BIMA
BULLETIN OF INTELLIGENT MACHINES AND ALGORITHMS
Vol. 1 No. 3 March 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v1i3.16

Machine Learning Based Cervical Cancer Risk Prediction with SHAP-Driven Feature Interpretation

Fachrizal Ardiansyah[1*], Raka Deny Abdi Putra[2], Budiman[3]

Informatics[1]
Universitas Informatika dan Bisnis Indonesia, Bandung, Indonesia[1,2,3]
https://unibi.ac.id/[1,2,3]
fachrizalardiansyah23@student.unibi.ac.id[1*], rakadenyabdiputra23@student.unibi.ac.id[2], budiman@unibi.ac.id

Abstract. Cervical cancer remains a critical public health problem, particularly in developing countries where early detection is often limited. This study presents a machine learning–based approach for cervical cancer risk prediction that emphasizes both predictive accuracy and interpretability. Several supervised algorithms, namely K-Nearest Neighbors, Random Forest, XGBoost, and CatBoost, were evaluated using the Cervical Cancer (Risk Factors) dataset from the UCI Machine Learning Repository following comprehensive data preprocessing and systematic hyperparameter optimization. The experimental results show that CatBoost achieved the best overall performance, with an optimized accuracy of 97.01% and improved sensitivity in detecting high-risk cases, supported by stable k-fold cross-validation results. To enhance clinical transparency, explainable artificial intelligence was incorporated via SHAP, revealing that key predictors such as the Schiller test, age, and reproductive factors played dominant roles in the model's decisions. These findings demonstrate that the proposed framework is not only accurate and stable but also interpretable and clinically relevant, making it well-suited to support early detection and decision-making in cervical cancer screening, especially in resource-limited healthcare settings.

Keywords: Cervical cancer, machine learning, risk prediction, explainable AI (XAI), CatBoost

1. INTRODUCTION

Cervical cancer remains one of the major global public health challenges, contributing substantially to female morbidity and mortality, particularly in developing countries [1]. According to data from the World Health Organization [2], cervical cancer ranked fourth among the leading causes of cancer-related deaths in women worldwide in 2022, with more than 660,000 new cases and approximately 350,000 deaths reported annually. In Indonesia, the burden of cervical cancer is even more alarming, primarily due to limited access to early detection services and low public awareness regarding the importance of routine cervical screening. As a result, a large proportion of cases are diagnosed at advanced stages, highlighting the urgent need for innovative, accurate, and practical early detection strategies that leverage artificial intelligence and predictive analytics to support modern public health systems [3], [4].

One of the main challenges in developing cervical cancer prediction models lies in the quality and heterogeneity of medical data [5]. Such datasets are often affected by class imbalance, missing values, and complex correlations among clinical, behavioral, and demographic variables. In addition, the selection of appropriate machine learning algorithms, including K-Nearest Neighbors, Random Forest, and Gradient Boosting methods, plays a critical role in determining predictive performance [6]. Feature selection and data preprocessing procedures are therefore essential for building models that are stable, interpretable, and clinically reliable [7]. Recent advances emphasize the importance of integrating explainable artificial intelligence (XAI) techniques to improve transparency and trust in medical predictive models, particularly in high-stakes decision-making environments such as cancer risk assessment [8].

In response to these challenges, this study aims to develop a cervical cancer risk prediction model to support early detection and clinical decision-making by comparing multiple machine learning algorithms. The proposed approach focuses not only on predictive accuracy but also on model interpretability, enabling healthcare professionals to understand better the underlying risk factors and the rationale for predictions [9], [10]. Furthermore, by utilizing clinical and lifestyle variables that are relatively easy to obtain, the proposed framework offers practical advantages for deployment in resource-limited settings. Such an approach has the potential to contribute to evidence-based screening programs and inform national health policies for cervical cancer

Vol. 1 No. 3 March 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v1i3.16

prevention, particularly in regions with limited laboratory infrastructure [11], [12].

The contribution of this research lies in integrating machine learning approaches with explainable artificial intelligence (XAI) principles to enhance transparency and trust in the application of predictive models in the medical domain. By delivering a comprehensive analysis of relevant risk factors and the underlying relationships among variables, the proposed approach is expected to support data-driven screening processes and facilitate faster and more targeted clinical decision-making. Moreover, the development of adaptive and context-aware predictive models offers practical advantages for real-world implementation. It holds the potential to strengthen cervical cancer prevention strategies, particularly in healthcare settings with limited diagnostic and laboratory resources.

Overall, this study emphasizes the critical role of artificial intelligence and predictive analytics in reinforcing early detection strategies for cervical cancer. By integrating medical data, statistical modeling, and machine intelligence, the proposed framework aims to provide an innovative, accurate, and practically applicable solution. In the context of the growing cancer burden at both national and global levels, this effort represents a strategic step toward a more proactive, predictive, and data-driven healthcare system, aligned with the broader vision of digital transformation in public health.

2. RELATED WORK

Machine learning for cervical cancer risk prediction is advancing rapidly, driven by the growing availability of medical data and advances in predictive analytics. Several previous studies have demonstrated that machine learning models can aid in the early detection of cervical cancer by analyzing clinical and behavioral risk factors. Despite the high accuracy often reported, various challenges remain related to interpretability, model generalization, and integrating prediction results in real-world clinical decision-making.

