Journal of Pathology Informatics 15 (2024) 100398

Contents lists available at ScienceDirect

Journal of Pathology Informatics

journal homepage: www.elsevier.com/locate/jpi

A precise machine learning model: Detecting cervical cancer using feature selection and explainable AI

Rashiduzzaman Shakil *, Sadia Islam, Bonna Akter

Department of Computer Science and Engineering, Daffodil International University, Dhaka, Birulia 1216, Bangladesh

ARTICLE INFO

Keywords:
Cervical cancer
SMOTE
ADASYN
Chi-square
LASSO
Machine learning
Decision tree
Explainable AI
SHAP

ABSTRACT

Cervical cancer is a cancer that remains a significant global health challenge all over the world. Due to improper screening in the early stages, and healthcare disparities, a large number of women are suffering from this disease, and the mortality rate increases day by day. Hence, in these studies, we presented a precise approach utilizing six different machine learning models (decision tree, logistic regression, naïve bayes, random forest, k nearest neighbors, support vector machine), which can predict the early stage of cervical cancer by analysing 36 risk factor attributes of 858 individuals. In addition, two data balancing techniques—Synthetic Minority Oversampling Technique and Adaptive Synthetic Sampling—were used to mitigate the data imbalance issues. Furthermore, Chi-square and Least Absolute Shrinkage and Selection Operator are two distinct feature selection processes that have been applied to evaluate the feature rank, which are mostly correlated to identify the particular disease, and also integrate an explainable artificial intelligence technique, namely Shapley Additive Explanations, for clarifying the model outcome. The applied machine learning model outcome is evaluated by performance evaluation matrices, namely accuracy, sensitivity, specificity, precision, f1-score, false-positive rate and false-negative rate, and area under the Receiver operating characteristic curve score. The decision tree outperformed in Chi-square feature selection with outstanding accuracy with 97.60%, 98.73% sensitivity, 80% specificity, and 98.73% precision, respectively. During the data imbalance, DT performed 97% accuracy, 99.35% sensitivity, 69.23% specificity, and 97.45% precision. This research is focused on developing diagnostic frameworks with automated tools to improve the detection and management of cervical cancer, as well as on helping healthcare professionals deliver more efficient and personalized care to their patients.

Introduction

Cervical cancer arises primarily in the cervix cells located in the lower region of the uterus. Infection with human papillomaviruses (HPVs), transmitted through sexual contact, is the primary cause of cervical cancer. According to data, cervical cancer ranks first or second among women for cancer-related deaths.[1] A significant proportion of the recorded cases in 2018 originate from both impoverished and industrialized nations.[2] Cervical cancer constitutes 6.5% of all malignancies in women. Around 342,000 cervical cancer deaths and 604,127 new cases of the illness are expected to occur globally in 2020.[3] The prevalence of cervical cancer is a significant public health issue, especially in affluent nations: The number of newly identified instances of invasive cervical cancer in Europe is 54,517 each year, with 24,874 women losing their lives to this type of cancer.[4]

Vaccinations are commonly included in standard immunization regimens. Approximately, 90% of European girls are expected to have had the full HPV vaccination by the age of 15 by 2030.[5] Moreover, the World Health Organization (WHO) strongly emphasizes the need of European decision-makers intensifying their efforts to eradicate cervical cancer by utilizing existing preventive measures.[6] Systematic screenings and prompt detection, like other medical conditions, can significantly reduce the likelihood of mortality.[7] Given the lack of symptoms during the first phases of the disease, the identification of cervical cancer may provide challenges. Systematic annual examinations often uncover alterations in the cervix. The most common sign of cervical cancer is abnormal bleeding, which can change in severity as the disease progresses. In approximately 90% of cases, the advanced phases of the illness exhibit distinct hallmarks.[8] Common signs of this condition include reddish discharge, spotting, contact bleeding, and postmenopausal bleeding. In addition to the symptoms, the existence of bloody discharge, which usually has an unpleasant smell, is another crucial indication of cervical cancer. When the disease has evolved to a certain point, it can affect neighboring organs, which can lead to lower abdominal discomfort.[9]

To reduce vast mortality rate, early detection of the disease is essential. Early-stage cervical cancer can be cured at an average rate of 80% with either radical surgery or radiation.[10]

* Corresponding author.
E-mail addresses: rashiduzzaman.diucse@gmail.com (R. Shakil), sadia15-3980@diu.edu.bd (S. Islam), bonna.diucse@gmail.com (B. Akter).

http://dx.doi.org/10.1016/j.jpi.2024.100398
Received 17 July 2024; Received in revised form 12 September 2024; Accepted 17 September 2024
Available online 26 September 2024

One of the most well-known branches of artificial intelligence (AI), machine learning, has been quite successful in many different areas.[11] Machine learning algorithms are now being used more and more to predict cervical cancer and its early stages by utilizing several forms of data, such as cytology (pap smear), histology, scanning, and clinical information. AI has demonstrated significant potential in accurately predicting cervical cancer by examining several datasets linked to the condition. The development of explainable AI (XAI) originated from the critical need for ensuring responsible deployment of machine learning models. It acknowledges the issue that intricate algorithms can frequently function as 'black boxes', rendering their decision- making procedures obscure and potentially resulting in biases or ethical concerns.

This study focuses on the utilization of XAI approaches, specifically SHAP, to enhance the comprehension of cervical cancer predictions. Recognition of this information is essential for healthcare professionals to accurately evaluate the model's results and optimize the prediction process for better accuracy and clinical significance. This study utilized machine learning models to predict the probability of negative and positive cervical cancer for clinical use.

The contribution of this study are as follows:

- Introducing two robust feature selection strategies, including the Chi-square test and Least Absolute Shrinkage and Selection Operator (LASSO), to improve the predictive accuracy and efficiency of cervical cancer prediction models.
- In this research, a comparative result analysis was presented between two distinct data balancing techniques, namely synthetic minority oversampling technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN).
- In addition, we conducted a novel study where we integrated the explainable AI model, SHAP, to improve the interpretability and transparency of our cervical cancer prediction models.
- This study utilized six machine learning models, such as decision trees (DTs), logistic regression (LR), naive bayes (NB), random forest (RF), k nearest neighbor (KNN), and support vector machines (SVMs). Each of these models contains distinct advantages and qualities, rendering them ideal for certain elements of the prediction task.
- The model's performance metrics were measured in terms of accuracy, specificity, sensitivity, F1-score, precision, false-positive rate (FPR), and false-negative rate (FNR) to thoroughly evaluate their ability to make accurate predictions.

Background and literature review

Nowadays, cervical cancer is one of the most alarming diseases for women in society. Therefore, researchers are presenting their own methodology utilizing AI to identify the early stages of cervical cancer.

Mehmood et al.[12] proposed a novel approach named CervDetect for detecting cervical cancer using machine learning algorithms. They used pearson correlation to identify relevant features. The RF and shallow neural networks were utilized to accurately predict cervical cancer and achieved an accuracy of 93.6% and mean square error of 0.07111. Arora et al.[13] introduced a method for classifying Herlev pap smear image datasets using SVMs. In their study, bilateral filtering was used for noise removal and local Gaussian fitting energy segmentation for cell and nuclei segmentation, achieving 95% accuracy.

Sabanci et al.[14] classify patients based on their survey responses to identify the high risk of cervical cancer using multilayer perceptron, BayesNet, and KNN methods. They achieved the highest result of false positives in the BayesNet and KNN algorithms. A study on the risk factors associated with cervical cancer conducted by Razali et al.[15] They utilized classification algorithms namely NB, DT, and KNN. This study applied mean to handling the missing values and SMOTE for data balancing. They evaluated the effectiveness of several classifiers using 10-fold cross-validation, showcasing the potential of data mining techniques in medical diagnosis and treatment planning.

Parikh et al.[16] developed an approach to predicting cervical cancer utilizing machine learning algorithms and various factors present in the dataset. Their study includes the establishment of three models: KNN, DT, and RF. Specially, the KNN achieved the highest accuracy among the three models. This research showed the capability to identify individuals at risk of cervical cancer. Ou et al.[17] employed six machine learning classifiers to predict postoperative pathological risk factors in 1260 early-stage cervical cancer patients, who had a radical hysterectomy. The RF classifier demonstrated the highest level of accuracy in predicting deep stromal infiltration, achieving an accuracy rate of 70.8% and an AUC (area under the receiver operating characteristic (ROC) curve) value of 0.767.

Malli et al.[18] proposed an automated machine learning approach using KNN and ANN to predict cervical cancer. The technique analyzes the color and shape features of the cervix cell nucleus and cytoplasmic segments using a fuzzy-based technique. They achieved 88.04% accuracy with kNN and 54% with ANN for cell classification in cervical cancer prediction. Latha et al.[19] compared the data mining algorithms for classifying and identifying the stage of cervical cancer. These algorithms were analyzed by accuracy, sensitivity, and specificity, and J48 was found to be the most effective with 93.03% specificity and sensitivity over 80%. A DT accurately identifies attributes closely related to cancer staging, with sensitivity surpassing 70%.

