International Journal on Engineering Technology (InJET)

Volume 3, issue no. 1, Dec 2025, Page 207 - 220
Received Date: Aug 19, 2025
Accepted Date: Sep 10, 2025
ISSN: 3021-940X (Print)
DOI: https://doi.org/10.3126/injet.v3i1.87023

Beyond Accuracy and Classification: XAI Driven Interpretability in Cervical Cancer

Saugat Kafle[1], Prakash Paudel[2], Mohan Bhandari[1*]

[1] *Department of Computer Science, Samriddhi College, Bhaktapur, Nepal, saugatkafle77@gmail.com*
[*1] *Department of Computer Science, Samriddhi College, Bhaktapur, Nepal,mail2mohan@gmail.com*
[2] *Department Of IT, NCIT, Lalitpur, Nepal, mail2mohanbhandari@gmail.com*

Abstract

Cervical cancer, a prevalent malignancy linked to HPV infection, necessitates accurate and timely diagnosis to mitigate its high mortality rate. Traditional diagnostic methods, such as Pap smears and colposcopy, are often laborious and subjective, highlighting the need for advanced computational approaches. This study covers machine learning (ML) to enhance cervical cancer detection, evaluating models including Multi- Layer Perceptron (MLP), Gaussian Naive Bayes (GaussianNB), Bagging, Random Forest (RF) and K-Nearest Neighbors (KNN). The MLP classifier achieved 99.59% accuracy, while other algorithms surpassed 97% AUC, underscoring their clinical viability. To ensure interpretability, Explainable AI (XAI) techniques – SHAP, LIME and ELI5, are used, explaining feature contributions and decision pathways, thus nurturing clinician trust. The integration of high-accuracy ML models with transparent XAI frameworks not only improves diagnostic precision, but also facilitates the ethical deployment of AI in healthcare, paving the way for reliable, data-driven clinical decision making.

Keywords: Cervical cancer, eXplainable AI, LIME, SHAP, ELI5

1. Introduction

Cervical cancer, the fourth most common cancer among women worldwide, is caused almost entirely by human papillomavirus (HPV) (Fowler, et al., 2017).HPV is the common infection that is passed through sexual contact. There are two main types of cervical cancer: Squamous Cell Carcinoma and Adenocarcinoma. The most common type is Squamous Cell Carcinoma, which originates in the thin, flat Squamous Cells lining the cervix and accounts for about 70% of cases. Adenocarcinoma, which begins in the column-shaped glandular cells lining the cervical canal, is less common (about 25% of cases) and more difficult to diagnose because it develops higher up in the cervix.

Cervical cancer, a malignant tumor arising in the cervix, is primarily caused by persistent infection with high-risk strains of the human papillomavirus (HPV) (Cohen, et al., 2019). Cervical cancer is the fourth most frequently occurring malignancy in women, and results in an estimated 530,000 new cases annually with 270,000 deaths (Cohen, et al., 2019).It is responsible for damaging deep tissues of the cervix and can gradually reach other areas of the human body, such as the lungs, liver, and vagina, which can increase the difficulties involved (Ghoneim, et al., 2020).

Recently, many studies have been conducted on cervical cancer using modern techniques that provide prediction in the early stage. Using machine learning (ML) has contributed to early prediction. Using ML has contributed to early prediction (Osuwa & öztoprak, 2021).ML has improved the performance of analyses and the generation of accurate patient data (Mudawi & Alazeb, 2020).

Regions such as sub-Saharan Africa, Central America, and South-East Asia bear the highest burden, with profound societal consequences, including reduced workforce productivity, increased healthcare costs, and

* *Corresponding author*

emotional distress for affected families (Asadi, et al., 2020). On an individual level, cervical cancer disrupts quality of life, often progressing silently until advanced stages, underscoring the need for early detection through regular screening (Canavan, et al., 2000).

ML is a specific artificial intelligence (AI) branch that collects data from training data (Alsmariy, et al., 2020).However, accurate prediction of survival of cervical cancer patients is still challenging due to the heterogeneity of the cancer cells (Ding, et al., 2021). Incorporating a more evenly distributed set of training data allows these algorithms to better predict outcomes from novel data and set limits for their development. To adapt to different data distributions, quickly manage large datasets, and fine-tune algorithmic parameters, the oversampling approach employs powerful machine learning skills (Mujahid, et al., 2024). The objectives of the proposed study are:

- To develop an interpretable ML model for predicting cervical cancer using patient health data.
- To apply XAI techniques such as SHAP, LIME (Local Interpretable Model-Agnostic Explanations), and ELI5 (Explain Like I'm 5) to explain individual predictions and identify key influencing features.
- To enable greater trust and transparency for healthcare professionals in model-driven diagnostic support.