Tanimu et al. [13] classified cervical cancer using various machine learning algorithms, including Random Forest, Naïve Bayes, and Support Vector Machine, on the Cervical Cancer (Risk Factors) dataset from the UCI Machine Learning Repository. The study showed that Random Forest achieved the best performance with an accuracy of over 95%. However, this research still focuses on evaluating model performance without including an in-depth analysis of feature importance or the contribution of predictor variables to the classification results. As a result, even though the model achieves high performance, the clinical interpretation of the predictions remains limited.

Research by Chadaga et al. [14] introduced a stacked ensemble learning approach to predict outcomes of cervical cancer biopsies using demographic and epidemiological data. The results show an increase in accuracy compared to single models, confirming the superiority of ensemble learning in capturing the complexity of medical data. However, the study has not yet integrated an explainable AI (XAI) approach to make the model's decision-making mechanisms transparent. In addition, there has been no sensitivity analysis of the model to feature variations, which could provide a deeper understanding of the stability and reliability of the results.

Kılıçarslan et al. [15] addressed class imbalance by applying the Synthetic Minority Oversampling Technique (SMOTE) before building an ML model for cervical cancer prediction. The study shows that applying oversampling techniques can improve the model's sensitivity to positive cases of cervical cancer, which are generally underrepresented in the dataset. However, this study has not yet examined in depth the relationship between variables or feature importance, which could provide further insight into the main risk factors. Therefore, there is an opportunity to combine a data balancing approach with interpretive analysis so that the model results are not only accurate but also clinically meaningful.

Okyay et al. [16] evaluated cervical cancer risk using several machine learning algorithms, including Decision Tree, Random Forest, and Logistic Regression, on the Cervical Cancer (Risk Factors) dataset from UCI. This study makes an important contribution through a comparative analysis of algorithms and performance validation using various evaluation metrics. However, two main limitations remain open for further development. First, the study does not systematically highlight aspects of model optimization, such as hyperparameter tuning or more extensive cross-validation, that can improve the consistency of results across data subsets. Second, the study has not implemented explainable AI (XAI) based interpretation mechanisms to reveal the role of the most influential variables in determining risk levels. Thus, although the model's performance is quite high, the transparency and clarity of the reasons for the predictions still need to be strengthened so that the results can be adopted clinically.

Recent research by Dong et al. [17] introduces the SMART HPV model, a full-genotyping high-risk HPV testing approach developed for risk stratification and cervical cancer screening management. This study represents a significant step forward by conducting direct clinical validation in a real-world healthcare setting. However, the focus of this research is primarily on HPV molecular data rather than on integrating behavioral and sociodemographic data, leaving room for further exploration of multidimensional data.

Based on this review, previous studies still face several limitations that this study addresses. Most prior work has focused on improving the accuracy of prediction models but has not adequately considered interpretability, a crucial element in clinical applications. In addition, comprehensive model optimization approaches have not been widely applied, especially those that incorporate feature selection, hyperparameter tuning, and repeated

Vol. 1 No. 3 March 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v1i3.16

evaluation using k-fold cross-validation to ensure the stability and reliability of model performance. On the other hand, the application of explainable AI (XAI) techniques in cervical cancer risk prediction remains relatively limited, so the relationship between predictor variables and classification outcomes has not been fully elucidated. These limitations indicate the need to develop predictive approaches that are not only computationally efficient but also provide transparent, medically accountable insights.

3. METHODS

Figure 1 illustrates the stages of the research and describes the methodology used in this study in a systematic and structured manner. The research stages were designed to ensure that the data processing, model development, and evaluation were carried out in an integrated and replicable manner. Each stage was designed to support the next stage, resulting in a consistent, transparent, and scientifically valid analysis process.