Sun et al.[20] proposed a technique for identifying cervical cancer by employing a RF classifier along with ReliefF feature selection. The study utilizes preprocessing, segmenting, and extracting 20 features and ranks these features by importance. The RF classifier, utilizing the most significant top 13 features, outperformed NB, C4.5, and LR, achieving 94.44% accuracy and AUC score of 98.04%. Ilyas et al.[21] presented a novel ensemble classification approach to enhance the diagnostic precision of cervical cancer. By employing an ensemble method that combines the predictions of various classifiers including DT, SVM, RF, K-NN, NB, MLP, J48, and LR. This study highlights the capacity of machine intelligence to provide cost-effective and rapid diagnosis of cervical cancer.

Nithya et al.[22] presented an efficient framework for cervical cancer diagnosis by addressing challenges related to high-dimensional, redundant, and imbalanced datasets. Their goal is to achieve classification accuracy and computational efficiency by combining filter- and wrapper-based feature selection methods. Vidya et al.[23] analyzed three supervised machine learning methods: ID3, C4.5, and NB, for cervical cancer prediction using a novel NCBI dataset. ID3 and C4.5 build DTs iteratively, whereas NB quickly creates Bayesian statistical models. Overall, NB predicts cervical cancer better than other classifiers.

Tseng et al.[24] applied machine learning techniques to predict cervical cancer recurrence-proneness to improve clinical decision-making. They used medical records and pathology data to identify recurrence risk factors, finding that the C5.0 model is particularly effective. This study suggested pathological stage and pathological T for better surveillance and treatment based on these predictive labels. Chaudhuri et al.[25] suggested a three-stage hybrid feature selection approach and a stacked classification model for early detection of cervical cancer using machine learning techniques. In this research, six algorithms have been employed including LR, NB, SVM, ET, RF, and GDB, where NB outperformed with 97%.

Yang et al.[26] explored machine learning models to improve the diagnosis and prediction of cervical cancer using MLP and RF. They identified age, number of sexual partners, and hormonal contraception as major risk factors for cervical cancer. To conduct this study, they used 858 individuals information. Kalbhor et al.[27] developed a method for diagnosing cervical cancer by classifying cytology pap smear images. This study analyzed the efficacy of utilizing discrete cosine transform and Haar transform coefficients as characteristics for image classification. They employed seven ML classifiers where the RF classifier achieved the highest accuracy rate of 81.11%. Table 1. consists outcome of the related research work which are studied by several researcher.

Table 1
Comparative analysis with other existing work.

Author and reference	Dataset source	Data frequency	Applied feature selection	Best ML classifier	Performance outcome
Mehmood et al.[12]	UCI Machine Learning Repository	859	Random forest, Pearson correlation	RF and shallow neural network	ACC: 93.6%,
Arora et al.[13]	Herlev University Hospital. Denmark	917	Principal component analysis (PCA)	SVM	ACC: 95%
Sabanci et al.[14]	UCI Machine Learning Repository	858	No	NB	ACC: 97%
Razali et al.[15]	UCI Machine Learning Repository	858	No	RF	ACC: 96.40%
Parikh et al.[16]	UCI Machine Learning Repository	429	Hill climbing algorithm	KNN	ACC: 95.3%.
Ou et al.[17]	Wenzhou Medical University	1260	LinearSVC with L1 penalty	RF	ACC: 70.8%
Malli et al.[18]	UCI Machine Learning Repository	300	PCA	KNN	ACC: 88.04%
Latha et al.[19]	Indian Statistical Institute, Calcutta	203	CFS	J48	ACC: 93.03%
Sun et al.[20]	Herlev University Hospital, Denmark	917	ReliefF	RF	ACC: 94.44%
Ilyas et al.[21]	UCI Machine Learning Repository	858	Cross-validation and supervised feature selection	Ensemble classifier	ACC: 94%
Nithya et al.[22]	UCI Machine Learning Repository	Not mentioned	Information gain, Chi-square Correlation coefficient, RFE, GA, Sequential feature selection	RF	ACC: 85.5%
Vidya et al.[23]	National Center for Biotechnology Information (NCBI)	Not mentioned	PCA	NB	ACC: 81%
Tseng et al.[24]	Chung Shan Medical University Hospital Tumor Registry.	168	GainRatio	C5.0	ACC: 96%
Chaudhuri et al.[25]	UCI Machine Learning Repository	858	Multistage GA	GDB	ACC: 97%
Yang et al.[26]	Caracas University Hospital in Venezuela	858	GA	MLP	ACC: 97.2%
Kalbhor et al.[27]	Herlev University Hospital, Denmark	917	DCT, Haar Transform	RF	ACC: 81.11%
Shakil et al. (proposed methodology)	**UCI Machine Learning Repository**	**858**	**Chi-square and LASSO**	**DT**	**ACC: 97.60%**

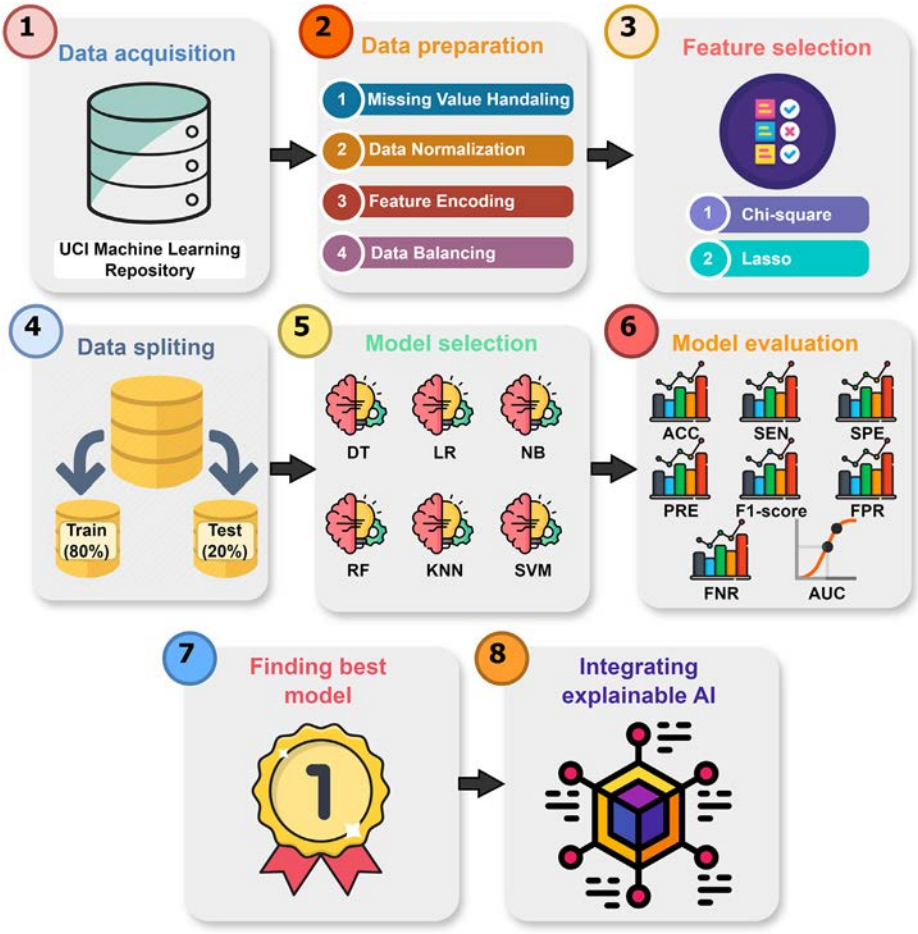

Fig. 1. Visualization the overall working procedure of this study.

Methodology

In this section, we are mentioned all required task sequentially. Fig. 1 illustrates the overall working flow that is divided into seven sections: data acquisition, data preparation, feature selection, data split, model selection, evaluation of the applied model, finding the best model, and integrating explainable AI.

Data collection

To conduct this research, we used cervical cancer (risk factors) dataset, which was collected from a renowned online repository called UCI machine learning repository.[28,29] The dataset was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela. This dataset comprises 858 individual records of medical patients, each with 36 distinct attributes, of which 55 belong to the cervical cancer class and 803 do not have cervical cancer. Fig. 2. depicts the ration of dataset between two classes. The primary feature of this dataset is its ability to distinguish cervical cancer patients based on their demographic information, habits, and historical medical records.