2. Literature Review

Recent research highlights the growing role of ML in cervical cancer prediction. Alsmariy et al. (2020) integrated SMOTE and PCA with ensemble voting, achieving 98.49% accuracy for Schiller test classification (Alsmariy, et al., 2020). Mehmood et al. (2021) developed CervDetect, a hybrid Random Forest (RF) and neural network model, which attained 93.6% accuracy with a low false-positive rate (6.4%) (Mehmood, et al., 2021). Deng et al. (2018) compared Support Vector Machines (SVM), XGBoost, and RF, identifying the latter two as superior for diagnosing cervical cancer based on four target variables (Hinselmann, Schiller, Cytology, and Biopsy) (Deng, et al., 2018) . Uddin et al. (2025) proposed an ensemble framework combining RF and Logistic Regression, achieving 99.75% accuracy while leveraging SHAP and LIME or interpretability (Uddin, et al., 2025). Chadaga et al. (2022) further advanced the field with a model yielding 98% accuracy and 100% AUC, validated through XAI techniques like SHAP and ELI5 (Chadaga, et al., 2022).Collectively, these studies underscore AI's potential to enhance early detection, particularly in underserved populations. Hasan et al. utilized the RF classifier in com- bination with the SMOTETomek data balancing technique (Synthetic Minority Oversampling Technique with Tomek links), which significantly enhanced the model's performance. Their approach achieved an impressive 99.85% accuracy along with perfect scores 100% precision, recall, and F1- score (Hasan, et al., 2022). Shakil et al. demonstrated that the DT model achieved superior performance when combined with Chi-square feature selection, reporting impressive results of 97.60% accuracy, 98.73% sensitivity, 80% specificity, and 98.73% precision. Even under data imbalance conditions, the DT maintained strong performance, achieving 97% accuracy, 99.35% sensitivity, 69.23% specificity, and 97.45% precision (Shakil, et al., 2024). Asadi et al. applied the Quest and C&R tree algorithms, achieving accuracy, sensitivity, specificity, and AUC values of 95.55%, 90.48%, 100%, and 95.20%, respectively. For the RBF model, these metrics were 95.45%, 90.00%, 100%, and 91.50%, while the SVM produced 93.33%, 90.48%, 95.83%, and 95.80%. The MLP model, in comparison, obtained 90.90%, 90.00%, 91.67%, and 91.50% (Asadi, et al., 2020). Tanimu et al. employed a DT model using features selected through Recursive Feature Elimination (RFE) in combination with the SMOTETomek sampling technique. This approach significantly improved classification performance, yielding an accuracy of 98.72% and a sensitivity of 100%, indicating that the model was highly effective in correctly identifying all positive cases (Tanimu, et al., 2022).

3. Methodology

The ML models and explainable AI (XAI) techniques were implemented using Python, including LIME, SHAP, and ELI5 to enhance model interpretability. The experiments were utilized on Google Colab, using a Google provided NVIDIA K80 GPU and 12 GB of RAM. The development environments included Python 3.7, Keras 2.5.0, and TensorFlow 2.5.0. SHAP provides both local and global interpretability for the model's predictions, and LIME analyzes feature im- portance, offering both local and global interpretability for the model's predictions. ELI5 was also used to illustrate and describe the model's internal decision-making process. The Fig 1 summarized the workflow of this study:

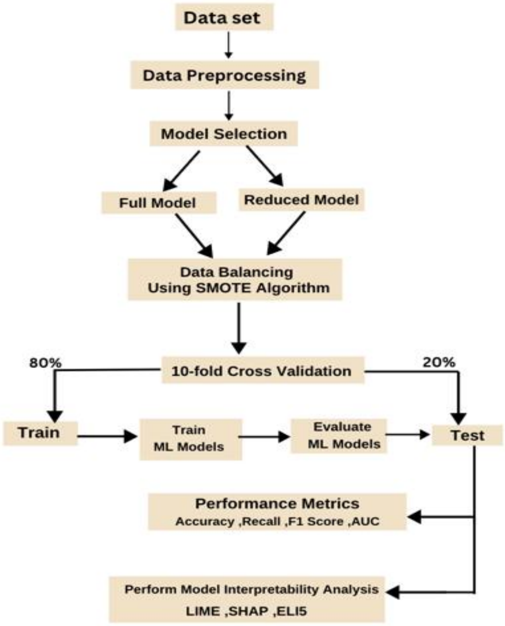
Data set
Data Preprocessing
Model Selection
Full Model
Reduced Model
Data Balancing Using SMOTE Algorithm
80%
10-fold Cross Validation
20%
Train
Train ML Models
Evaluate ML Models
Test
Performance Metrics Accuracy ,Recall ,F1 Score ,AUC
Perform Model Interpretability Analysis LIME ,SHAP ,ELI5

Figure 1. Overview of Data Processing, Modeling, and Evaluation Pipeline

3.1 Dataset Details

3.1.1 Description

The dataset employed in this study was sourced from a publicly available repository on Kaggle (Aziz, 2025).This dataset includes information from 835 patients who underwent various medical tests and health assessments. Its purpose is to assess the variables that could affect a cervical cancer diagnosis. The dataset includes several clinical and demographic characteristics linked to the people. The primary objective is to classify whether a patient is diagnosed with cervical cancer (Cancer) or not (Non-Cancer).

3.2 Statistical Overview

3.2.1 Number of Entries

The dataset consists of a total of 835 entries for most features, with a few features having fewer entries due to missing data (e.g., STDs: Time since first diagnosis has 71 entries).

3.2.2 Numerical Data

The dataset contains several continuous variables, such as Age, Number of sexual partners, and First sexual intercourse, which can take on a wide range of values. Certain temporal features, such as STDs: Time since first diagnosis and STDs: Time since last diagnosis, have been represented in years to maintain consistency and interpretability.

3.2.3 Categorical Data

The dataset includes a substantial number of binary categorical variables. These represent the presence or absence of specific conditions or behaviors — for instance, whether the patient smokes, uses hormonal contraceptives, or has been diagnosed with particular types of sexually transmitted diseases (STDs). Binary encoding is applied to such features, where 1 indicates “Yes” and 0 indicates “No.”

3.2.4 Data Preprocessing and Data Balancing

The dataset used in this study is obtained from a publicly accessible Kaggle repository and contains labeled samples suitable for supervised learning tasks. Although the data originates from a single source, it sufficiently captures variability within the problem domain, providing a meaningful representation of relevant clinical and demographic attributes. To ensure consistency and reliability, records containing missing values are excluded, and upsampling techniques are employed to address the inherent class imbalance, enabling more robust model training and performance evaluation.