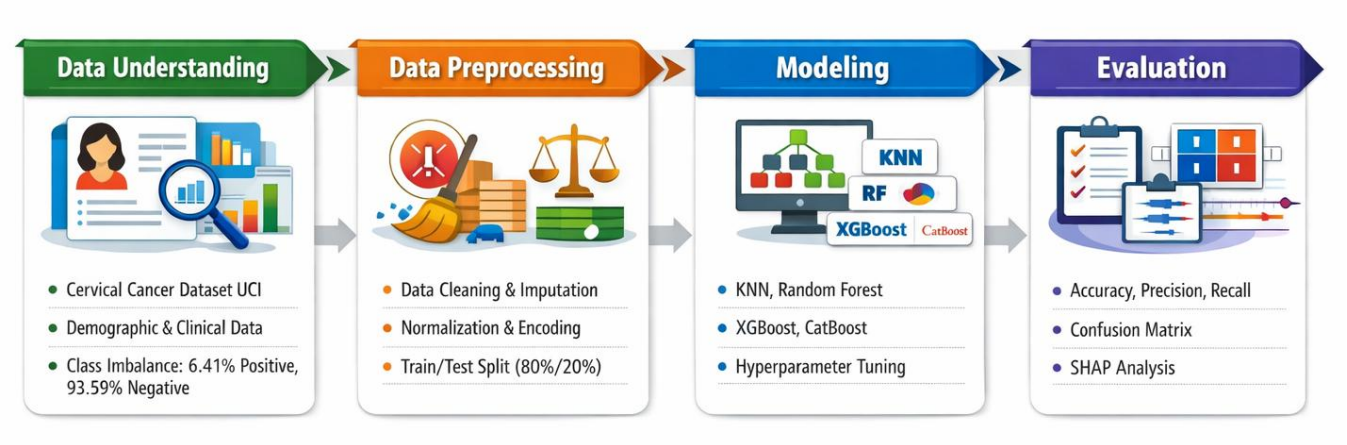

Fig.1. Research Methodology

3.1. Data Understanding

The dataset used in this study was obtained from the *Cervical Cancer (Risk Factors)* dataset available in the UCI Machine Learning Repository. This dataset contains clinical and behavioral data on female patients, including demographic characteristics, lifestyle habits, medical history, and previous screening results. The data collection process involved verifying the completeness, consistency, and integrity of the data to ensure that all attributes used in the modeling process met the required quality standards. The class distribution indicates that positive cases account for 6.41% of the data, while negative cases constitute 93.59%, reflecting a significant class imbalance. A complete list of variables contained in this research dataset is presented in Table 1.

TABLE 1. Dataset Feature

Attribute name	Description
Age	Indicates the age of the patient (woman)
Number of sexual partners	Expresses the total count of people with whom the woman is engaged in sexual activity together
First sexual intercourse	Denotes the age at which the woman had her first sexual intercourse
Number of pregnancies	Denotes the total number of childbirths by the woman
Smokes	Either the woman smokes (one) or not (zero)
Smokes (years)	Denotes the total smoking period (in years)
Smokes (packs/year)	Depicts the annual count of cigarette packets consumed
Hormonal contraceptives	Indicates whether the woman takes hormonal contraceptives or not
Hormonal contraceptives (years)	It indicates the duration for which the contraceptive method has been used.
IUD	Indicates whether the intrauterine contraceptive device was used (one) or not (zero)
IUD (years)	Denotes the total usage period of IUD (in years)
STDs	Indicates the presence of STDs—either yes (one) or no (zero)
STDs (number)	This numerical feature shows the overall count of STDs detected in the patient.
STDs: Condylomatosis	Presence of condylomatosis with the patient—expressed in ones (yes) and zeros (no)
STDs: cervical condylomatosis	Presence of cervical condylomatosis—expressed in ones (yes) and zeros (no)
STDs: vaginal condylomatosis	Presence of vaginal condylomatosis—expressed in ones (yes) and zeros (no)
STDs: vulva-perineal condylomatosis	Presence of vulvo-perineal condylomatosis—expressed in ones (yes) and zeros (no)
STDs: syphilis	Presence of syphilis—expressed in ones (yes) and zeros (no)
STDs: pelvic inflammatory disease	Presence of pelvic inflammatory diseases—expressed in ones (yes) and zeros (no)
STDs: genital herpes	Presence of genital herpes—expressed in ones (yes) and zeros (no)
STDs: molluscum contagiosum	Presence of molluscum contagiosum—expressed in ones (yes) and zeros (no)
STDs: AIDS	Presence of AIDS—expressed in ones (yes) and zeros (no)
STDs: HIV	Presence of HIV infection—expressed in ones (yes) and zeros (no)

Vol. 1 No. 3 March 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v1i3.16

Attribute name	Description
STDs: Hepatitis B	Presence of Hepatitis B virus infection in the patient—expressed in ones (yes) and zeros (no)
STDs: HPV	Presence of HPV in the patient—expressed in ones (yes) and zeros (no)
STDs: Number of diagnoses	Indicates the total number of times the STDs have been diagnosed
STDs: Time since first diagnosis	Indicates the total number of years since the first diagnosis
STDs: Time since last diagnosis	Specifies the length of time since the last diagnosis
Dx: Cancer	Indicates the presence of cancer after the diagnosis
Dx: CIN	Indicates the presence of CIN after the diagnosis
Dx: HPV	Indicates the presence of HPV after the diagnosis
Dx	It indicates the presence of cancer.
Hinselmann	Hinselmann or colposcopy is an intensive medical test that uses magnification to identify abnormalities in cervical cells, vagina, and vulva accurately.
Schiller	The Schiller iodine test is a medical test in which iodine solution is applied to the cervix to diagnose cervical cancer.
Cytology	The Pap smear test helps detect abnormal cells in the cervix.
Biopsy (TARGET)	A surgical procedure to examine a small amount of tissue removed from the cervix to determine whether a carcinogenic cell is present (one) or not (zero)

3.2. Data Preprocessing.

The preprocessing stage began with data cleaning, which aimed to handle missing values, inconsistencies, and outliers that could bias the learning process. Missing values were handled using statistical imputation techniques, such as mean or mode replacement, depending on the data type. At the same time, extreme outliers were identified using interquartile range (IQR) analysis and handled appropriately to maintain data validity. After cleaning, a standardization process was performed using the *StandardScaler method to normalize numerical features, ensuring that all variables were on* a comparable scale. This step was essential to prevent features with larger numerical ranges from dominating the learning process.

Subsequently, categorical variables were encoded. For nominal features, *one-hot encoding* was applied to generate binary representations, whereas ordinal features were encoded numerically according to their inherent order. The resulting dataset was then split into training and testing sets, with 80% for training and 20% for testing. The training set was used for model learning, while the testing set was used for independent performance evaluation to assess generalization capability.

