Original Article

DOI: 10.4274/haseki.galenos.2025.35220
Med Bull Haseki 2025;63(4):188-194

Evaluating Cervical Cancer Risk Using Machine Learning

Tugba Muhlise Okyay*, Ibrahim Yilmaz**, Macit Koldas**

*University of Health Sciences Türkiye, Hamidiye Faculty of Medicine, Department of Medical Biochemistry, Istanbul, Türkiye
**University of Health Sciences Türkiye, Istanbul Haseki Training and Research Hospital, Clinic of Medical Biochemistry, Istanbul, Türkiye

Abstract

Aim: Cervical cancer development is influenced by a complex interaction of socio-demographic, behavioral, and clinical factors, which can be systematically analyzed using large datasets. Therefore, this study aimed to evaluate the effectiveness of machine learning (ML) models applied to the University of California, Irvine (UCI), cervical cancer risk factors dataset in predicting cervical health outcomes and supporting early detection strategies.

Methods: This study was designed as a retrospective data analysis covering a random sampling of patients between 2012 and 2013 who attended the gynecology service at Hospital Universitario de Caracas in Caracas, Venezuela. The publicly available UCI cervical cancer risk factors dataset was utilized for the analysis. A correlation heatmap was generated to explore the relationships among various risk factors. To address the class imbalance present in the dataset, the synthetic minority over-sampling technique (SMOTE) was applied. Subsequently, different ML classifiers were trained and evaluated to predict cervical cancer outcomes with improved accuracy.

Results: The correlation analysis revealed strong correlations among smoking-related measures and diagnostic variables, indicating internal consistency. After applying SMOTE, the dataset achieved a balanced distribution of healthy and diseased individuals. The ensemble classifiers demonstrated high accuracy, up to 97%, and precision, with random forest and light gradient boosting machine performing particularly well. However, the recall for cancer detection was lower: 0.80, indicating potential missed diagnoses.

Conclusion: The findings support the integration of ML in clinical diagnostics for cervical cancer, highlighting its potential for improving early detection and patient outcomes while also emphasizing the need for ongoing refinement in model performance.

Keywords: Cervical cancer, risk factors, machine learning, gynecology, diagnosis

Introduction

Cervical cancer remains one of the most significant global health concerns among women, especially in developing regions (1). Cervical cancer accounts for approximately 6.5% of all malignancies in women. Despite advances in screening and vaccination programs, the high incidence and mortality rates highlight the urgent need for improved strategies in prevention and early detection (2,3). Infection with human papillomavirus (HPV), primarily transmitted through sexual contact, is the leading cause of cervical cancer, and vaccination against HPV has become an essential preventive measure supported by global health authorities (4-6).

Early detection is critical to reducing mortality, yet asymptomatic progression in the early stages makes timely diagnosis (Dx) challenging (7). Traditional screening methods such as Pap smears and HPV tests, while effective, may be limited in sensitivity, accessibility, or cost in certain healthcare settings. For example, Pap smears may yield false-negative results in up to 50% of cases, leading to delayed diagnosis (8). In addition, in many low- and middle-income countries, limited access to trained

Corresponding Author: Macit Koldas, MD, University of Health Sciences Türkiye, Istanbul Haseki Training and Research Hospital, Clinic of Medical Biochemistry, Istanbul, Türkiye
E-mail: macitkoldas@gmail.com **ORCID:** orcid.org/0000-0001-8967-2708
Received: 11.06.2025 **Accepted:** 04.09.2025 **Publication Date:** 02.10.2025

Cite this article as: Okyay TM, Yilmaz I, Koldas M. Evaluating cervical cancer risk using machine learning. Med Bull Haseki. 2025;63(4):188-194

personnel and laboratory infrastructure further reduces the effectiveness of routine screening programs (9). Advances in artificial intelligence and machine learning (ML) offer opportunities to improve prediction and stratification of high-risk individuals (10).

We hypothesized that the integration of socio-demographic, behavioral, and medical data into ML based models would enhance the accuracy of cervical cancer risk prediction compared to traditional screening methods alone. Therefore, the aim of this study was to evaluate multiple ML algorithms on the University of California, Irvine (UCI) cervical cancer risk factors dataset (11), addressing class imbalance through the use of the synthetic minority over-sampling technique (SMOTE) (12). This approach is expected to contribute to clinical practice by supporting earlier identification of high-risk patients, thereby enabling timely interventions and ultimately reducing cervical cancer-related morbidity and mortality.