Prior to model development, several preprocessing steps are performed to enhance data quality and analytical integrity. Instances with missing information are eliminated to maintain completeness. In particular, the attributes "STDs: Time since first diagnosis" and "STDs: Time since last diagnosis" contain substantial missing entries and are therefore removed to prevent potential bias or distortion in model outcomes. The full model includes all available clinical, demographic, and behavioral predictors potentially influencing cervical cancer diagnosis, whereas the reduced model considers only the most significant predictors identified through statistical analysis and feature selection, emphasizing variables that contribute most strongly to predictive performance. After separating the target variable ("Dx:Cancer") from the feature set, remaining missing values are replaced with the median of each respective column to ensure internal consistency. The dataset is then divided into training (80%) and testing (20%) subsets using a stratified split to preserve class proportions across both sets. The Hyperparameter values is selected empirically, and a fixed random seed (42) is used for all models. To bring all features onto a common scale, the StandardScaler method is applied, fitting only on the training data and subsequently transforming the test data to avoid any form of data leakage.

Given the highly imbalanced nature of the dataset, the RandomOverSampler technique is applied exclusively to the training set to balance class distribution. This process synthetically increases the minority class (cancer cases) by duplicating existing samples until both classes contain 817 instances each, resulting in a balanced dataset of 1,634 samples. While this resampling approach improves class representation, it inherently carries a risk of overfitting, since the duplicated samples do not introduce new information. To mitigate this issue, model performance is evaluated using the unseen test set, which retains the original class distribution. This ensures that the evaluation metrics accurately reflect the model's true generalization capability rather than memorization of synthetic data.

4. Related Theory

A diverse collection of ML models was employed to construct a robust classification system for diagnosing cervical cancer. By leveraging the unique strengths of various algorithms, the system enhances predictive accuracy and overall model reliability. The models used in this study include MLP, RF, KNN, GaussianNB, and the Bagging classifier contributing distinct capabilities to the classification task.

The MLP is the most known and most frequently used type of neural network (Popescu, et al., 2009).The MLP is well-suited for detecting complex, non-linear patterns in data. The MLP model configured with a hidden layer of five neurons and trained for 300 iterations effectively captures complex relationships among the clinical and demographic features. The MLP neural networks provide a lot of flexibility and have proven useful and reliable in a wide range of classification and regression problems (Bataineh, et al., 2022).

RF are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Biau & Scornet, 2016). It works by building multiple decision trees and averaging their predictions. This reduces overfitting and variance, making the model more reliable and better at distinguishing between cancerous and non-cancerous cases.

The KNN algorithm classifies data points by looking at the nearest neighbors in the dataset. In our case, with 15 neighbors, KNN groups patients with similar characteristics together. KNN classification has the remarkable property that, under very mild conditions, the error rate of a KNN algorithm tends towards being Bayes optimal as the sample size tends towards infinity (Lei, et al., 2009).

The Gaussian NB is a variant of Naive Bayes that follows Gaussian normal distribution and supports Continuous data (Sawant & Khadapkar, 2022).It is based on the assumption that the features follow a normal (Gaussian) distribution. With a variance smoothing parameter, this model reduces sensitivity to noise and outliers. It's particularly useful when the features are conditionally independent, making it an efficient and probabilistic method for classifying data.

Finally, the Bagging model, which stands for Bootstrap Aggregating, helps reduce overfitting by training multiple base models on different subsets of the data and combining their results. Bagging improves when probabilistic estimates in conjunction with no-pruning are used, as well as when the data was backfit (Bauer & Kohavi, 1999).

Together, these models form an ensemble system that balances strength and diversity, ensuring improved performance by reducing bias and variance. The MLP captures complex, non-linear relationships, the Random Forest handles feature interactions and reduces overfitting. KNN utilizes local data structures, GaussianNB provides probabilistic classification and Bagging stabilizes predictions. By combining these models, the final system becomes robust, efficient, and capable of making highly accurate predictions about whether a patient has cervical cancer or not.

4.1 Model Performance and Performance Metrics

This study evaluates machine learning models using key performance metrics. Accuracy measures the proportion of correctly classified cases. Precision shows how many predicted positive cases are actually correct, while recall indicates how well the model identifies actual positives. The F1-score balances precision and recall, providing a reliable measure of overall performance (Alvarez, 2002). This study employs specific performance metrics to assess the effectiveness of the ML models.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \qquad \text{(Equation 1)}$$

$$\text{Precision}=\frac{TP}{TP+FP} \qquad \text{(Equation 2)}$$

$$\text{Recall}=\frac{TP}{TP+FN} \qquad \text{(Equation 3)}$$

$$F_1\text{Score}=\frac{Precision*Recall}{Precision+Recall} \qquad \text{(Equation 4)}$$

In this study, performance is evaluated using key metrics: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These values help calculate accuracy, precision, recall, and F1-score.

SHAP, LIME and ELI5 were used to improve the interpretability of the ML models and provide insights into their decision-making processes. SHAP helped to determine the contribution of each feature to the model’s predictions by assigning importance values. LIME was employed to generate interpretable surrogate models for specific predictions, offering clearer explanations of how certain inputs affected the outcomes. ELI5 was used to simplify the explanation of feature importance and model behavior, enhancing the transparency of the model’s decisions. Together, these methods helped to build a better understanding of the models, fostering greater trust in their predictions.

5. Result and Discussion

5.1 Overall Performance of Machine Learning Models

The evaluation metrics of five machine learning algorithms GaussianNB, Bagging, KNN, RF, and MLP illustrate their strengths and limitations based on accuracy, precision, recall, F1-score.

The GaussianNB achieves 97.14% accuracy with perfect recall but slightly lower precision (94.44%), meaning it identifies all positive cases correctly but misclassifies some negative cases as positive, resulting in false positives. Bagging and KNN perform comparably, achieving 97% accuracy and an AUC of 1.00, demonstrating strong overall classification. However, KNN can be sensitive to noisy data and requires careful parameter tuning.RF, while generally robust against noise and overfitting, performs slightly worse with 94% accuracy and an AUC of 0.94, suggesting that the model may not fully capture the subtle patterns in this dataset. The MLP model shows near-perfect performance with 99.54% accuracy, 0.9919 precision, 100% recall, and an AUC of 1.00, highlighting its strong capability to model complex nonlinear relationships. Nevertheless, such near-perfect results may indicate overfitting, and the model’s generalization to entirely new clinical data must be interpreted cautiously.