3.3. Modeling

The modeling stage involved implementing and comparing several supervised *machine learning* algorithms, including *KNN, Random Forest*, *XGBoost*, and Cat*Boost*. Each model was trained on the preprocessed data and optimized via hyperparameter tuning via grid search and cross-validation to identify the configuration that yields the best predictive performance. The modeling process was conducted in Python using the *scikit-learn* framework to ensure reproducibility and methodological rigor.

In the K-Nearest Neighbors (KNN) algorithm, classification is determined based on the majority label among the k nearest data points in the feature space. The similarity between data points is measured using the Euclidean distance metric, formulated as:

$$d(x_i, x_j) = \sqrt{\sum_{n=1}^{p} (x_i - x_j)^2} \quad (1)$$

Equation (1) represents the Euclidean distance measurement used in the K-Nearest Neighbors (KNN) algorithm to calculate the degree of proximity between two data samples in a p-dimensional feature space. Mathematically, the distance between samples xi and xj is calculated as the square root of the sum of the squares of the differences of each corresponding feature component. Thus, the more conceptually accurate notation refers to the differences of each feature-k of the two samples. This distance value is the basis for determining the nearest neighbor k that has a minimum distance to the test data.

$$\hat{y} = mode(y_i),\ i \in K \quad (2)$$

Furthermore, as stated in equation (2), the final prediction class $\hat{y}$ is obtained through a majority voting mechanism, which is by selecting the label that appears most frequently among the selected k neighbors. This approach confirms that classification decisions in KNN are entirely dependent on the local structure of data distribution around the observation point.

$$\text{Obj} = \sum_{i=1}^{n} L\left(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)\right) + \Omega(f_t) \quad (3)$$

Equation (3) describes the objective function in the Extreme Gradient Boosting (XGBoost) algorithm, which is designed to minimize prediction errors iteratively by adding new decision trees at each iteration t. The objective function consists of two main components, namely the loss function L, which measures the difference between the actual value yi and the updated prediction $y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)$, and the regularization component $\Omega(f_t)$, which controls the model's complexity. The function $f_t(x_i)$ represents the contribution of the new tree added to

Vol. 1 No. 3 March 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v1i3.16

improve the residual error from the previous iteration. With this approach, XGBoost not only focuses on accuracy, but also maintains a balance between bias and variance through model complexity control.

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (4)$$

The regularization component in Equation (4) is formulated as a combination of penalties on the number of tree leaves T and penalties on the weight of each leaf wj. The parameter γ regulates the penalty on the complexity of the tree structure, while λ controls the amount of regularization on the leaf weight. This mechanism aims to prevent the model from becoming too complex and overfitting to the training data, thereby improving its generalization ability on new data.

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \lambda \sum_{t=1}^{T} |f_t|^2 \quad (5)$$

Equation (5) states the loss function in the CatBoost algorithm, which is also based on the boosting technique. The function is the sum of the total prediction loss across all samples and the regularization component for model complexity, expressed as a penalty on the tree function ft. Unlike conventional boosting approaches, CatBoost employs a categorical feature processing strategy and an ordered boosting mechanism to mitigate prediction shift and bias from target leakage. Therefore, this mathematical formulation aims to minimize prediction errors while maintaining model stability, leading to more consistent performance, especially on medical datasets with complex feature distributions and potential class imbalances.

3.4. Evaluation

Finally, the evaluation phase was carried out to assess the predictive capability and robustness of each model. Performance metrics such as *accuracy*, *precision*, *recall*, and F1-score were used to comprehensively assess the model's effectiveness in distinguishing between high- and low-risk individuals. Additionally, confusion matrix analysis was applied to evaluate classification errors and identify potential biases in prediction outcomes. To enhance interpretability, *feature-importance analysis and explainable AI (XAI) techniques, including SHAP (Shapley Additive exPlanations), were employed to identify and visualize the contributions* of individual features to model decisions.

Through this systematic methodological framework, which combines rigorous data preprocessing, robust model optimization, and transparent interpretative analysis, the study aims to develop a predictive model that is not only accurate and reliable but also interpretable and clinically meaningful to support data-driven cervical cancer risk assessment.

4. RESULTS AND DISCUSSIONS

4.1. Results

The evaluation results show that all applied machine learning algorithms achieved a relatively high level of accuracy in predicting cervical cancer risk. However, analysis of the evaluation metrics, particularly precision, recall, and F1-score in the minority class, revealed differences in performance characteristics between algorithms. This confirms that accuracy alone is insufficient to assess the effectiveness of models in medical contexts where classification errors are sensitive. Algorithms are shown in Table 2

TABLE 2. Results of All Models

Model	Biopsy	Precision	Recall	F1-Score	Accuracy
KNN	No	0.95	0.99	0.97	0.9401
	Yes	0.50	0.10	0.17	
Random Forest	No	0.97	0.99	0.98	0.9641
	Yes	0.75	0.60	0.67	
XGBoost	No	0.97	0.98	0.97	0.9521
	Yes	0.62	0.50	0.56	
CatBoost	No	0.98	0.97	0.98	0.96
	Yes	0.64	0.70	0.67	

The K-Nearest Neighbors (KNN) model achieved an accuracy of 94.01%, which appears numerically competitive. However, this performance was dominated by the model's ability to classify the majority class (non-risk), as reflected in the very low recall value for the high-risk class (0.10). This condition shows that KNN is less effective in detecting positive cases of cervical cancer, which is the main focus of early detection systems. This phenomenon indicates the limitations of KNN in handling class imbalance and high sensitivity to data distribution and parameter selection κ.