Materials and Methods

Dataset Description

Ethics committee approval was not required for this study, as the data used does not contain personally identifiable information. Therefore, ethics committee approval was not obtained. The dataset used in this study was obtained from a publicly available cervical cancer screening database, containing clinical and behavioral attributes of female patients. The dataset, sourced from the UCI ML repository, includes 858 instances with 32 attributes capturing demographic, sexual, and clinical risk factors.

- Age
- Number of sexual partners
- First sexual intercourse age
- Number of pregnancies
- Smoking status and history
- Smokes (packs/year)
- Sexually transmitted diseases (STDs) and HPV infection status
- Diagnostic test results (Hinselmann, Schiller, cytology, biopsy)

The target variable was a binary classification indicating the presence or absence of cervical cancer or precancerous conditions such as cervical intraepithelial neoplasia (CIN) and HPV-positive status. Missing values were handled by imputing missing values. Categorical variables were encoded using one-hot encoding (13).

A snapshot of the dataset used in this study is presented in Figure 1.

Handling Imbalanced Data

Given the rarity of positive biopsy cases in the dataset, the class distribution was heavily skewed. We applied SMOTE to synthetically balance the classes and ensure fair model evaluation. synthetic minority over-sampling technique works by generating synthetic samples of the minority class rather than simply duplicating existing instances. It achieves this by selecting a minority class sample and interpolating it with one of its k-nearest neighbors in the feature space (14). This process introduces new, plausible samples and helps the model learn a more generalized decision boundary, thereby reducing the bias toward the majority class and improving the classifier's ability to detect minority class instances. synthetic minority over-sampling technique is particularly beneficial when used prior to training classification models, as it provides a balanced dataset without introducing exact duplicates, which could otherwise lead to overfitting (15).

Exploratory Data Analysis

Exploratory data analysis (EDA) was conducted using Python's data visualization libraries seaborn and matplotlib to visualize the distributions of patient groups or characteristics, including those pertaining to patients (16). Histograms were generated to assess the normality of the distribution of variables in the dataset. In this study, we conducted an EDA to investigate the relationships between cervical cancer Dx and HPV status using a heatmap visualization. This approach allows us to visualize the interactions and correlations between these critical variables.

Age	Num.	Fir.	Num.	Smokes	Smo.	Smo.	STD.	Dx:Cance	Dx:CIN	Dx:HPV	Dx	Hinselmar	Schiller	Citology	Biopsy
18	4.0	15.0	1.0	0.0	0.0	0.0	0	0	0	0	0	0	0	0	0
15	1.0	14.0	1.0	0.0	0.0	0.0	0	0	0	0	0	0	0	0	0
34	1.0	?	1.0	0.0	0.0	0.0	0	0	0	0	0	0	0	0	0
52	5.0	16.0	4.0	1.0	37.0	37.0	0	1	0	1	0	0	0	0	0
46	3.0	21.0	4.0	0.0	0.0	0.0	0	0	0	0	0	0	0	0	0

Age → Age, Num. → Number of sexual partners, Fir. → First sexual intercourse, Num. → Num of pregnancies, Smokes → Smokes, Smo. → Smokes (years), Smo. → Smokes (packs/year), STD. → STDs: Number of diagnosis, Dx:Cancer → Dx:Cancer, Dx:CIN → Dx:CIN, Dx:HPV → Dx:HPV, Dx → Dx, Hinselmann → Hinselmann, Schiller → Schiller, Citology → Citology, Biopsy → Biopsy

Figure 1. A snapshot of the dataset for this study, filtered data with first 5 rows

Machine Learning Model Training and Evaluation

For predictive modeling, the dataset was split into training and testing subsets using an 80/20 split ratio. The LazyPredict library was employed to facilitate the comparison of multiple regression models, including multi-layer perceptron regressor, extreme gradient boosting (XGBoost), elastic net with cross-validation, and Lasso, using Python and scikit-learn (17).

Software and Tools

All analyses were conducted using Python 3.10 within a Jupyter Notebook environment. The primary libraries utilized were pandas and NumPy for data manipulation, as well as matplotlib (18) and seaborn for visualization. For model development and evaluation, scikit-learn, XGBoost, and Light Gradient Boosting Machine (LightGBM) were employed, providing robust tools for building and assessing ML models (19).