Compared to previous studies, our MLP model achieves 99.54% accuracy, outperforming Alsmariy et al. with 98.49% accuracy and Mehmood et al. having 93.6% accuracy, and is nearly comparable to Uddin et al. of 99.75% accuracy. While Alsmariy et al. relied on SMOTE and PCA combined with ensemble voting, and Mehmood et al. used a hybrid RF–neural network, our approach attains strong performance by focusing on careful preprocessing, median imputation, scaling, and RandomOverSampler applied only on the training set, without the need for dimensionality reduction or complex ensembles. This allows the model to learn robust patterns directly from the available features while minimizing the risk of data leakage or overfitting. Moreover, our approach ensures interpretability through methods like LIME and SHAP, providing clear insights into feature contributions for cervical cancer prediction. Overall, these results demonstrate that our MLP model is among the best-performing approaches on this dataset, achieving a strong balance between accuracy, generalization, and interpretability.

Overall, the differences reflect dataset characteristics, preprocessing choices, model selection, and evaluation protocols, highlighting the importance of consistent methodology for reliable clinical predictive modeling.

Table 1. Performance Metrics of Different Algorithms

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
GaussianNB	0.9714	0.9444	1.0000	0.9714	1.0000
Bagging	0.9700	0.9700	0.9700	0.9700	1.0000
KNN	0.9700	0.9700	0.9700	0.9700	1.0000
RF	0.9400	0.9500	0.9400	0.9400	0.9400
MLP	0.9959	0.9919	1.0000	0.9950	1.0000

5.2 XAI Insights

5.2.1 Baggier Classifier

a) LIME:

As shown in Figure 2, the Bagging model reveals that the prediction toward Cervical Cancer (67% probability) is mainly influenced by several diagnostic and sexually transmitted disease (STD) related features. A positive diagnosis of Human Papillomavirus (Dx:HPV) strongly contributes to the "No Cancer" prediction, aligning with clinical expectations that HPV negativity reduces cancer risk. Conversely, positive findings in general diagnosis (Dx) and Cervical Intraepithelial Neoplasia (Dx:CIN) push the model's prediction toward "Cancer," consistent with early precancerous cellular changes. Additional factors like use of intrauterine devices (IUD) and histories of pelvic inflammatory disease (STDs:PID) or molluscum contagiosum (STDs:MOL) play moderate roles, indicating how reproductive and infection-related variables collectively influence cervical cancer likelihood.

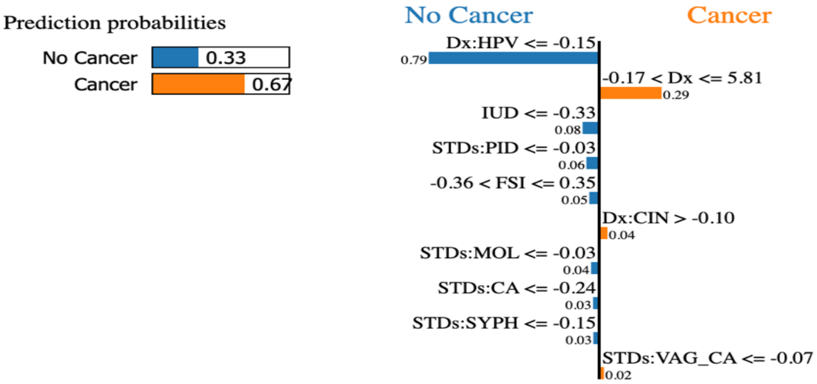

Figure 2. Bagging LIME Tabular Plots

b) ELI5:

Table 2 shows the feature importance of the Bagging Classifier using ELI5. Among the evaluated factors, a history of STDs such as syphilis (0.012) and HIV (0.009) were the most influential in predicting cervical cancer, highlighting the known association between persistent infections and cervical cellular changes. Cervical intraepithelial neoplasia (CIN) and HPV infection had lower importance (0.004 and 0.002, respectively), suggesting that while they contribute to risk, the model emphasizes previous STD exposure as a stronger predictor in this dataset. Interestingly, age at first sexual intercourse had negligible importance (0.000), indicating that in this cohort, early sexual activity alone was less predictive than documented STD history.

Table 2. Explaining Bagging Classifier with ELI5

Feature	Importance
STDs: syphilis	0.0120

STDs:HIV	0.0087
STDs:CIN	0.0042
STDs HPV:	0.0024
First sexual intercourse	0.0000

c) SHAP

Figure 3 shows the SHAP summary plot for the Bagging model indicates the most influential features are Age and Number of sexual partners, with SHAP values ranging between approximately -0.1 and 0.2 for Age and -0.3 and 0.3 for Number of sexual partners. The negative SHAP values for Age indicate that younger individuals tend to reduce the model's output, whereas the mixed SHAP values for Number of sexual partners suggest that an increase in partners can either increase or decrease the prediction depending on other factors. Other features such as Dx:CIN, Smokes, Num of pregnancies, and First sexual intercourse have minimal impact, as their SHAP values are clustered around zero, indicating they do not contribute significantly to the prediction.