Data preprocessing

Data preprocessing is a crucial part of machine learning; it can improve data quality, enhance the performance of models, and ensure data consistency. During this phase, at first, we checked the missing values by applying the isnull() function and handling the missing values using mean.[30] Fig. 3 visualizes the number of missing values of each features. Secondly, a min-max scaler was used to normalize this dataset, and the range of the normalization was from 0 to 1. Thirdly, we employed feature encoding to convert our target class (biopsy class) into 0 and 1, where cervical cancer is 1 and not 0. Furthermore, the SMOTE and ADASYN are two different data balancing techniques that had been used to eliminate the class imbalance problem. SMOTE aims to balance class distribution by generating synthetic examples for the minority class.[31] ADASYN also balances class distribution, but it focuses on generating synthetic data points that are harder to classify, thus improving the learning of complex decision boundaries.[32] Moreover, the whole dataset was separated into training and testing with a ratio of 80:20.

Feature selection

Feature selection is the process of identifying the most significant features from a large number of attributes that are not relevant for this particular disease. In this study, we employed two different feature selection techniques, named LASSO and Chi-square. LASSO feature selection utilizes regression technique to identify and select the most relevant features for classifying cervical cancer presence or absence.[33] Table 2 contain the selected features by Chi-square and LASSO. Initially, this dataset consists of 36 features. During feature selection phase, Chi-square feature selection has 30 features and 17 features was selected by LASSO feature selection. This approach operates by imposing a penalty on the coefficients of the regression model, hence causing certain coefficients to be reduced to zero. In addition, it can improve the performance of the models and interpretability by removing less-essential elements. Within the framework of binary classification, the LASSO technique assists in differentiating between malignant and benign instances by prioritizing the most influential variables. This ultimately contributes to a more precise and efficient diagnosis of cervical cancer.

$$\underset{\beta}{\text{minimize}}\left(\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log\left(h_\beta(X_i)\right) + (1 - y_i)\log\left(1 - h_\beta(X_i)\right)\right] + \lambda \sum_{j=1}^{p}|\beta_j|\right) \tag{1}$$

Here, n is the total number of samples in dataset, p is the feature. Let X be the $n \times p$ feature matrix, y be the binary target vector, and β be the vector of coefficients.

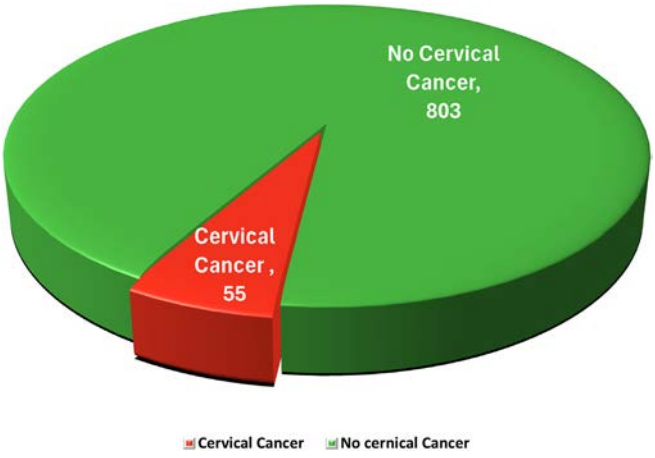

Fig. 2. Ratio of cervical cancer and no cervical cancer based on overall dataset.

where:

- y_i is the main target value for the i-th sample.
- β_0 is the intercept.
- X_{ij} is the value of the j-th feature for the i-th sample.
- β_j is the coefficient for the j-th feature.
- λ is the regularization parameter that controls the amount of shrinkage applied to the coefficients.

The degree of independence between each characteristic and the target variable is assessed by computing the Chi-square statistic.[34] A higher degree of dependence on the target variable indicates that the features are more significant and so are chosen.

For Chi-square feature selection in a binary classification task, where both feature X and the target Y *are binary*, the Chi-square statistic (χ^2) measured how much obeserved distribution of the feature values differs from the expected distribution. The Chi-square statistic is calculated by the following formula:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \tag{2}$$

where:

The number of rows (categories of feature X) indicated by r. c is the number of columns (binary classes of target Y). O_{ij} is the observed frequency count of the i-th category of X and j-th class of Y. E_{ij} is the expected frequency count, calculated as:

$$E_{ij} = \frac{\sum_{k=1}^{r} O_{kj} \sum_{l=1}^{c} O_{il}}{N} \tag{3}$$

where N is the total number of observations.

Applied machine learning model

In order to detect the presence of cervical cancer, this research employed six machine learning algorithms including DT, LR, NB, RF, KNN, and SVM. The details of applied model are mentioned below:

Decision tree: DT is a supervised learning model that is used for both classification and regression tasks. For binary classification of cervical cancer, the model splits the data into subsets based on the values of input features, recursively creating branches until reaching the leaf nodes, which represent the output class labels. The splitting is done using criteria like Gini impurity or entropy (information gain).[35] Gini impurity is a measure of how often a randomly chosen element would be incorrectly labeled if

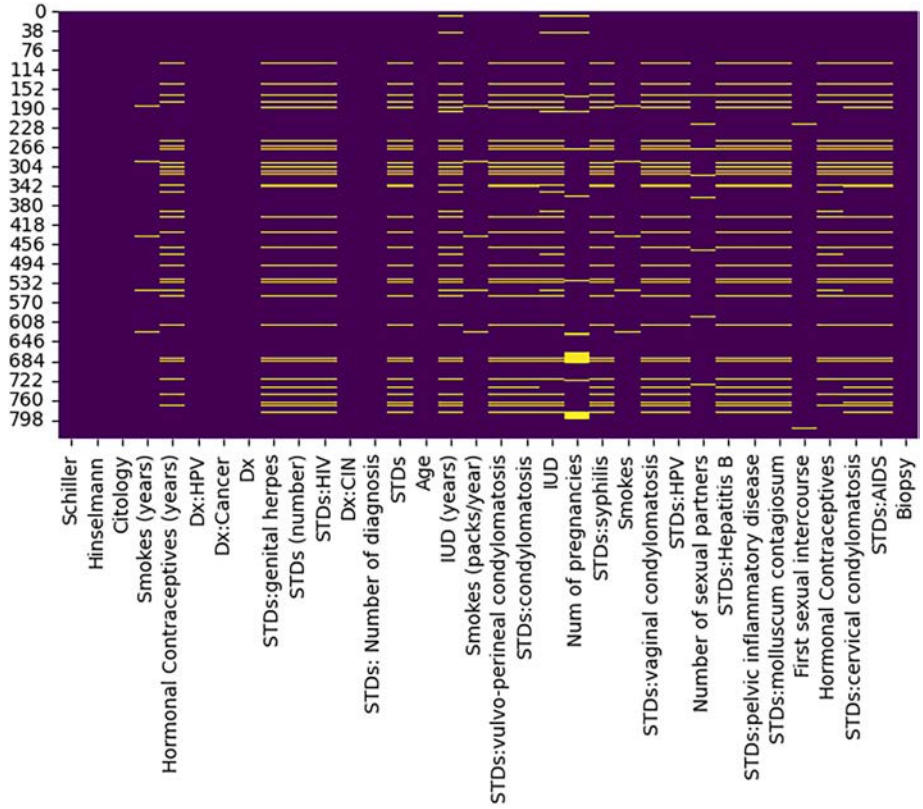

Fig. 3. Number of missing values of each attributes.

Table 2
Features rank according to the feature section technique.

Method used	Features
Chi-square	Schiller, Hinselmann, Citology, Smokes (years), Hormonal contraceptives (years), Dx:HPV, Dx:Cancer, Dx, STDs:genital herpes, STDs (number), STDs:HIV, Dx:CIN, STDs: Number of diagnosis, STDs, Age, IUD (years), Smokes(packs/year), STDs: Vulvo-perineal condylomatosis, STDs:condylomatosis, IUD, Num of pregnancies, STDs: Syphilis, Smokes, STDs:vaginal condylomatosis, STDs:HPV, Number of sexual partners, STDs: Hepatitis B, STDs:pelvic inflammatory disease, STDs:molluscum contagiosum, First sexual intercourse
LASSO	Schiller, Hinselmann, Citology, Dx:HPV, Dx:Cancer, STDs:genital herpes, Dx:CIN, Hormonal Contraceptives (years), Number of sexual partners, Num of pregnancies, First sexual intercourse, STDs:vulvo-perineal condylomatosis, Smokes(packs/year), STDs: Vaginal condylomatosis, Number of diagnosis, IUD, STDs:Hepatitis B

it was randomly labeled according to the distribution of labels in the subset. It is calculated as:

$$\text{Gini} = 1 - \sum_{i=1}^{C} p_i^2 \tag{4}$$

where p_i is the probability of a randomly chosen element being classified as class i, and C is the number of classes (in binary classification, $C=2$).

