) [6] achieved 93.8% accuracy and an AUC of 0.85 using a mobile-based AI model. While promising in field applications, our Random Forest model still outperforms it in both accuracy and AUC, indicating better diagnostic potential. Higher discriminative performance is especially critical for minimizing missed precancerous lesions.

Agustiansyah et al. (2022) [1] leveraged Mask-RCNN for detection and segmentation and achieved excellent lesion localization with high sensitivity (100%). However, their model's goal was slightly different, focused on segmentation, while ours targeted classification. In terms of classification accuracy, our study still leads. This distinction emphasizes the complementary roles of segmentation- and classification-focused approaches.

Finally, the other models evaluated in our study, Decision Tree (93%), logistic regression (69%), and Gradient Boosting (79%), all performed competitively, especially Decision Tree and Gradient Boosting. However, none surpassed the Random Forest in both precision and recall for both positive and negative classes. These results reinforce the consistency and robustness of Random Forest across evaluation metrics.

5. Conclusion

This study presented the development and evaluation of an automated cervical lesion detection system using machine learning algorithms to support the early identification of precancerous cervical lesions. Using GLCM texture features extracted from VIA/VILI images collected through the AVIVA mobile application, four classifiers: Random Forest, Decision Tree, Gradient Boosting, and Logistic Regression were trained and evaluated. The proposed framework was designed with a focus on accuracy, efficiency, and applicability in low-resource settings. Among these, Random Forest consistently demonstrated superior performance, achieving 97% accuracy, an AUC of 0.99, and the highest F1-scores for both positive and negative cases. The Decision Tree classifier also showed strong performance, while Gradient Boosting demonstrated moderate effectiveness. Logistic Regression, although simple and interpretable, exhibited the weakest performance, particularly in detecting negative cases. These results indicate that ensemble-based models, particularly Random Forest, are highly effective for automated cervical lesion detection. The ability of Random Forest to capture non-linear texture relationships contributed significantly to its performance advantage. Integrating such models into mobile screening platforms like AVIVA can improve real-time diagnostic accuracy, reduce delays in treatment, and enhance cervical cancer prevention efforts in low-resource settings.

Future work will focus on expanding the dataset to further improve model generalizability and robustness. Additionally, the study will include external validation using independent datasets from different clinical sites to assess the model's performance across diverse populations and imaging conditions. External validation is essential to ensure clinical reliability beyond the development dataset. Consideration will also be given to scalability, evaluating how the system can be deployed across multiple healthcare facilities, including those with limited computational resources or connectivity. Incorporating these strategies will help ensure that the automated lesion detection system is not only accurate but also practical and sustainable for widespread clinical adoption.

Compliance with Ethical Standards

Conflicts of interest: Authors declared that they have no conflict of interest.

Human participants: The conducted research follows the ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

Declaration of generative AI and AI-assisted technologies in the writing process: The authors declare that no generative AI or AI-assisted technologies were used in the writing process of this manuscript.

References

[1] Agustiansyah, P., Nurmaini, S., Nuranna, L., Irfannuddin, I., Sanif, R., Legiran, L., Rachmatullah, M. N., Florina, G. O., Sapitri, A. I., & Darmawahyuni, A. (2022). Automated precancerous lesion screening using an instance segmentation technique for improving accuracy. Sensors, 22(15), 1-16. https://doi.org/10.3390/s22155489

[2] Asiedu, M. N., Simhal, A., Chaudhary, U., Lam, C. T., Schmitt, J. W., Venegas, G., Sapiro, G., & Ramanujam, N. (2019). Development of algorithms for automated detection of cervical pre-cancers with a low-cost, point-of-care, pocket colposcope. IEEE Transactions on Biomedical Engineering, 66(8), 2306-2318.