Statistical Analysis

Model performance metrics were calculated to comprehensively assess the predictive ability and efficiency of each algorithm. Each model was evaluated in terms of accuracy, which reflects the overall correctness of predictions; sensitivity (recall), which measures the ability to correctly identify positive cases; specificity, which quantifies the correct identification of negative cases; precision, which reflects the proportion of true positives among predicted positives; F1-score, which balances precision and recall; and receiver operating characteristic (ROC) area under curve (AUC), which provides a global measure of model discrimination capability (20). In addition, the computational efficiency of each model was assessed by recording the time taken (in seconds) to complete both training and prediction phases. This was particularly important in evaluating the scalability of the models for real-world applications where rapid decision-making may be required.

Results

Exploratory Data Analysis

A correlation heatmap (Figure 2) illustrated both weak and strong relationships among variables, emphasizing the multifactorial nature of cervical cancer risk.

- Smokes, smokes (years), and smokes (packs/year) show strong correlations ($r \approx 0.69$-0.72), indicating internal consistency across these smoking-related measures.
- Diagnostic variables (Dx: Cancer, Dx: HPV, Dx: CIN, and overall Dx) are also strongly correlated ($r > 0.68$), reflecting overlapping diagnostic criteria or comorbidity.
- Visual inspection outcomes (Hinselmann, Schiller, cytology) are moderately to strongly correlated with biopsy, with Schiller showing the strongest correlation ($r = 0.74$), suggesting predictive value for biopsy-confirmed cases.
- Age correlates moderately with number of pregnancies ($r = 0.56$) and first sexual intercourse ($r = 0.37$), consistent with expected life course patterns. Several variables, including STDs, number of diagnoses, and number of sexual partners, display very weak correlations with most other variables ($r < 0.1$), implying limited linear association in this sample.

The distribution indicates that HPV diagnosis (Dx: HPV) is the most frequently observed condition, slightly surpassing cancer diagnoses (Dx: Cancer), while CIN (Dx: CIN) is relatively less common (Figure 3).

A significant class imbalance is evident. The number of healthy individuals greatly outnumbers diseased individuals, indicating a highly skewed dataset. This imbalance in class frequencies may necessitate data augmentation techniques to mitigate class bias and enhance model learning. After applying SMOTE, a balancing technique, the classes are now evenly distributed, with almost equal numbers of healthy and diseased individuals (Figure 4).

Machine Learning on SMOTE Balanced Data

Table 1 below shows the performance of various classification models used to predict cervical cancer based on raw data. The highest accuracy and ROC-AUC scores (97%) were achieved by both the RandomForestClassifier and LGBMClassifier. RandomForestClassifier delivered this high performance with a training time of just 0.74 seconds, while LGBMClassifier achieved similar results even faster (0.33 seconds), making both models efficient and effective. Other strong performers include XGBClassifier, DecisionTreeClassifier, BaggingClassifier, and ExtraTreesClassifier, each achieving 96% accuracy and ROC-AUC, with DecisionTreeClassifier standing out for its extremely low training time (0.02 seconds).

Model Performance of Categorized Data

The ML model demonstrated strong classification performance in predicting cervical health outcomes across four categories: Healthy, Cancer, CIN, and HPV infection. The overall accuracy achieved was 93%, with particularly high precision in detecting cancer (1.00) and high recall for HPV (0.93). However, the recall for the cancer class was slightly lower (0.80), suggesting an increase in false negatives, while CIN prediction maintained balanced precision and recall (Table 2).

A precision of 1.00 for the cancer class indicates no false positives. While this is desirable, the recall is only 0.80, implying that 20% of true cancer cases were missing.