Entropy used to evaluate best feature and threshold for splitting.

$$\text{Entropy} = -\sum_{i=1}^{C} p_i \log_2(p_i) \tag{5}$$

where p_i is the probability of a randomly chosen element being classified as class i, and C is the number of classes.

Logistic regression (LR): One statistical model that is commonly employed for binary classification is LR. The likelihood that an input x falls into the positive category (cervical cancer) is represented by it. After applying the logistic (sigmoid) function to a linear combination of the input features, the output is a probability value between 0 and 1.[36]

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}} \tag{6}$$

where β_0 is the intercept, β_i are the coefficients for each feature x_i, and $X = (x_1, x_2, \ldots, x_n)$ is the feature vector.

Naïve bayes (NB): The NB classifier is a probabilistic classification method that operates under the 'naive' premise that all features are independent of the class label. It determines the posterior probability of every class and forecasts the class with the highest probability for binary classification of cervical cancer.[37]

$$P(Y|X) \propto P(Y) \prod_{i=1}^{n} P(x_i|Y) \tag{7}$$

where $P(Y)$ is the prior probability of the class, and $P(x_i|Y)$ is the likelihood of feature x_i given class Y.

Random forest: RF is an ensemble learning technique that builds numerous DTs during training and produces the most frequent prediction among them for classification tasks.[38] It enhances precision and mitigates overfitting by averaging many complex DT trained on different subsets of the same dataset.

$$\hat{y} = \text{mode}(\{h_1(x), h_2(x), \ldots, h_B(x)\}) \tag{8}$$

where $h_i(x)$ is the prediction of the i-th tree for input x.

K-nearest neighbor (KNN): KNN is a non-parametric, instance-based learning algorithm that classifies a sample based on the majority class

among its k nearest neighbors in the feature space. For binary classification of cervical cancer, it considers the closest k samples in the training data.[39]

$$\hat{y} = \text{mode}(\{y_i : x_i \in NN_k(x)\}) \tag{9}$$

where $N_k(x)$ represents the set of k nearest neighbors of x, and y_i is the class label of neighbor x_i.

Support vector machine (SVM): SVM is a separating hyperplane-defined discriminative classifier. SVM uses support vectors that are closest to a hyperplane in high-dimensional space to increase the margin between cervical cancer binary classes.[40]

$$f(x) = \text{sign}(w \cdot x + b) \tag{10}$$

where **w** is the weight vector, **x** is the feature vector, and b is the bias term. The function sign determines the class label based on the sign of the result.

Explainable artificial intelligence (XAI)

XAI is of significant importance in the healthcare field, because it is necessary to comprehend the reasoning behind the decisions made by machine learning models in order to facilitate their acceptance and use in clinical settings. Shapley Additive Explanations (SHAP), a type of XAI approach, improves the clarity and comprehensibility of AI models.[41] These techniques assist in understanding the influence of factors like age, smoking habits, and HPV status on the model's decision to classify a sample as high- or low-risk in the context of cervical cancer classification. Model behavior can be better understood with the use of XAI, which in turn improves patient outcomes, enhances integrity in behavior, and secures compliance with regulations. Therefore, the incorporation of XAI in healthcare not only enhances the precision and dependability of AI applications but also encourages their adoption and utilization in clinical environments.

The SHAP value can be evaluated by following formula:

$$\phi_i = \phi_0 + \sum_{j=1}^{p} \omega_j z_{ij} \tag{11}$$

where:

- ϕ_i represents the SHAP value for the i-th feature.
- ϕ_0 is the base value, which is the average prediction of the model.
- ω_j are weights that represent the contribution of each feature.
- z_{ij} are binary variables (0 or 1) that indicate whether the i-th feature for the j-th instance is active.

Performance evaluation metrices

Performance evaluation metrics are the integral part of assessing the effectiveness and efficiency of applied models. It offers a clear understanding of a model's performance in comparison to other models, enabling the selection of the most suitable model for a given task. In this research, we assess seven performance evaluation metrics to determine the model efficiency, and all of the metric formulas are mentioned below:

Accuracy: Accuracy is a metric that evaluates the overall correctness of a model by determining the proportion of correctly predicted cases, including both true positives (TPs) and true negatives (TNs), out of the total instances.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{12}$$

Sensitivity: Sensitivity quantifies the accuracy of the model in correctly identifying actual positive cases. It is alternatively referred to as recall or the genuine positive rate.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

Specificity: Specificity is a metric that quantifies the accuracy of a model in correctly identifying negative cases.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{14}$$

Precision: Precision is defined as the fraction of positive forecasts that are correct.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{15}$$

F1-score: The F1-score represents the harmonic mean of precision and recall. It gives a single statistic that balances precision and recall, which is particularly beneficial when both are required.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{16}$$

False-positive rate (FPR): The FPR quantifies the ratio of negative cases that are erroneously identified as positive.

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{17}$$

False-negative rate (FNR): The FNR quantifies the percentage of TP instances that are inaccurately identified as negative.

$$\text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TP}} \tag{18}$$

The proper meaning of TP, TN, FP, and FN are mentioned sequentially in below:

- TP: TP occurs when a model predicts cervical cancer and the actual condition is also cervical cancer.
- TN: TN would happen if a model predicts no cervical cancer and the actual condition is also no cervical cancer.
- FP: This occurs when a model predicts cervical cancer, but the actual condition is no cervical cancer.
- False negative (FN): The FN would happen if a model predicts no cervical cancer but the actual condition is cervical cancer.

Result

In this section, we present the performance outcomes of six machine learning algorithms—DT, LR, NB, RF, KNN, and SVM—applied to the binary classification of cervical cancer. We evaluate each model based on key metrics, including accuracy, sensitivity, specificity, precision, F1-score, FPR, FNR, and AUC. Additionally, SHAP is an explainable AI technique that have been utilized for model interpretability, offering insights into the significance of various features. These results aim to highlight the comparative effectiveness of each algorithm in accurately distinguishing between healthy and cancerous cases, thereby informing the potential application of these models in clinical settings.

Table 3 represents the performance results of applied models without data balancing. It highlights the strengths and weaknesses across different metrics. The DT and RF algorithms both achieved high accuracy (97%), with DT excelling in sensitivity (99.35%) and RF showing a strong balance between sensitivity (97.50%) and specificity (85.71%). The LR performed accuracy rates (94.01%) with 96.82% precision. The NB algorithm achieved 92.21% accuracy, which is the lowest among all applied machine learning algorithms to classify cervical cancer patients. However, KNN and SVM both showed accuracy of 95.81%, with SVM having a slightly higher specificity (80%) compared to KNN (71.43%).

The summarizes performance evaluation results after data balancing with SMOTE is presented in Table 4. The LR algorithm achieved an

Table 3
Performance evaluation result without balancing.

Algorithm name	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)	FPR (%)	FNR (%)
DT	97	99.35	69.23	97.45	98.39	30.77	0.65
LR	94.01	96.82	50	96.82	96.82	50	3.18
NB	92.21	96.75	38.46	94.90	95.82	61.54	3.25
RF	97	97.50	85.71	99.36	98.42	14.29	2.50
KNN	95.81	96.88	71.43	98.73	97.79	28.57	3.12
SVM	95.81	96.29	80	99.36	97.80	20	3.71

Table 4
Performance evaluation result after data balancing using SMOTE.

Algorithm name	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)	FPR (%)	FNR (%)
DT	86.23	98.57	29.62	87.97	92.96	70.38	1.43
LR	94.61	99.33	52.94	94.90	97.06	47.06	0.67
NB	86.22	98.55	27.59	86.62	92.20	72.41	1.45
RF	95.21	98.69	57.14	96.18	97.42	42.86	1.31
KNN	95.21	99.34	56.25	95.54	97.41	43.75	0.66
SVM	96.41	99.35	64.29	96.82	98.06	35.71	0.65

accuracy of 94.61%, precision of 94.90%, and an F1-score of 97.06%. NB showed an accuracy of 86.22%, sensitivity of 98.55%, specificity of 27.59%, and an F1-score of 92.20%. KNN demonstrated a sensitivity of 99.34%, specificity of 56.25%, and precision of 95.54%. SVM exhibited the highest accuracy at 96.41%, sensitivity of 99.35%, specificity of 64.29%, precision of 96.82%, and an F1-score of 98.06%. These results highlight the effectiveness of SMOTE in improving classification metrics across different algorithms, with SVM achieving the highest overall accuracy.

Table 5 presents the performance evaluation results after data balancing using ADASYN. The DT algorithm achieved 87.43% accuracy, 98.75% sensitivity, 29.62% specificity, and a precision of 87.97%. LR showed an accuracy of 89.22%, 93.99% of F1-score and 89.81% precision, and an FPR is 66.67%. RF gained a sensitivity rate of 99.34%, specificity of 56.25%, precision of 95.54%, and the FPR is 43.75%. The SVM algorithm achieved the highest accuracy at 96.41%, with a sensitivity of 99.35%, 64.29% specificity, 96.82% precision, and F1-score is 98.06%. Among the algorithms, SVM had the highest accuracy, whereas NB had the lowest accuracy of 85.63%.