Instead, Random Forest showed a more balanced performance improvement, with an accuracy of 96.41% and a positive class recall of 0.60. The combination of decision trees allows Random Forest to capture nonlinear patterns more effectively. The higher macro-average F1-score compared to KNN indicates that this model has better generalization across both classes, making it more suitable for prediction.

The XGBoost model achieved an accuracy of 95.21%, with fairly good performance but relatively lower on minority classes compared to Random Forest. Although the gradient boosting mechanism allows the model to learn residual errors iteratively, the positive class recall of 0.50 indicates

Vol. 1 No. 3 March 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v1i3.16

that the model is still not optimal at identifying all high-risk cases. This may be due to XGBoost's sensitivity to regularization parameters and model complexity, which requires further optimization to balance bias and variance.
CatBoost's performance was the most consistent among non-optimization models, with an accuracy of 96% and a positive class recall of 0.70. This advantage demonstrates the effectiveness of the boosting approach in reducing prediction shift and in learning feature interactions more stably. The relatively high macro-average F1-score indicates that CatBoost is more adaptive to imbalanced data distributions, making it a strong candidate for cervical cancer risk prediction systems.
Based on the comparative evaluation of all implemented machine learning models, CatBoost demonstrated the most consistent predictive performance, particularly in balanced accuracy and robustness in identifying high-risk cases. Consequently, CatBoost was selected as the primary model for further optimization through hyperparameter tuning and for interpretability analysis using explainable artificial intelligence (XAI) techniques.

TABLE 3. Results of CatBoost Optimized

Biopsy	Precision	Recall	F1-Score	Accuracy
No	0.98	0.99	0.98	0.9701
Yes	0.78	0.70	0.74	

Improvement was observed in Table 3 for CatBoost, using hyperparameter tuning with GridSearchCV, with an accuracy of 97.01% and positive-class precision and recall of 0.78 and 0.70, respectively. These results show that the hyperparameter tuning process plays an important role in improving model sensitivity without sacrificing specificity. Additionally, the k-fold cross-validation results show an average accuracy of 95.45% with a low standard deviation (0.0140), indicating the stability and consistency of the model's performance across variations in the training data. These findings reinforce the model's validity and minimize the risk of overfitting.

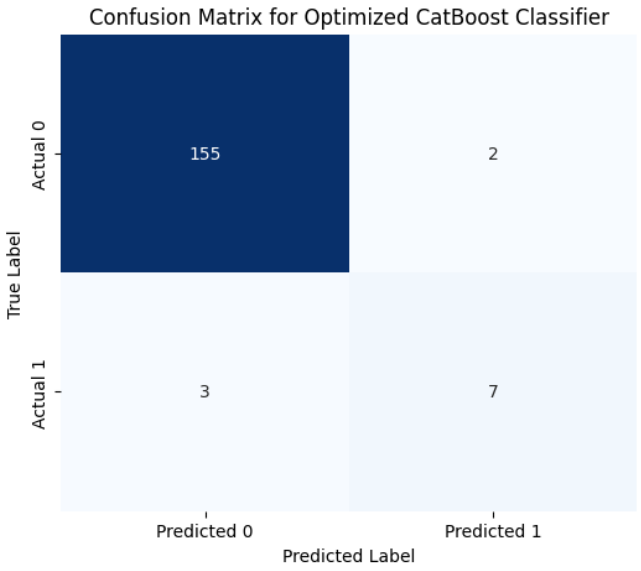

Fig.2. Confusion Matrix for optimized CatBoost model

Fig 2 shows the confusion matrix of the optimized CatBoost model, which is used to evaluate its performance in classifying cervical cancer into two categories: class 0 (no risk) and class 1 (risk). In this matrix, the vertical axis shows the actual labels, while the horizontal axis shows the model's predicted labels. Based on the results shown, 155 samples in the actual class 0 are correctly predicted as class 0, thus falling into the true negative category. In addition, 2 samples from the actual class 0 are predicted as class 1, which fall into the false positive category. In the positive class, the model correctly identified 7 samples as class 1 (true positives). In comparison, 3 samples from the actual class 1 were predicted as class 0, which are false negatives. This distribution of values shows that the model correctly classified most samples in the majority class. In the minority class, the model detected some positive cases, though there were still classification errors. This confusion matrix provides a detailed overview of the model's prediction patterns for each class and serves as the basis for calculating evaluation metrics such as precision, recall, and F1-score.