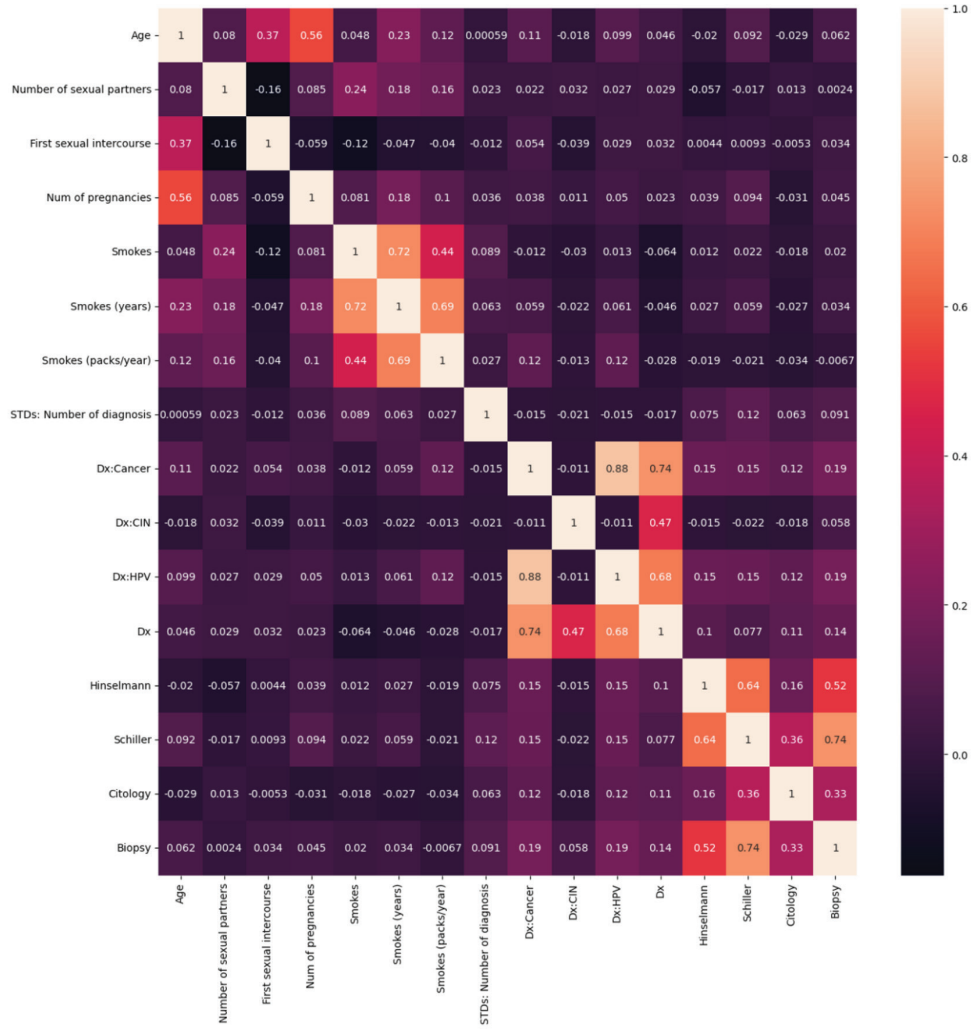

Figure 2. Multivariate correlation structure of risk factors associated with cervical cancer diagnosis

Dx: Diagnosis, CIN: Cervical intraepithelial neoplasia, HPV: Human papillomavirus, STDs: Sexually transmitted diseases

In clinical diagnostics, false negatives for cancer can have severe implications, leading to delayed Dx and treatment. For HPV, the recall of 0.93 is excellent; however, a precision of 0.78 means there's a relatively higher false positive rate.

Discussion

This study provides a detailed evaluation of ML approaches for predicting cervical cancer risk based on clinical and lifestyle data. Our analysis revealed strong internal consistency among smoking-related variables, confirming their collective importance as predictive features. This finding emphasizes that integrating multiple related behavioral factors can enhance the discriminatory power of ML models, supporting targeted risk assessment strategies.

Interestingly, reproductive variables such as high parity and early age at first sexual intercourse did not show a direct association with cervical cancer in our dataset, despite being reported as risk factors in previous epidemiological studies (21,22). There are several possible explanations for the lack of observed association between reproductive factors and cervical cancer risk in our analysis. First, the relationship may not be detectable in smaller or demographically homogeneous samples. Second, unmeasured confounders such as HPV infection, socio-economic status, smoking habits, contraceptive use, or access to cervical cancer screening may influence or obscure the true relationship between reproductive behaviors and cancer risk. Lastly, issues related to data quality, such as self-reported information, missing values,

or inaccuracies in key variables like age at first intercourse, could contribute to the attenuation of expected associations. This discrepancy suggests that contextual variables such as HPV status, socio-economic conditions, screening access, and smoking habits may modulate the impact of reproductive behaviors. It also highlights the importance of considering dataset-specific characteristics and potential confounders when applying ML models to real-world clinical data.