The study assessed the performance of the binary classification model by utilizing a confusion matrix, which is an essential tool for summarizing the predicted accuracy of the model. The confusion matrix is a tabular representation of a 2×2 table that presents the TP, TN, FP, and FN.

Figs. 4 and 5 depict the confusion matrix for Chi-square and LASSO feature selection for the six ml algorithms. In Fig. 4, DT has TP (155), TN (8), FP (2), and FN (2). The SVM classifier gained TN (9), TP (152), FP (5), and FN (1). In addition, in Fig. 5, DT and KNN have the same TP, TN, FP, and FN values of 152, 9, 5, and 1. Similarly, the performance of the LR algorithm is comparable to that of the RF.

Table 6 presents the performance evaluation results utilizing Chi-square feature selection. The DT algorithm achieved an accuracy of 97.60%, with a sensitivity of 98.73%, a specificity of 80%, and a precision of 98.73%. LR demonstrated an accuracy of 95.81%, an F1-score of 97.73%, a precision of 96.18%, an FPR of 40%, and an FNR of 0.66%. RF showed 98.70% sensitivity, 61.54% specificity, 96.82% precision, and 36.46% of FPR. NB performed with 92.21% accuracy, whereas it has 94.90% precision, 95.81% F1-score, 61.54% FPR, and 3.25% FNR. SVM had an accuracy of 96.41% with a sensitivity of 99.35%, and the specificity, precision, and F1-score were gradually 64.29%, 96.82%, and 98.06%, respectively. Among the algorithms, DT had the highest accuracy, making it the best performer. SVM, with an accuracy of 96.41%, ranked second. NB and KNN had the lowest accuracy at 92.21%.

The performance evaluation results utilizing LASSO feature selection is present in Table 7. The DT and KNN algorithm have the similar performance result. The accuracy of 96.41%, where the sensitivity, specificity, precision, and F1-score are 99.35%, 64.29%, 96.82%, and 98.06%, respectively. The LR and RF algorithm gained the similar accuracy rate is 95.21%, and other performance evaluation results: sensitivity (99.34%), specificity (56.25%), FPR (43.75%), FNR (0.66%), and precision (95.54%). SVM indicate the lowest performance to classify the cervical cancer with an accuracy of 88.62%, 99.29% sensitivity, 33.33% specificity, and 91.90% F1-score. Among the algorithms, both DT and KNN had the highest accuracy at 96.41%, making them the best performers. NB with an accuracy of 95.81%, ranked in the second. LR and RF had the lowest accuracy at 95.21%.

Fig. 6 demonstrates overall performance outcome of Chi-square and LASSO feature selection. In this figure, the performance evaluation results -accuracy, sensitivity, specificity, precision, F1-score, FPR, and FNR are visualized in a bar graph. It is clear that DT algorithm gained 97.60% accuracy in Chi-square, which outreached other applied ml model performance.

In this study, we analyzed the performance of binary classification of cervical cancer using the ROC curve—a plot that illustrates the trade-off between the TPR and the FPR at various threshold settings—on the x and y axes. The TPR measures the proportion of actual positive cases (cervical cancer) correctly identified by the model, whereas the FPR represents the proportion of actual negative cases incorrectly classified as positive.

Figs. 7 and 8 depict the ROC score for Chi-square and LASSO feature selection for applied machine learning models. For Chi-square feature selection, the LR, NB, RF, KNN, and SVM are performed gradually at 0.94,

Table 5
Performance evaluation result after data balancing using ADASYN.

Algorithm name	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)	FPR (%)	FNR (%)
DT	87.43	98.57	29.62	87.97	92.96	70.38	1.43
LR	89.22	98.60	33.33	89.81	93.99	66.67	1.4
NB	85.63	98.54	26.67	85.99	91.84	73.33	1.46
RF	95.21	99.34	56.25	95.54	97.41	43.75	0.66
KNN	91.62	99.31	40.91	91.72	95.36	59.09	0.69
SVM	96.41	99.35	64.29	96.82	98.06	35.71	0.65

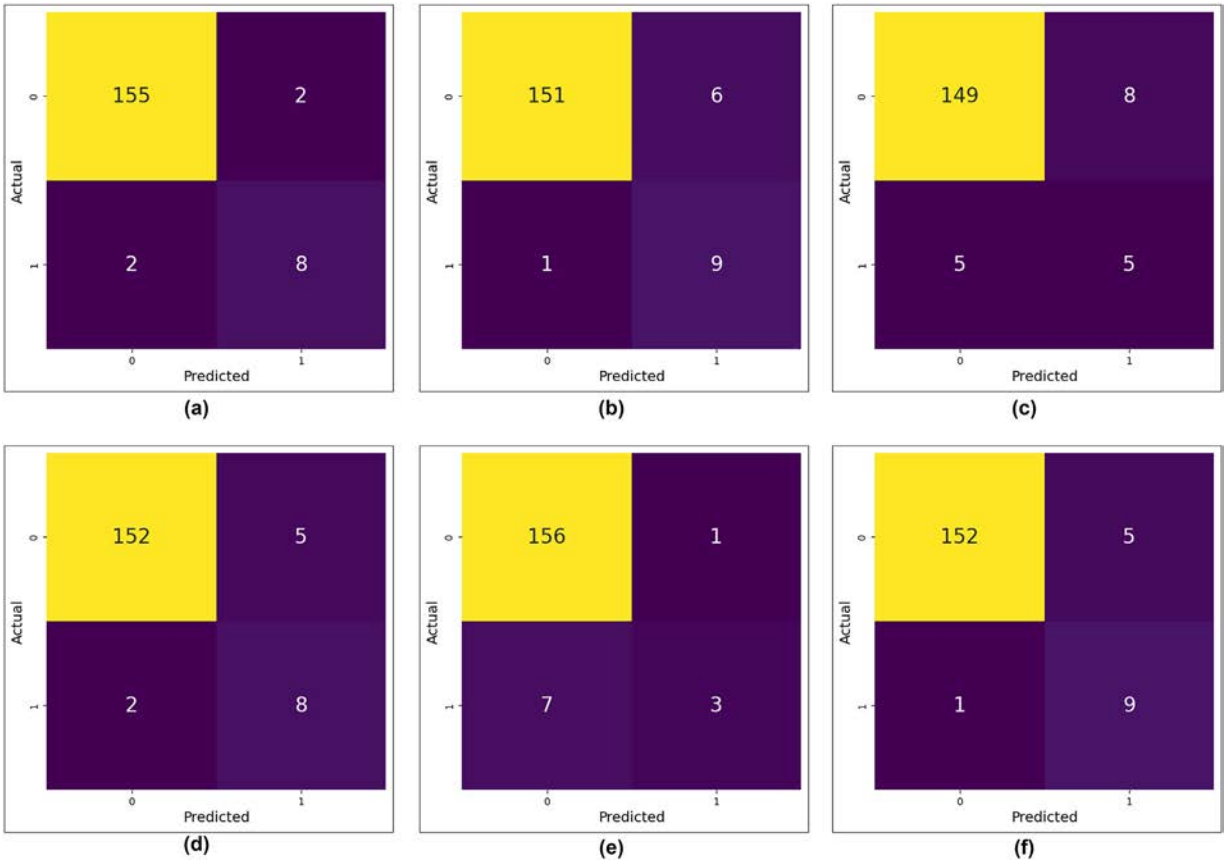

Fig. 4. Confusion matrix of six machine learning algorithm in Chi-square feature selection, (a) DT (b) LR (c) NB (d) RF (e) KNN (f) SVM.