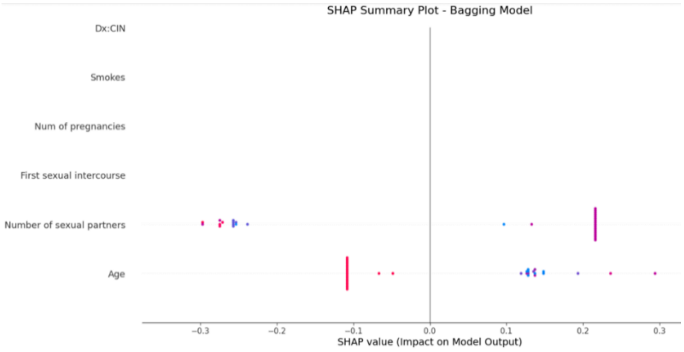

Figure 3. SHAP Analysis For Bagging Classifier

5.2.2 K-Nearest Neighbors Classifier

a) LIME

As shown in Figure 4 the KNN model shows a clear prediction toward the No Cancer class with a probability of 1.0. The model identifies the absence of Human Papillomavirus (STDs:HPV) as the strongest indicator of no cervical cancer, consistent with the clinical understanding that HPV infection is the primary cause of cervical abnormalities. Additionally, fewer years of Hormonal Contraceptive use (HC) and the absence of an Intrauterine Device (IUD) further reinforce the "No Cancer" prediction, as prolonged exposure to such reproductive factors has been associated with cervical tissue changes. Minor opposing influences, such as molluscum contagiosum (STDs:MOL) and condylomatosis (STDs:CA), slightly contribute toward the cancer class, reflecting their potential but weaker associations with cervical pathology.

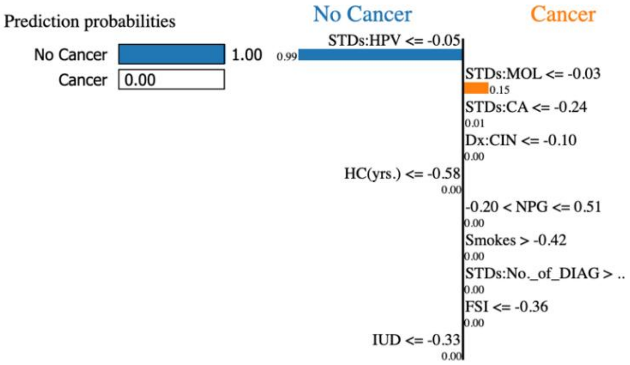

Figure 4. KNN LIME Tabular Plots

b) ELI5

Table 3 presents the feature importance of the KNN Classifier using ELI5. The model identifies a confirmed HPV diagnosis (0.202) and cytological diagnosis (0.153) as the most significant predictors of cervical cancer, reflecting the central role of HPV infection and abnormal cell findings in disease development. Smoking intensity (0.045) and history of STDs including HPV (0.039) also contribute, supporting the known link between lifestyle factors, persistent infections, and cancer risk. CIN diagnosis had minimal importance (0.005), suggesting that in this dataset, prior cellular changes were less influential than direct HPV infection and cytology results.

Table 3. Explaining KNN Classifier with ELI5

Feature	Importance
Dx: HPV	0.2020
Dcx	0.1525
Smokes (packs/year)	0.0448
STDs: HPV	0.0389
Dx:CIN	0.0054

c) SHAP

Figure 5 shows The SHAP summary plot for the KNN model indicates the SHAP values that highlight Age and Number of sexual partners as the dominant factors, with SHAP values ranging between -0.2 and 0.4 for Age and -0.4 and 0.3 for Number of sexual partners. The greater spread in the SHAP values for Age compared to the other models suggests that KNN relies more on this feature for making predictions. The remaining features, including Dx, Dx:CIN, Smokes, Num of pregnancies, and others contribute very little to the model's output, as indicated by their SHAP values close to zero.

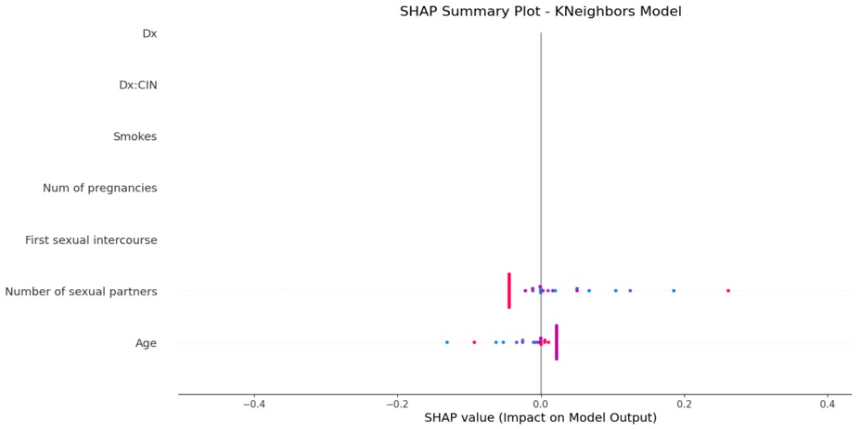

Figure 5. SHAP Analysis for KNN Classifier

5.2.3 MultiLayer Perceptron Classifier

a) LIME

The LIME explanation in Figure 6, shows that the MLP model shows a Cancer prediction probability of 0.99 and No Cancer probability of 0.01. Key contributors to the Cancer outcome include Diagnosis (Dx), cervical intraepithelial neoplasia (Dx:CIN), pelvic inflammatory disease (STDs:PID), molluscum contagiosum (STDs:MOL), and HIV infection (STDs:HIV), all of which are medically linked to increased cervical cancer risk. In contrast, lower human papillomavirus infection (STDs:HPV), reduced smoking intensity (Smokes (packs/yr.)), and early STD diagnosis (STDs:FIRST_DIAG) slightly favored the No Cancer class. However, the dominance of high-risk infection and diagnosis features led the model to strongly predict Cancer.