BIMA
BULLETIN OF INTELLIGENT MACHINES AND ALGORITHMS
Vol. 1 No. 3 March 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v1i3.16

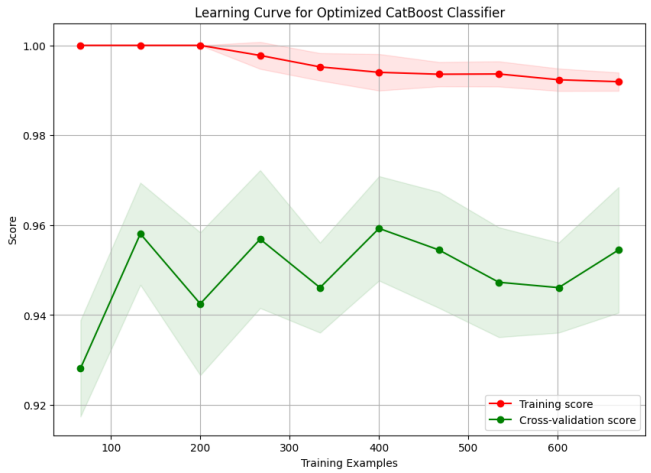

Fig.3. Learning curve for hyperparameter tuning catboost

Analysis of the learning curve in Fig. 3 illustrates the relationship between the amount of training data (training examples) and model performance, as measured by the training and cross-validation scores. In general, the training score is very high and relatively stable, approaching the maximum value as the amount of training data increases. This shows that the model consistently learns patterns from the training data. On the other hand, the cross-validation score is lower than the training score but remains within a relatively stable range as the amount of training data increases. With smaller data sizes, there are greater fluctuations in the validation score, as indicated by the wider shaded region. As the amount of training data increases, this variation tends to decrease, and the validation curve becomes more stable. The distance between the training and validation curves remains relatively constant, with no significant differences, indicating that the model does not exhibit underfitting or extreme overfitting. Overall, this curve pattern shows that the model's performance tends to remain stable as the amount of training data increases, although the improvement from adding data is not particularly dramatic.

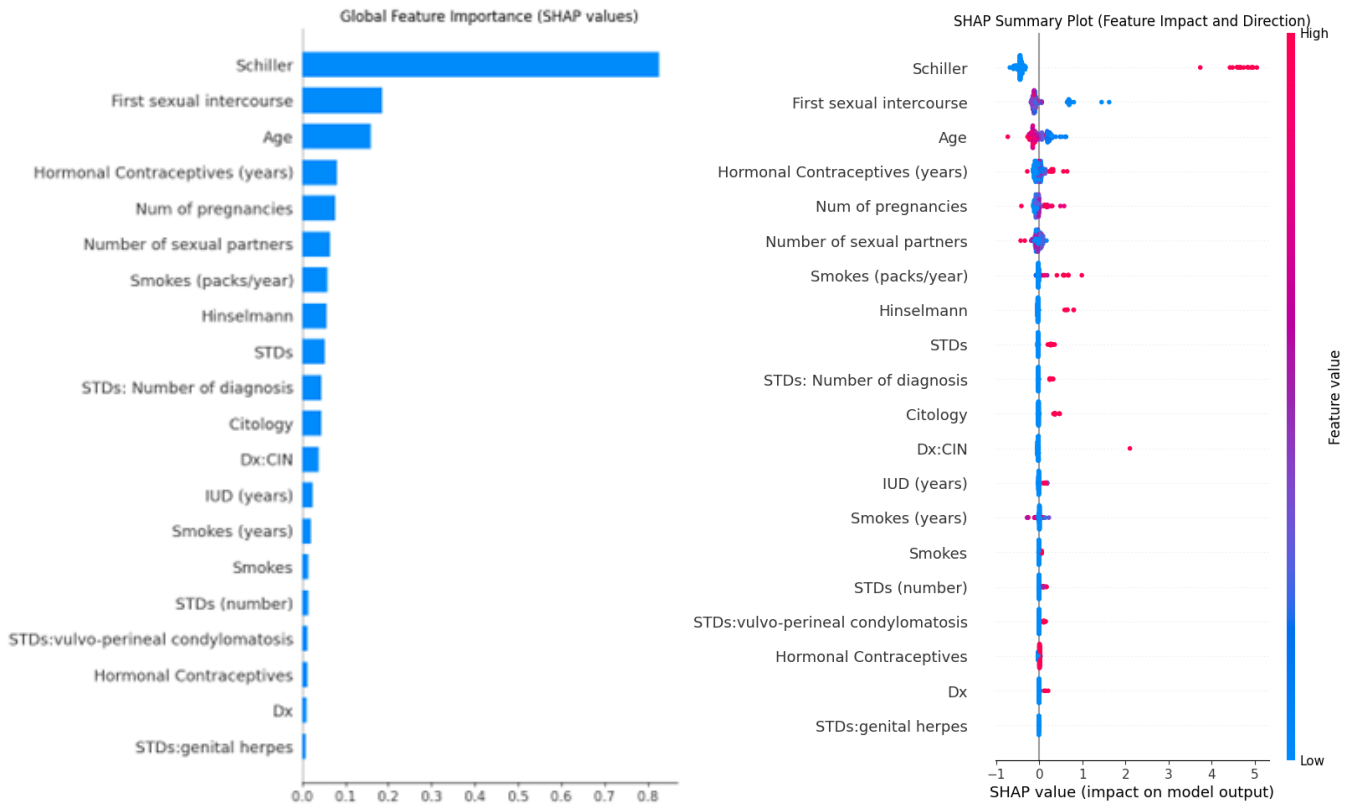

Fig.4. XAI global feature importance

Fig 4 shows that the Schiller test feature makes the dominant contribution to the model output, far exceeding that of the other features. This is in line with clinical practice, where Schiller test results are used as an important indicator in screening for precancerous cervical lesions. Features of age at first sexual intercourse, age, and duration of hormonal contraceptive use emerged as factors with a significant influence on risk prediction. These findings are consistent with epidemiological literature stating that early sexual exposure and hormonal factors play a role in increasing the risk of cervical cancer. Other behavioral and reproductive history features, such as the number of pregnancies, the number of sexual partners, and smoking habits, also show relevant contributions, even though their level of influence is lower. Interestingly, features related to sexually transmitted diseases (STDs)

Vol. 1 No. 3 March 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v1i3.16

and cytology test results had a smaller contribution in the global model. This can be interpreted as indicating that, in the context of the dataset used, clinical screening factors and behavioral characteristics play a more dominant role in distinguishing between high- and low-risk individuals.