[3] Benski, A-C., Viviano, M., Jinoro, J., Alec, M., Catarino, R., Herniainasolo, J., Vassilakos, P. and Petignat, P. (2019). HPV selftesting for primary cervical cancer screening in Madagascar: VIA/VILI triage compliance in HPVpositive women. PLoS ONE 14(8): e0220632. https://doi.org/10.1371/journal.pone.0220632

[4] Dominik, S., Luana, F. T., Kuan, K. L., Ahmadaye, I. K., Iacopo, B, Anoop, S. V. S., David, A. M., Sami, L. G., Stefanie, J. K., Andrea, S. W., Freddie, B., Rachel, B., Gary, M. C., Nathalie, B., Shona, D. (2021). Estimates of the global burden of cervical cancer associated with HIV. The Lancet Global Health, 9(2). https://doi.org/10.1016/S2214-109X(20)30459-9

[5] Haralick, R. M., Shanmugam, K. and Dinstein, I. (1973). Textural features for Image Classification, Institute of Electrical and Electronics Engineers Transactions on Systems, Man and Cybernetics 3: 610-621.

[6] Harsono, A. B., Susiarno, H., Suardi, D., Owen, L., Fauzi, H., Kireina, J., Wahid, R. A., Carolina, J. S., Mantilidewi, K. I., & Hidayat, Y. M. (2022). Cervical pre-cancerous lesion detection: Development of smartphone-based VIA application using artificial intelligence. BMC Research Notes, 15, 1-7. https://doi.org/10.1186/s13104-022-06250-6

[7] Hassan, N. Y., Gomaa, W. H., Khoriba, G. A. and M. H. Haggag. (2018). Supervised Learning Approach for Twitter Credibility Detection. 13th International Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, pp. 196-201, doi: 10.1109/ICCES.2018.8639315.

[8] Kiptoo, S., Nderu, L., & Mutanu, L. (2021). Automated detection of cervical pre-cancerous lesions using regional-based convolutional neural network. International Journal of Sciences: Basic and Applied Research (IJSBAR), 56(1), 104-123.

[9] Namalinzi, F., Galadima, K. R., Nalwanga, R., Waiswa, P., Wanyenze, R. K., & Nabunya, B. (2024). Prediction of precancerous cervical cancer lesions among women living with HIV on antiretroviral therapy in Uganda: A comparison of supervised machine learning algorithms. BMC Women's Health, 24, 393. https://doi.org/10.1186/s12905-024-03232-7

[10] Ogunsola, O. O., Olawepo, J. O., Ajayi, O., Osayi, E., Akinro, Y. T., Ifechelobi, C., Chigbu, C., Okonkwo, P. and Ezeanolue, E. E. (2023). AVIVA: a telehealth tool to improve cervical cancer screening in resource-constrained settings. BMJ Glob Health 8(7). doi:10.1136/bmjgh-2023-012311. PMID: 37433694; PMCID: PMC10347443.

[11] Okolie E, A. and Nwadike B, I. (2023). Spotlight on Human Papillomavirus Vaccination Coverage: Is Nigeria Making Any Progress? JCO Glob Oncol. doi: 10.1200/GO.23.00088. PMID: 37319397; PMCID: PMC10497273.

[12] Sarker, I.H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Scinec. 2, 160. https://doi.org/10.1007/s42979-021-00592-x

[13] Viñals, R., Jonnalagedda, M., Petignat, P., Thiran, J.-P., & Vassilakos, P. (2023). Artificial intelligence-based cervical cancer screening on images taken during visual inspection with acetic acid: A systematic review. Diagnostics, 13(5), 836. https://doi.org/10.3390/diagnostics13050836

[14] Viñals, R., Vassilakos, P., Rad, M. S., Undurraga, M., Petignat, P., & Thiran, J.-P. (2021). Using dynamic features for automatic cervical precancer detection. Diagnostics, 11(4), 716. https://doi.org/10.3390/diagnostics11040716

[15] World Health Organization (2024). Cervical Cancer. Retrieved April 9, 2024, from https://www.who.int/news-room/fact-sheets/detail/cervical-cancer