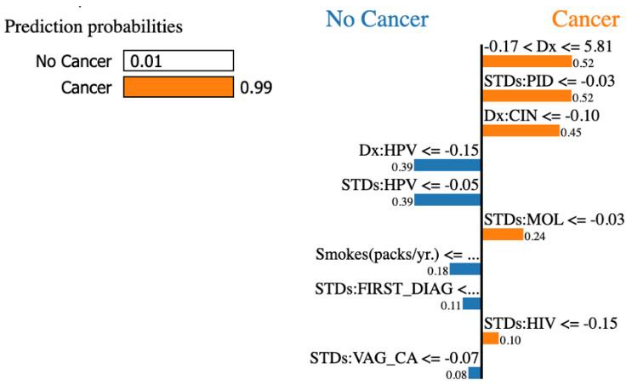

Figure 6. MLP LIME Tabular Plots

b) ELI5:

Table 4 shows the feature importance of the MLP Classifier using ELI5. A general diagnosis (0.128) and confirmed HPV infection (0.119) emerged as the strongest predictors, emphasizing that documented disease presence and viral infection are key indicators of cervical cancer risk. History of STDs including HPV (0.039) and smoking intensity (0.034) also contributed moderately, aligning with known lifestyle and infection-related risk factors. CIN diagnosis had minimal influence (0.004), suggesting that prior cellular abnormalities were less predictive than active HPV infection or overall diagnostic findings in this dataset.

Table 4. Explaining MLP Classifier with ELI5

Feature	Importance
Dx	0.1281
Dx: HPV	0.1195
STDs: HPV	0.0389
Smokes (packs/year)	0.0336
Dx:CIN	0.0042

c) SHAP:

Figure 7 shows the MLP model follows a similar pattern but captures non-linear relationships among features. STDs: HPV and Number of Sexual Partners remain influential, with SHAP values up to +0.12, while Age and First Sexual Intercourse show more dispersed effects. For instance, First Sexual Inter- course at 16 has a +0.05 SHAP value, slightly increasing risk. Unlike Random Forest, the MLP model assigns importance more evenly across features, relying on complex interactions for predictions.

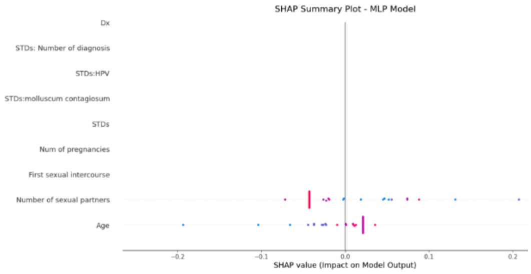

Figure 7. SHAP Analysis for MLP Classifier

5.2.4 GaussianNB Classifier

a) LIME:

As in Figure 8 the GaussianNB model predicted Cancer with a probability of 1.00. The most influential features driving this prediction were the presence of human papillomavirus infection (STDs:HPV), pelvic inflammatory disease (STDs:PID), and recent STD diagnosis (STDs:LAST_DIAG), which are medically recognized as major precursors of cervical malignancy. Additional factors such as molluscum contagiosum (STDs:MOL) and cervical intraepithelial neoplasia (Dx:CIN) further strengthened the cancer indication. The absence of protective responses like negative Hinselmann test, low number of pregnancies (NPG), and negative Schiller test also supported the malignant classification.

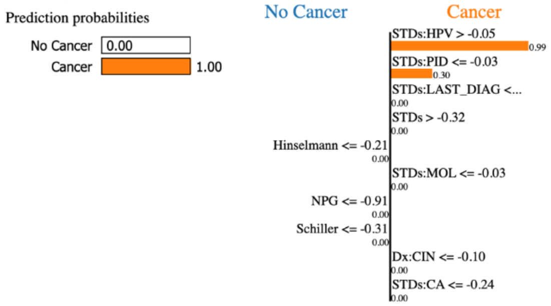

Figure 8. GaussianNB LIME Tabular Plots

b) ELI5:

Table 5 presents the feature importance of the GaussianNB Classifier using ELI5. Confirmed HPV infection (0.167) and general diagnostic history (0.124) were the most influential predictors, underscoring the central role of viral infection and documented disease in cervical cancer risk. Smoking intensity (packs/year, 0.045) and duration of hormonal contraceptive use (0.035) also contributed, reflecting known lifestyle and hormonal risk factors. History of STDs including HPV (0.035) had a moderate impact, while its lower ranking compared to direct HPV diagnosis highlights the stronger predictive value of confirmed infection in this cohort.

Table 5. Explaining GaussianNB Classifier with ELI5

Feature	Importance
Dx: HPV	0.1672
Dx	0.1238
Smokes(packs/year)	0.0452
Hormonal Contraceptive(years)	0.0354
STDs: HPV	0.0348

c) SHAP:

Figure 9 shows the Gaussian NB model, a similar trend is observed, with Age and Number of sexual partners remaining the most influential factors. The SHAP values for Age range between -0.2 and 0.2, while those for Number of sexual partners range between -0.4 and 0.3, showing slightly greater variation than in the Bagging model. The more dispersed SHAP values for Age suggest that this model considers different age groups with more variability in their effect on predictions. Features like Dx:CIN, Smokes, Num of pregnancies and others remain insignificant, with SHAP values near zero.

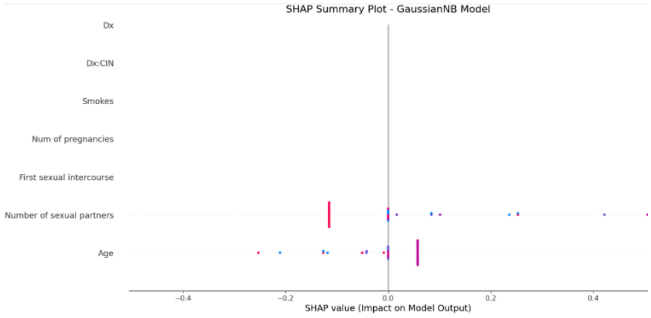

Figure 9. SHAP Analysis for GaussianNB Classifier

5.2.5 Random Forest Classifier

a) LIME:

As in Figure 10, the RF model predicted a No Cancer probability of 0.73 and a Cancer probability of 0.27. Important protective indicators include the absence or low levels of human papillomavirus infection (Dx:HPV, STDs:HPV), limited intrauterine device usage (IUD(yrs.)), and absence of pelvic inflammatory disease (STDs:PID), all of which reduced the cancer likelihood. In contrast, mild influence toward Cancer was contributed by general diagnostic indicators (Dx), presence of vaginal condylomatosis (STDs:VAG_CA), and HIV infection (STDs:HIV). These patterns indicate that the model associated HPV and other sexually transmitted infections with elevated cancer risk, but found stronger evidence for a non-cancerous condition in this instance.