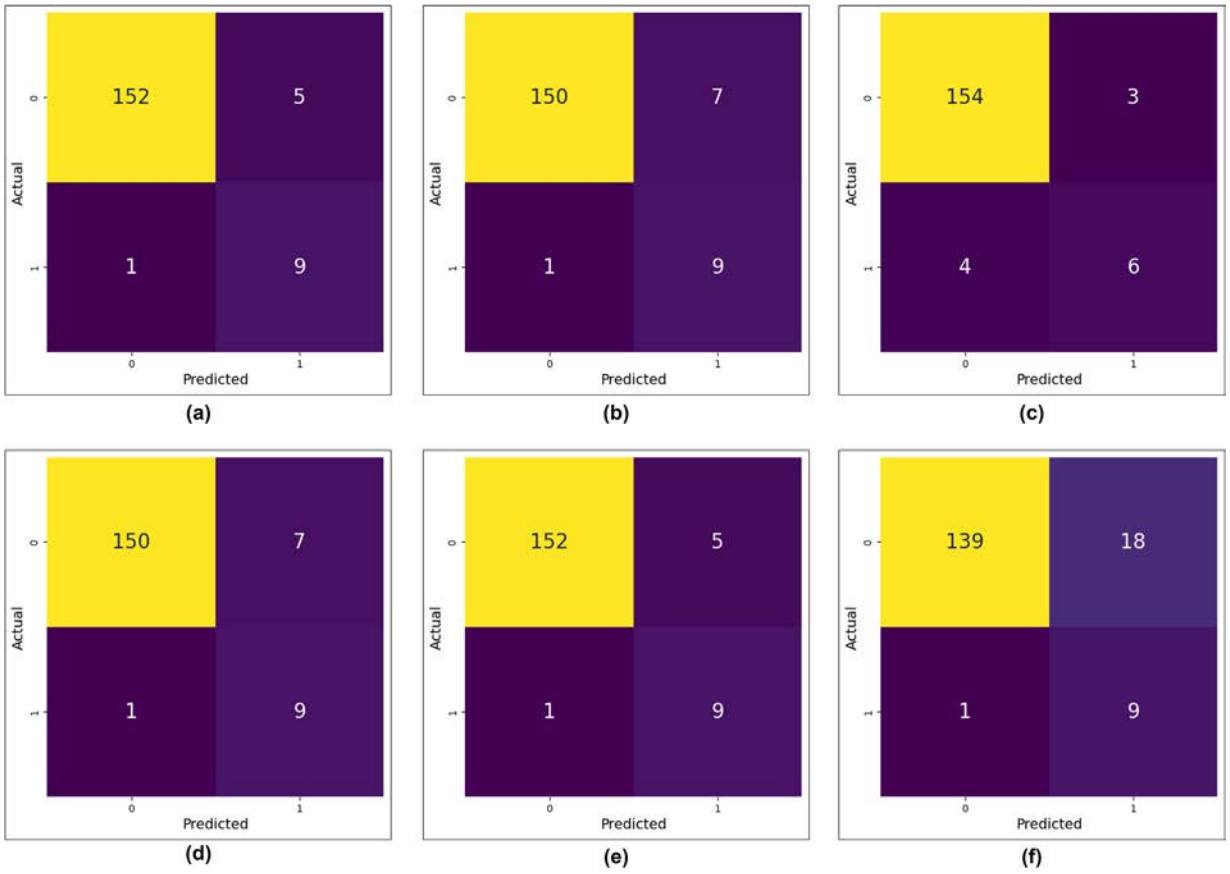

Fig. 5. Confusion matrix of six machine learning algorithm in LASSO feature selection, (a) DT (b) LR (c) NB (d) RF (e) KNN (f) SVM.

Table 6
Performance evaluation result utilizing Chi-square feature selection.

Algorithm name	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)	FPR (%)	FNR (%)
DT	97.60	98.73	80	98.73	98.73	20	1.27
LR	95.81	99.34	60	96.18	97.73	40	0.66
NB	92.21	96.75	38.46	94.90	95.81	61.54	3.25
RF	95.81	98.70	61.54	96.82	97.75	36.46	1.3
KNN	95.21	95.71	75	99.36	97.50	25	4.29
SVM	96.41	99.35	64.29	96.82	98.06	35.71	0.65

Table 7
Performance evaluation result utilizing LASSO feature selection.

Algorithm name	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1-score (%)	FPR (%)	FNR (%)
DT	96.41	99.35	64.29	96.82	98.06	35.71	0.65
LR	95.21	99.34	56.25	95.54	97.40	43.75	0.66
NB	95.81	97.47	66.67	98.08	97.78	33.33	2.53
RF	95.21	99.34	56.25	95.54	97.40	43.75	0.66
KNN	96.41	99.35	64.29	96.82	98.06	35.71	0.65
SVM	88.62	99.29	33.33	85.54	91.90	66.67	0.71

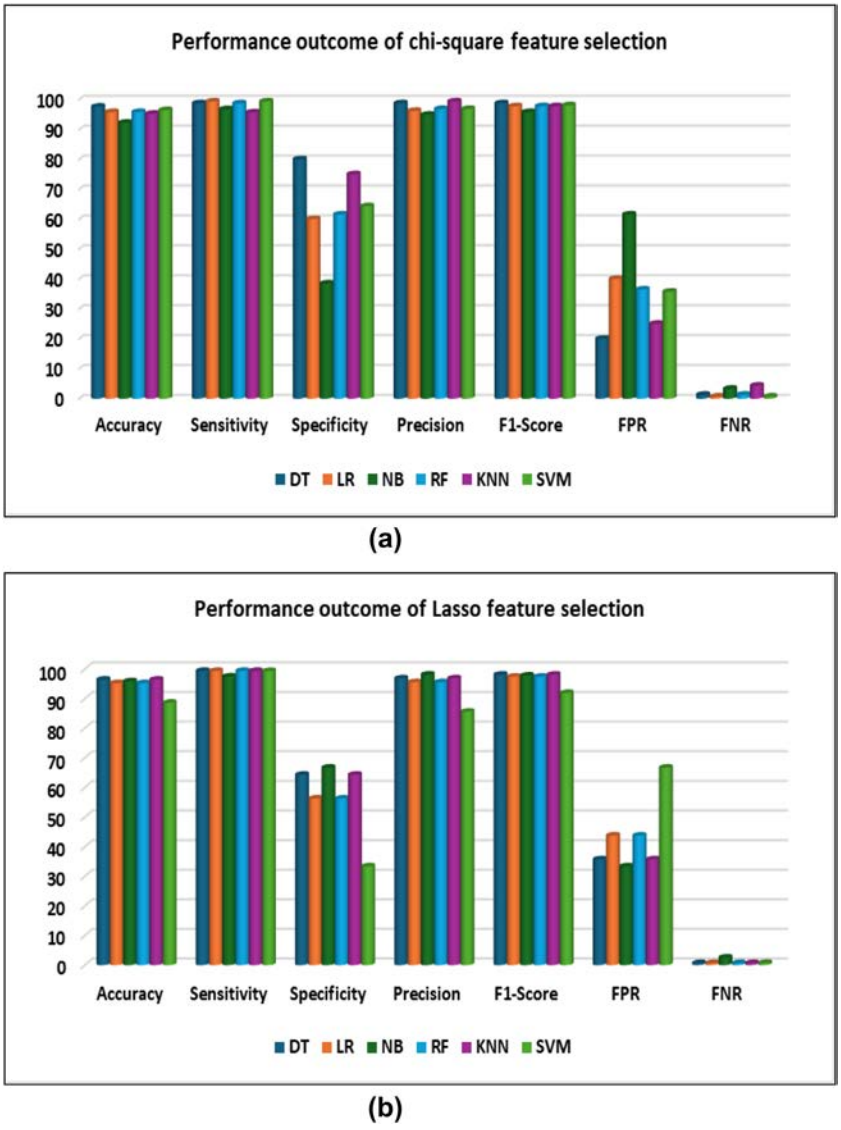

Fig. 6. Comparative result visualization (a) Chi-square feature selection (b) LASSO feature selection.