Overall, the evaluation results show that all applied machine learning models achieve high accuracy in predicting cervical cancer risk. Still, there are significant differences in each algorithm's ability to detect clinically most crucial minority classes. The evaluation confirms that high-accuracy models do not always have adequate sensitivity to high-risk cases, as seen with KNN, which tends to be biased towards the majority class. Random Forest and XGBoost show more balanced performance improvements, but still have limitations in optimizing positive-class recall. Among all models, CatBoost displayed the most consistent performance, both before and after optimization, with a better balance between accuracy, sensitivity, and model stability. The application of hyperparameter tuning with GridSearchCV further improved CatBoost's performance and yielded a stable model, as evidenced by cross-validation evaluation and learning curve analysis. Furthermore, integrating the SHAP-based XAI approach strengthens the model's validity by demonstrating that predictions are grounded in clinically relevant factors, such as Schiller test results, age, and reproductive behavior history. These findings confirm that the proposed approach is not only computationally superior but also highly interpretable and medically relevant for supporting cervical cancer early detection systems.

4.2. Discussions

The evaluation results in this study show that the ensemble learning approach, particularly CatBoost optimized using GridSearchCV, achieves the best performance, with an accuracy of 97.01% and a good balance between precision and recall in the high-risk class. Compared to previous studies discussed in the Related Work section, this achievement shows a significant improvement in both classification performance and model stability.

Compared with previous studies, Tanimu et al. [13] reported that Random Forest achieved an accuracy of over 95% on the Cervical Cancer (Risk Factors) dataset. Although this value is comparable to the results in this study, that study focused on overall accuracy and did not include an in-depth analysis of evaluation metrics for minority classes. In this study, Random Forest showed good performance, achieving an accuracy of 96.41%. However, it still fell short of the optimized CatBoost, particularly in detecting positive cases of cervical cancer, as indicated by higher recall and F1-score values. This confirms that model evaluation in the medical domain cannot rely solely on accuracy but must also account for sensitivity to high-risk classes.

Research by Chadaga et al. [14] proposes a stacked ensemble learning approach that has been proven to improve prediction performance compared to single models. However, this approach requires a more complex model architecture and lacks an evaluation of model stability via a cross-validation scheme. In contrast to that study, this study shows that optimized CatBoost can achieve equivalent or better performance with more controlled complexity. This is reinforced by a k-fold cross-validation accuracy average of 95.45% with a low standard deviation, indicating the model's consistency and reliability across variations in the training data.

Regarding the handling of class imbalance, Kılıçarslan et al. [15] reported increased model sensitivity by applying the SMOTE technique. However, this approach may introduce synthetic bias due to the use of artificial data. In contrast to this approach, the results of this study show that CatBoost can handle class imbalance inherently without explicit oversampling techniques, as reflected in a positive class recall of 0.70 and a stable F1-score. This advantage supports the research objective of preserving the original data distribution, ensuring the model remains clinically relevant.

Okyay et al. [16] evaluated several machine learning algorithms, including Logistic Regression and Random Forest, achieving accuracies ranging from 93% to 96%. These results are in line with the performance of non-optimized models in this study, such as KNN and Random Forest. However, the study by Okyay et al. does not describe the hyperparameter tuning process. In this context, this study demonstrates methodological progress by showing that hyperparameter tuning can improve the performance of the CatBoost algorithm beyond the results reported in that study.

Meanwhile, Dong et al. [17] reported high performance of the SMART HPV model, developed from HPV genotype data and clinically validated. Although this approach excels at using molecular biomarkers with high precision, a direct comparison shows that this study achieved a competitive level of accuracy using only clinical and behavioral risk factor data, which are easier to obtain. Furthermore, this study enhances its practical value by applying explainable artificial intelligence (XAI) based on SHAP (Shapley Additive exPlanations), which enables transparent interpretation of each predictor variable's contribution to determining the risk level of cervical cancer. These advantages make the model not only computationally accurate but also clinically accountable, especially in healthcare settings with limited molecular laboratory facilities.

Overall, compared to previous studies, the evaluation results in this study not only show an increase in accuracy but also rely on a combination of adaptive ensemble learning, systematic hyperparameter optimization, and the integration of the XAI (SHAP) approach, which provides a deep understanding of the relationship between risk

Vol. 1 No. 3 March 2026
E-ISSN: 3132-5115
DOI:10.65780/bima.v1i3.16

factors and prediction results. This study provides a stronger, more transparent, and more relevant empirical contribution to the development of machine-learning-based cervical cancer risk prediction systems compared to previous studies, in terms of both model performance and clinical acceptability.