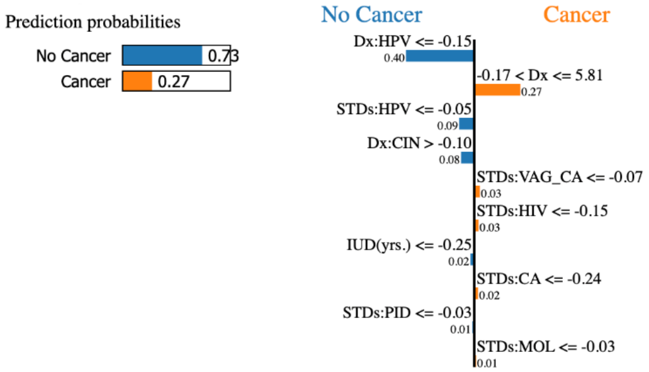

Figure 10. RF LIME Tabular Plots

b) ELI5:

Table 6 shows the feature importance of the RF Classifier using ELI5. Overall diagnostic history (0.302) was by far the most influential predictor of cervical cancer, highlighting that confirmed medical diagnoses carry the strongest risk signal. Schiller test results (0.038) contributed moderately, reflecting its role in detecting abnormal cervical cells. Age (0.001), history of STDs such as condylomatosis (0.001), and number of sexual partners (0.000) had minimal impact, suggesting that in this dataset, demographic and behavioral factors were less predictive than direct diagnostic findings.

Table 6. Explaining RF Classifier with ELI5

Feature	Importance
Dx	0.3020
Schiller	0.0378
Age	0.0012
STDs: condylomatosis	0.0010
Number of sexual partners	0.0000

c) SHAP

Figure 11 shows the RF model highlights STDs: HPV, Age, Number of Sexual Partners, and First Sexual Intercourse as key predictors of cervical cancer. For example, STDs: HPV (Yes) has a SHAP value of +0.12, indicating a strong positive impact, while Age = 30 has a SHAP value of -0.05, slightly reducing risk. Features like Dx and STDs: Number of Diagnoses have minimal impact. The model mainly relies on STD-related factors, with demographic features playing a smaller role.

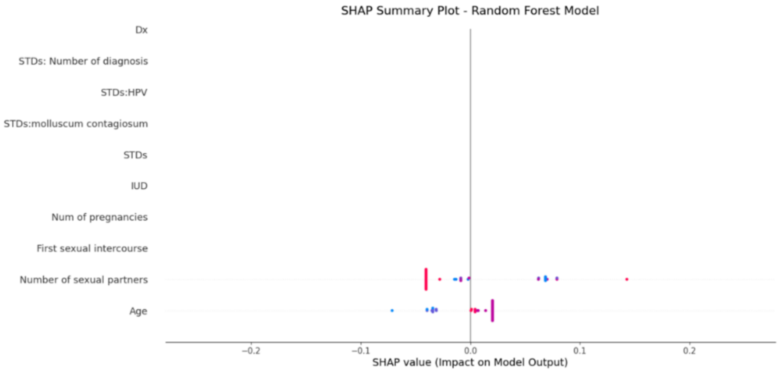

Figure 11. SHAP Analysis for RF Classifier

6. Comparative Analysis

The Table 7 presents the training accuracies of five ML models evaluated using 5-fold cross-validation. Among these models, the MLP consistently achieved the highest performance, with all folds maintaining near perfect accuracy and an overall average of 99.59%. The GaussianNB model demonstrated stable results with fold accuracies ranging between 97.10% and 97.18%, yielding an average accuracy of 97.14%. Similarly, the Bagging classifier performed robustly, recording accuracies between 96.90% and 97.10%, with an average of 97.00%. The RF classifier showed slightly lower performance compared to Bagging and GaussianNB, with fold values fluctuating between 93.90% and 94.10% and achieving an average of 94.00%. Meanwhile, the KNN algorithm produced competitive results, consistently maintaining accuracies around 97.00% with an average of 97.00%.

Table 7: Comparison of Training Accuracy with K-fold Cross Validation

Algorithm	GaussianNB	Bagging	KNN	RF	MLP
K1	97.10	96.10	97.10	93.90	99.61
K2	97.18	97.10	97.00	94.10	99.56
K3	97.15	97.00	97.00	94.00	99.60
K4	97.10	97.00	97.00	94.00	99.59
K5	97.18	97.00	97.00	94.00	99.59

Overall, the results indicate that MLP outperformed all other models, followed by GaussianNB, Bagging, and KNN, while RF recorded the lowest training accuracy among the tested classifiers. This highlights the strength of neural-network-based approaches for capturing complex relationships in the dataset, whereas ensemble-based methods and distance-based classifiers offered strong but comparatively lower performance.

7. Conclusion

This study presents a comparative analysis of various machine learning algorithms for cervical cancer diagnosis, emphasizing their predictive performance through multiple evaluation metrics. The results, summarized in that the MLP Classifier achieves perfect classification with an accuracy of 99.59%, Recall and AUC of 1.00, indicating its superior generalization capability.