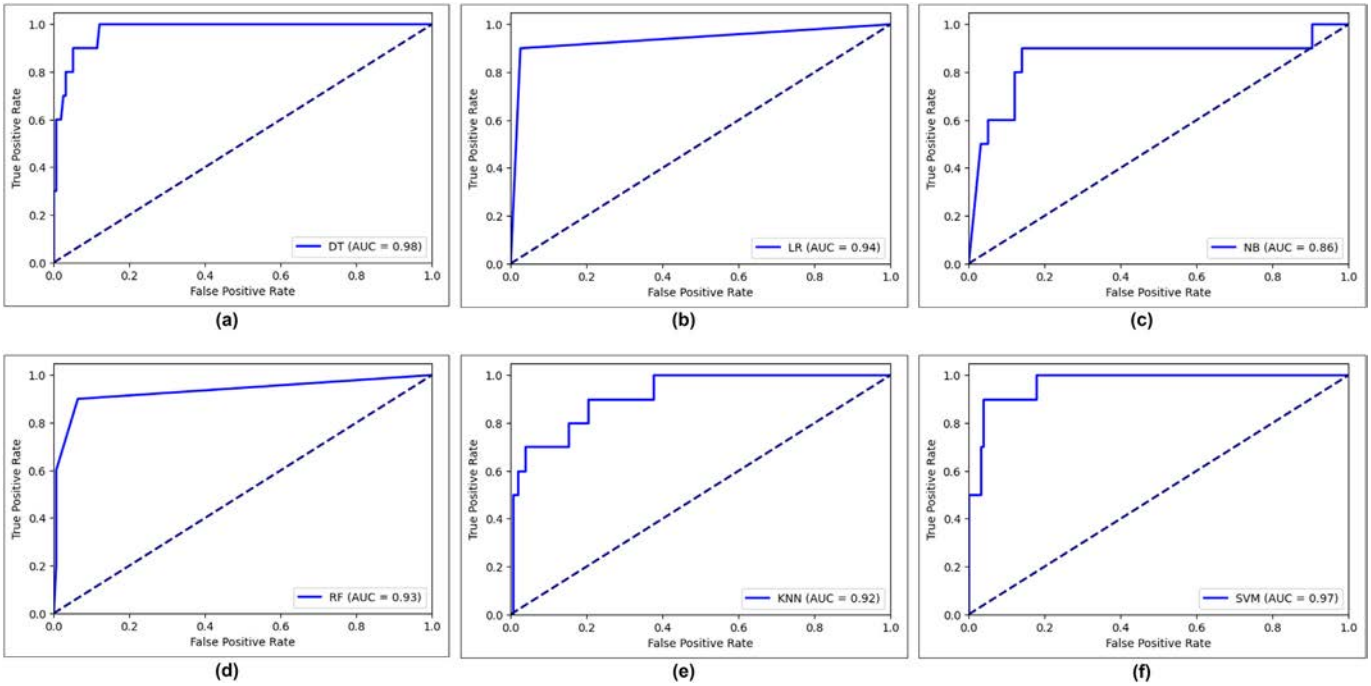

Fig. 7. Receiver operating characteristic (ROC) curve of six machine learning algorithm in Chi-square feature selection, (a) DT (b) LR (c) NB (d) RF (e) KNN (f) SVM.

0.86, 0.93, 0.92, and 0.97, whereas DT outperformed with a 0.98 AUC score. Furthermore, after performing LASSO feature selection, the AUC scores of LR and RF have the same result of 0.89. By contrast, the DT algorithm gained 0.97, which is superior.

Fig. 9, depicts the SHAP summary plot, which consists of a comprehensive visualization of feature importance in a DT model designed to classify cervical cancer. The *x*-axis represents the SHAP values, which quantify the impact of each feature on the model's output, with positive values indicating a higher likelihood of cervical cancer and negative values indicating a lower likelihood, whereas the *y*-axis lists the features analyzed. In this figure, all features rank indicated by color code from low feature value (blue) to high feature value (red). 'Hormonal Contraceptives (years),' which exhibits the highest positive impact to successfully classify cervical cancer and its effect on the model's output, with SHAP values extending up to 1. In addition, 'Age' also shows a significant positive impact, particularly at higher values just over 0.5. 'Number of pregnancies', 'First sexual intercourse', and 'Smoke(packs/year)' have varied impacts, suggesting complex relationships with the target variable. Other features like 'IUD (years)', 'Smokes (years)', and 'STDs' show moderate effects.

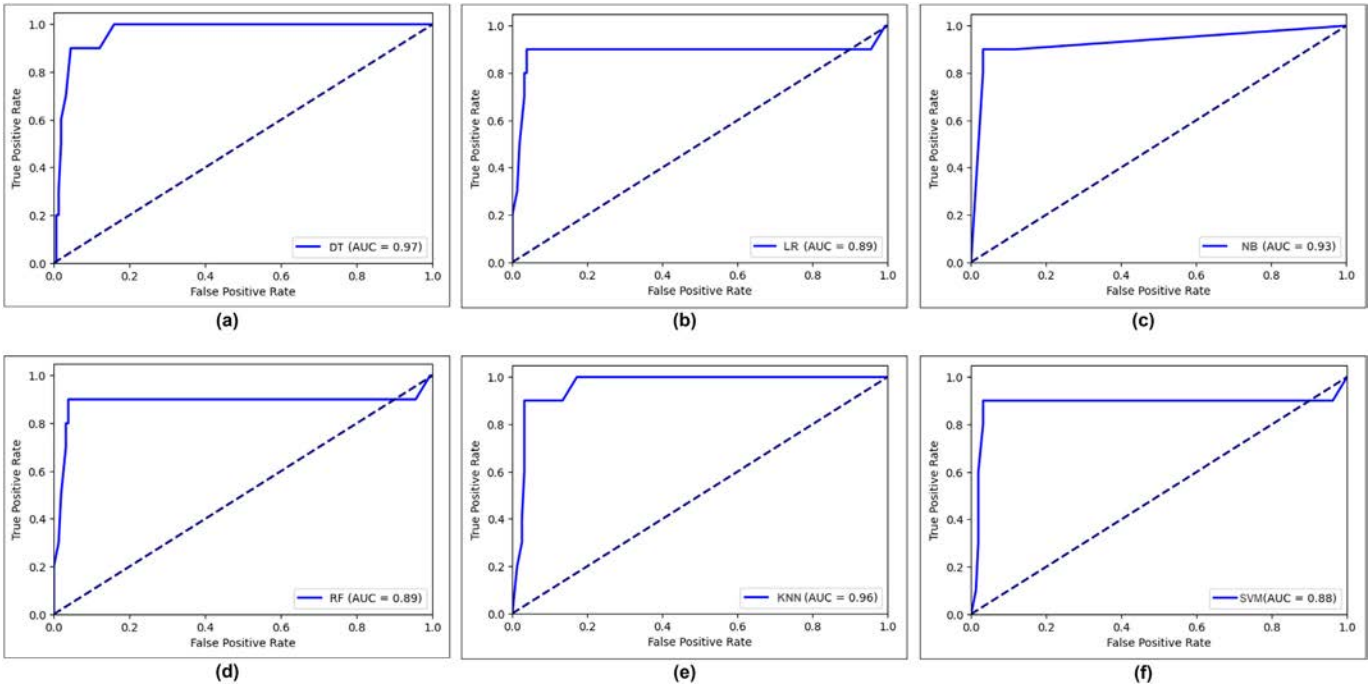

Fig. 8. Receiver operating characteristic (ROC) curve of six machine learning algorithm in LASSO feature selection, (a) DT (b) LR (c) NB (d) RF (e) KNN (f) SVM.

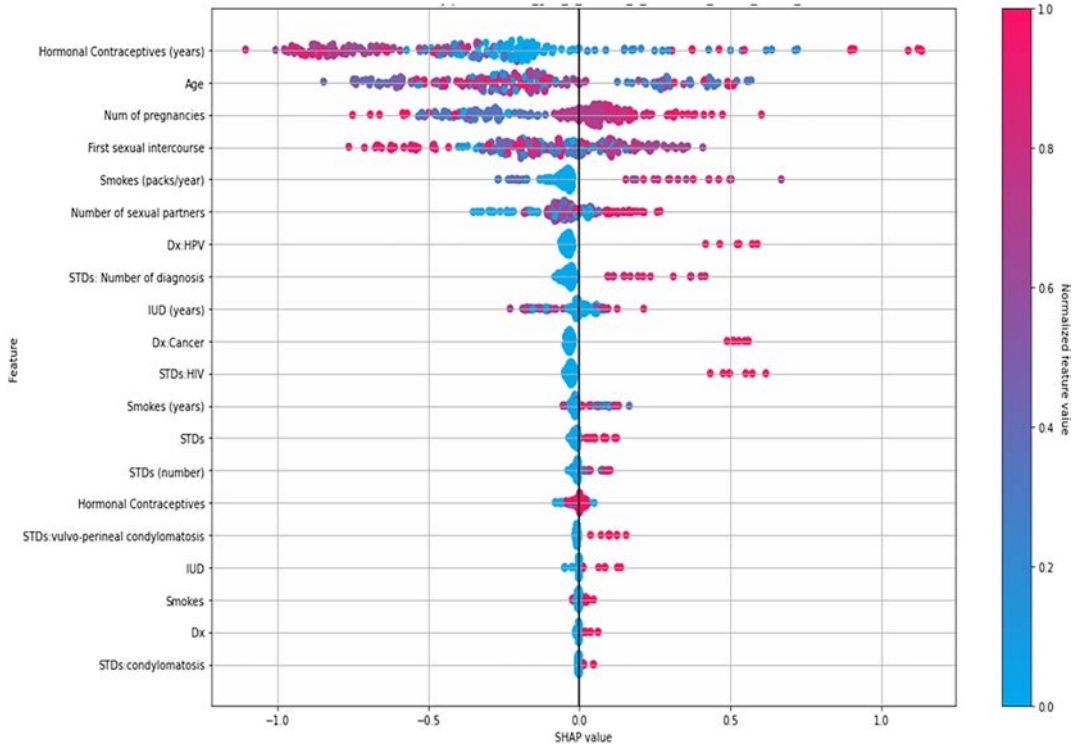

Fig. 9. SHAP explainability values of each feature.