5. CONCLUSIONS

This study successfully achieved its primary objective of developing an accurate, stable, and interpretable machine-learning–based model for cervical cancer risk prediction to support early detection and clinical decision-making. The evaluation results demonstrate that although all applied machine learning algorithms achieved high overall accuracy, differences emerged in their ability to identify high-risk cases, which is the most critical requirement in medical screening applications. These findings confirm the premise stated in the introduction: accuracy alone is insufficient to assess model effectiveness in clinically sensitive contexts, and sensitivity-oriented metrics, such as recall and F1-score for the minority class, must be carefully considered.

Among the evaluated models, CatBoost consistently outperformed KNN, Random Forest, and XGBoost in terms of balanced performance. The optimized CatBoost model achieved the highest accuracy of 97.01%, with positive class precision of 0.78 and recall of 0.70, indicating a substantial improvement in detecting high-risk cervical cancer cases without sacrificing accuracy. The stability of the model was further confirmed through k-fold cross-validation, which yielded an average accuracy of 95.45% with a low standard deviation, and through learning curve analysis, which showed good convergence between training and validation performance. These results demonstrate that the proposed preprocessing and hyperparameter optimization strategy effectively minimizes overfitting while ensuring strong generalization capability.

Furthermore, the integration of explainable artificial intelligence using SHAP fulfilled the interpretability objective emphasized in the introduction. The XAI analysis revealed that model predictions were driven by clinically meaningful factors, such as Schiller test results, age at first sexual intercourse, age, and duration of hormonal contraceptive use, which are consistent with established medical and epidemiological evidence. This transparency enhances trust in the model and supports its potential adoption as a decision-support tool rather than a black-box classifier. In conclusion, by combining adaptive ensemble learning, systematic hyperparameter optimization, rigorous evaluation, and explainable AI, this study provides a predictive framework that is not only computationally accurate but also interpretable and clinically relevant. The proposed approach effectively addresses the limitations identified in previous studies. It aligns with the research goals outlined in the introduction and offers a practical, data-driven solution to strengthen early detection strategies for cervical cancer, particularly in healthcare settings with limited diagnostic resources.

REFERENCES

[1] H. Sung *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] World Health Organization, "Cervical Cancer," World Health Organization, Mar. 05, 2024. https://www.who.int/news-room/fact-sheets/detail/cervical-cancer (accessed Jan. 22, 2026).

[3] M. Arbyn *et al.*, "Accuracy and effectiveness of HPV mRNA testing in cervical cancer screening: A systematic review and meta-analysis," *The Lancet Oncology*, vol. 23, no. 7, pp. 950–960, 2022.

[4] T. E. Sangers *et al.*, "Towards successful implementation of artificial intelligence in skin cancer care: A qualitative study exploring the views of dermatologists and general practitioners," *Archives of Dermatological Research*, vol. 315, no. 5, pp. 1187–1195, 2023.

[5] B. He et al., "Prediction Models for Prognosis of Cervical Cancer: Systematic Review and Critical Appraisal," Frontiers in Public Health, vol. 9, May 2021, doi: https://doi.org/10.3389/fpubh.2021.654454.

[6] L. Akter *et al.*, "Prediction of cervical cancer from behavior risk using machine learning techniques," *SN Computer Science*, vol. 2, no. 3, Art. no. 177, 2021.

[7] J. Dunn and P. Balaprakash, *Data Science Applied to Sustainability Analysis*. Amsterdam, Netherlands: Elsevier, 2021.

[8] G. Kostopoulos, G. Davrazos, and S. Kotsiantis, "Explainable artificial intelligence-based decision support systems: A recent review," *Electronics*, vol. 13, no. 14, Art. no. 2842, 2024.

[9] M. M. Uddin *et al.*, "The role of machine learning in transforming healthcare: A systematic review," *Information Systems Research*, vol. 1, no. 1, 2024.

[10] N. A. Wani *et al.*, "Synergizing fusion modeling for accurate cardiac prediction through explainable artificial intelligence," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 1, pp. 1504–1512, 2024.

[11] P. E. Castle, "Looking back, moving forward: Challenges and opportunities for global cervical cancer prevention and control," *Viruses*, vol. 16, no. 9, Art. no. 1357, 2024.

[12] C. Yue *et al.*, "Machine learning in early screening for high-grade cervical intraepithelial neoplasia using blood testing," *BMC Medical Informatics and Decision Making*, 2025.

[13] Tanimu, Jesse Jeremiah, et al. "A machine learning method for classification of cervical cancer." *Electronics* 11.3 (2022): 463.

[14] Chadaga, Krishnaraj, et al. "Predicting cervical cancer biopsy results using demographic and epidemiological parameters: A custom stacked ensemble machine learning approach." *Cogent Engineering* 9.1 (2022): 2143040.

[15] Kılıçarslan, Serhat, Maruf Gögebakan, and Cemil Közkurt. "Cervical cancer prediction using SMOTE algorithm and machine learning approaches." *Journal of the Institute of Science and Technology* 13.2 (2023): 747-759.

[16] Okyay, Tugba Muhlise, Ibrahim Yilmaz, and Macit Koldas. "Evaluating Cervical Cancer Risk Using Machine Learning." *The Medical Bulletin of Haseki* (2025).

[17] Dong, Binhua, et al. "Development, validation, and clinical application of a machine learning model for risk stratification and management of cervical cancer screening based on full-genotyping hrHPV test (SMART-HPV): a modelling study." *The Lancet Regional Health–Western Pacific* 55 (2025).