Among the other models, GaussianNB, Bagging, and KNN classifiers also exhibit robust performance, each achieving an AUC of 1.00 and an accuracy exceeding 97.00%, making them strong candidates for clinical deployment. The RF Classifier, while performing slightly lower with an accuracy of 94.00%, maintains high precision and recall values, suggesting its reliability for predictive modeling in medical diagnostics.

To enhance interpretability, both SHAP and ELI5 were employed to analyze feature importance, providing transparency into model decision-making. The SHAP analysis identified key contributing features, offering a detailed breakdown of their impact on predictions, while ELI5 provided additional insights into model behavior through permutation-based explanations. The integration of these explainability techniques strengthens the trustworthiness of the models, facilitating their adoption in real-world medical applications.

Acknowledgements

The authors sincerely appreciate the support provided by the Research Unit at Samriddhi College in carrying out this study.

References

Ding, D. et al., 2021. achine learning-based predic- tion of survival prognosis in cervical cance. *Bioinformatics 22.1,* p. 331.

Alsmariy, R., Healy, G. & Abdelhafez, H., 2020. Predicting cervical cancer using machine learning methods. *International Journal of Advanced Computer Science and Applications,* Volume 11.

Alsmariy, R., Healy, G. & Abdelhafez, H., 2020. Predicting Cervical Cancer using Machine Learning Methods. *International Journal of Advanced Computer Science and Applications.*

Alvarez, S. A., 2002. An exact analytical relation among recall, precision, and classification accuracy in information retrieval. *Technical Report BCCS.*

Asadi, Salehnasab & Ajori, 2020. Supervised Algorithms of Machine Learning for the Prediction of Cervical Cancer. *J Biomed Phys Eng.,* 1 Aug.pp. 513-522.

Aziz, K., 2025. *Kaggle.* [Online]
Available at: www.kaggle.com/datasets/khuzaimaaziz/cervical-cancer-dataset-csv

Bataineh, A., Kaur, D. & Jalali, S., 2022. Multi-layer perceptron training optimization using nature inspired computin. *IEEE access,* pp. 36963-36977.

Bauer, E. & Kohavi, R., 1999. An empirical compari- son of voting classification algorithms. *Springer Nature Link,* pp. 105-139.

Biau, G. & Scornet, E., 2016. A random forest guided tour. *Test 25.3,* pp. 197-227.

Canavan, Timothy, Doshi & Nipa, 2000. Cervical cancer. *American family physician,* Volume 61, pp. 1369-1376.

Chadaga, et al., 2022. Predicting cervical cancer biopsy results using demographic and epidemiological parameters: A custom stacked ensemble machine learning approach. *Cogent Engineering,* Volume 9.

Cohen, P. A., Jhingram, A., Oaknin, A. & Denny, L., 2019. Cervical Cancer. *The Lancet Journal,* pp. 169-182.

Deng, Luo & Wang, 2018. *Analysis of risk factors for cervical cancer based on machine learning methods.* s.l., IEEE, pp. 631--635.

Falyo, D. & Holland, B., 2017. *Medical and psychosocial aspects of chronic illness and disability.* s.l., s.n.

Fowler, J., Maani, E., Gasalberti, D. & Jack, B., 2017. Cervical Cancer. *europemc.*

Ghoneim, A., Muhammad, G. & Hossain, M. S., 2020. Cervical cancer classification using convolutional neural networks and extreme learning machines. *Future Generation Computer Systems ,* pp. 643-649.

Hasan, Roy & Nitu, 2022. Cervical cancer classification using machine learning with feature importance and model explainability. In: *2022 4th international conference on electrical, computer & telecommunication engineering (ICECTE).* s.l.:IEEE.

Lei, Z., Jiang, Y., Zhao, P. & Wang, J., 2009. s.l., Springer, pp. 431-438.

Mehmood, M., Rizwan, M., Gregus ML, M. & Abbas, S., 2021. Machine learning assisted cervical cancer detection. *Frontiers in public health.*

Michaud, E. J., Liu, Z. & Tegmark, M., 2023. Precision Machine Learning. *Entropy.*

Mudawi, N. A. & Alazeb, A., 2020. A Model for Predicting Cervical Cancer Using Machine Learning Algorithms. *Sensors 22.11.*

Mujahid, et al., 2024. Data oversampling and imbalanced datasets: An investigation of performance for machine learning and feature engineering. *Journal of Big Data,* Volume 11, p. 87.

Osuwa, A. A. & öztoprak, H., 2021. *Importance of continuous improvement of machine learning algorithms from a health care management and management information systems perspective.* s.l., IEEE, pp. 1-5.

Popescu, M., Balas, V., Popescu, L. & Mastorakis, N., 2009. Multilayer percep- tron and neural networks. *WSEAS Transactions on Circuits and Systems,* pp. 579-588.

Sawant, N. & Khadapkar, D., 2022. Comparison of the performance of GaussianNB Algorithm, the K Neighbors Classifier Algorithm, the Logistic Regression Algorithm, the Linear Discriminant Analysis Algorithm, and the Decision Tree Classifier Algorithm on same dataset. *International Journal for Research in Applied Science & Engineering Technology (IJRASET).*

Shakil, Islam & Akter, 2024. A precise machine learning model: Detecting cervical cancer using feature selection and explainable AI. *Journal of Pathology Informatics.*

Skyler, J. et al., 2017. *Differentiation of diabetes by pathophysiology, natural history, and prognosis.* s.l., s.n.

Tanimu, J. et al., 2022. A Machine Learning Method for Classification of Cervical CancerA Machine Learning Method for Classification of Cervical Cancer. *MDPI,* p. 463.

Uddin, et al., 2025. *Ensemble Machine Learning-Based Approach to Predict Cervical Cancer with Hyperparameter Tuning and Model Explainability.* s.l., Springer.

William and Bacon, M. A. a. B. A. a. C. et al., 2017. Cervical cancer: a global health crisis. *American Cancer Society.*