Conclusion

Despite substantial medical advancements, clinicians are finding it more difficult to eradicate the prevalence of cervical cancer mortality rates. Hence, this study presents a machine learning-based system that can detect the early stage of cervical cancer patients depending on the risk factors of cervical cancer. In this research, 858 individuals with 36 distinct attributes were used, which were collected from the UCI machine learning repository. During the data preparation stage, we eliminated missing values by calculating mean, and applied SMOTE and ADASYN techniques to address the issue of data imbalance. Subsequently, Chi-square and LASSO feature selection were used to evaluate the most significant features, which are mostly correlated to identify the particular disease. In addition, six supervised machine learning models have been performed in this research, where the DT algorithm outperformed with 97.60% accuracy, 98.73% sensitivity and 80% specificity for Chi-square feature selection. By contrast, DT gained 97% accuracy, 99.35% sensitivity and 69.23% specificity for the imbalanced dataset. Furthermore, the explainable AI technique (SHAP) was used to extend and check the validity of the model outcome precisely. In the future, we want to extend this study by including a larger real-time dataset and reducing the computation cost. In addition, we will choose some different techniques of XAI to make this ML model more robust.

Declaration of competing interest

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

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68(6):394–424.
2. Arbyn M, Weiderpass E, Bruni L, et al. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. Lancet Glob Health 2020;8(2):e191–e203.
3. W. H. Organization, et al. One-Dose Human Papillomavirus (HPV) Vaccine Offers Solid Protection Against Cervical Cancer. [consultado el 22 de mayo 2023], Disponible en: https://www.who.int/news/item/11-04-2022-one-dose-human-papillomavirus-(hpv)-vaccine-offers-solid-protection-against-cervical-cancer 2022.
4. Lebanova H, Stoev S, Naseva E, et al. Economic burden of cervical cancer in Bulgaria. Int J Environ Res Public Health 2023;20(3):2746.
5. Bruni L, Diaz M, Barrionuevo-Rosas L, et al. Global estimates of human papillomavirus vaccination coverage by region and income level: a pooled analysis. Lancet Glob Health 2016;4(7):e453–e463.
6. W. H. Organization, et al. *Behavioural and Cultural Insights at the Who Regional Office for Europe: Annual Progress Report 2022, Tech. Rep.* World Health Organization. Regional Office for Europe. 2023.
7. Pimple S, Mishra G, Shastri S. Global strategies for cervical cancer prevention. Curr Opin Obstet Gynecol 2016;28(1):4-10.
8. Okunade KS. Human papillomavirus and cervical cancer. J Obstet Gynaecol 2020;40(5):602–608.
9. Issah F, Maree JE, Mwinituo PP. Expressions of cervical cancer-related signs and symptoms. Eur J Oncol Nurs 2011;15(1):67–72.
10. Ali MM, Ahmed K, Bui FM, et al. Machine learning-based statistical analysis for early stage detection of cervical cancer. Comput Biol Med 2021;139, 104985.
11. Spagnoletti BRM, Bennett LR, Keenan C, et al. What factors shape quality of life for women affected by gynaecological cancer in south, south east and east Asian countries? A critical review. Reprod Health 2022;19(1):70.
12. Mehmood M, Rizwan M, Gregus ml M, Abbas S. Machine learning assisted cervical cancer detection. Front Public Health 2021;9, 788376.
13. Arora A, Tripathi A, Bhan A. Classification of cervical cancer detection using machine learning algorithms. 2021 6th International Conference on Inventive Computation Technologies (ICICT). IEEE; 2021. p. 827–835.
14. Unlersen MF, Sabanci K, Özcan M. Determining cervical cancer possibility by using machine learning methods. Int J Latest Res Eng Technol 2017;3(12):65–71.
15. Razali N, Mostafa SA, Mustapha A, Abd Wahab MH, Ibrahim NA. Risk factors of cervical cancer using classification in data mining. J Phys Conf Ser 2020;1529.IOP Publishing. p. 022102.
16. Parikh D, Menon V. Machine learning applied to cervical cancer data. Int J Math Sci Comput 2019;5(1):53–64.
17. Ou Z, Mao W, Tan L, et al. Prediction of postoperative pathologic risk factors in cervical cancer patients treated with radical hysterectomy by machine learning. Curr Oncol 2022;29(12):9613–9629.
18. Malli PK, Nandyal S. Machine learning technique for detection of cervical cancer using k-nn and artificial neural network. Int J Emerg Trends Technol Comput Sci (IJETTCS) 2017;6(4):145–149.
19. Latha DS, Lakshmi P, Fathima S. Staging prediction in cervical cancer patients–a machine learning approach. Int J Innov Res Pract 2014;2(2):14–23.

20. Sun G, Li S, Cao Y, Lang F. Cervical cancer diagnosis based on random forest. Int J Perform Eng 2017;13(4):446.
21. Ilyas QM, Ahmad M. An enhanced ensemble diagnosis of cervical cancer: a pursuit of machine intelligence towards sustainable health. IEEE Access 2021;9:12374–12388.
22. Nithya B, Ilango V. Machine learning aided fused feature selection based classification framework for diagnosing cervical cancer. 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC). IEEE; 2020. p. 61–66.
23. R. Vidya, G. Nasira, Predicting cervical cancer using machine learning techniques—an analysis, Glob J Pure Appl Math 12 (3).
24. Tseng C-J, Lu C-J, Chang C-C, Chen G-D. Application of machine learning to predict the recurrence-proneness for cervical cancer. Neural Comput Appl 2014;24:1311–1316.
25. Chaudhuri AK, Ray A, Banerjee DK, Das A. A multi-stage approach combining feature selection with machine learning techniques for higher prediction reliability and accuracy in cervical cancer diagnosis. Int J Intel Syst Appl 2021;10(5):46.
26. Yang W, Gou X, Xu T, Yi X, Jiang M. Cervical cancer risk prediction model and analysis of risk factors based on machine learning. Proceedings of the 2019 11th International Conference on Bioinformatics and Biomedical Technology; 2019. p. 50–54.
27. Kalbhor M, Shinde SV, Jude H. Cervical cancer diagnosis based on cytology pap smear image classification using fractional coefficient and machine learning classifiers. TELKOMNIKA (Telecommun Comput Elect Control) 2022;20(5):1091–1102.
28. Cervical Cancer (Risk Factors) Dataset. accessed: 2024-06-20. URL: https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors March 02, 2017.
29. Fernandes K, Cardoso JS, Fernandes J. Transfer learning with partial observability applied to cervical cancer screening. Pattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Faro, Portugal, June 20–23, 2017, Proceedings 8. Springer; 2017. p. 243–250.
30. L. O. Joel, W. Doorsamy, B. S. Paul, On the performance of imputation techniques for missing values on healthcare datasets, arXiv preprint arXiv:2403.14687.
31. Ali MS, Hossain MM, Kona MA, Nowrin KR, Islam MK. An ensemble classification approach for cervical cancer prediction using behavioral risk factors. Healthcare Anal 2024;5, 100324.
32. Munshi RM. Novel ensemble learning approach with SVM-imputed ADASYN features for enhanced cervical cancer prediction. PLoS One 2024;19(1), e0296107.
33. Li X, Ning R, Xiao B, et al. A multi-variable predictive warning model for cervical cancer using clinical and SNPS data. Front Med 2024;11, 1294230.
34. Chauhan R, Goel A, Alankar B, Kaur H. Predictive modeling and web-based tool for cervical cancer risk assessment: a comparative study of machine learning models. MethodsX 2024;12, 102653.
35. Hu J, Liu G, Liu Y, Yuan M, Zhang F, Luo J. Predicting lower limb lymphedema after cervical cancer surgery using artificial neural network and decision tree models. Eur J Oncol Nurs 2024, 102650.
36. S. Devi, R. Gangarde, S. Deokar, S. F. Muqeemuddin, S. R. Awasthi, S. Shekhar, R. Sonchhatra, S. Joshi, Public health nurse perspectives on predicting nonattendance for cervical cancer screening through classification, ensemble, and deep learning models, Public Health Nurs.
37. Narayana T, Nalini N. Prediction of fetal heart disease detection using naive bayes classifier and comparing with linear regression classifier. AIP Conference Proceedings. AIP Publishing; 2024.
38. Shakya S, Joby P. Heart disease prediction using fog computing based wireless body sensor networks (WSNS). IRO J Sustain Wireless Syst 2021;3(1):49–58.
39. Akter B, Shakil R, Rajbongshi A, Sara U, Barman MR. Utilization of five-distinct dataset to diagnose and predict heart disease: a machine learning approach. 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE; 2022. p. 1–6.
40. Wang H, Shao Y. Fast generalized ramp loss support vector machine for pattern classification. Pattern Recogn 2024;146, 109987.
41. Islam T, Sheakh MA, Tahosin MS, et al. Predictive modeling for breast cancer classification in the context of bangladeshi patients by use of machine learning approach with explainable ai. Sci Rep 2024;14(1):8487.