The evaluation of different ML classifiers demonstrated that ensemble methods, particularly RandomForestClassifier and LightGBMClassifier, outperformed simpler probabilistic models. The superior accuracy and ROC-AUC of these methods indicate that capturing complex, non-linear relationships between clinical features is crucial for reliable prediction. While the high precision for cancer detection minimizes false positives, the relatively lower recall underscores the need for caution in clinical interpretation, as some true cases may still be missed. In contrast, models such as DecisionTreeClassifier and ExtraTreeClassifier provide a balance between performance and computational efficiency, suggesting potential utility for rapid or mobile-based screening tools. These results are in line with existing literature that highlights the efficacy of ML in improving early detection of cervical cancer and HPV-related abnormalities (23).

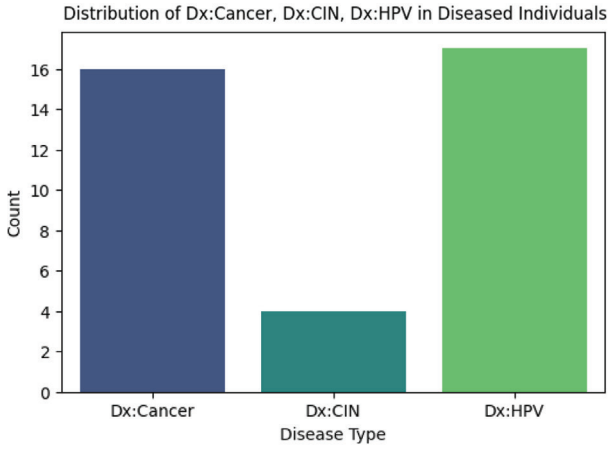

Figure 3. Distribution of diagnoses: Cancer, CIN and HPV among diseased individuals

Dx: Diagnosis, CIN: Cervical intraepithelial neoplasia, HPV: Human papillomavirus

Furthermore, the application of SMOTE to address class imbalance proved essential for ensuring adequate representation of minority cases. Our findings reinforce that data preprocessing techniques directly impact model reliability and generalizability, particularly in medical datasets, where diseased cases are often underrepresented (24). This supports the broader integration of ML pipelines into digital health solutions, potentially improving early detection in resource-limited settings and complementing existing clinical workflows.

The high precision in the cancer class (1.00) indicates that the model effectively minimizes false positives, which is crucial in avoiding unnecessary psychological and medical interventions. Conversely, the lower recall for cancer (0.80) implies a potential risk of missed cancer diagnoses, which is critical in a clinical setting. The model's strong performance in HPV prediction is also consistent with research indicating that behavioral and screening features are highly predictive of HPV status. According to Schiffman et al. (25), the integration of HPV typing into

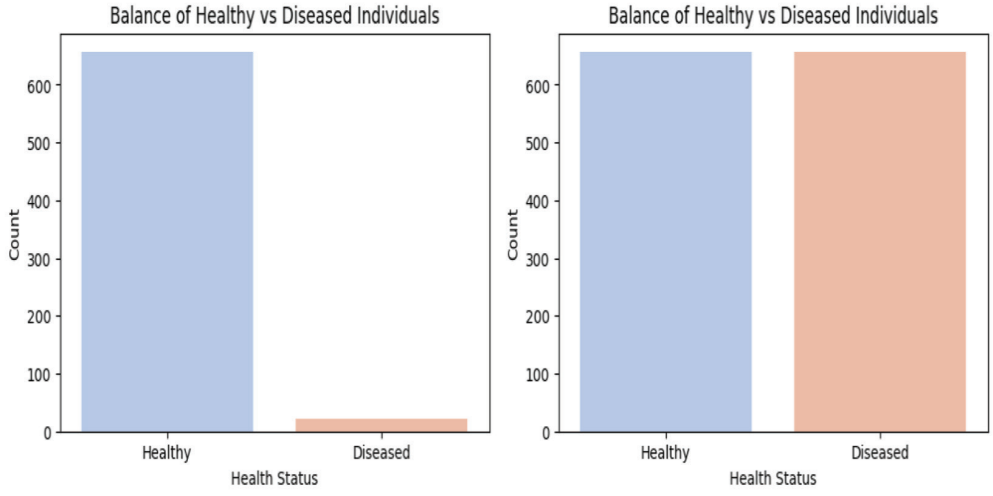

Figure 4. Distribution of Healthy vs. Diseased individuals in the dataset before SMOTE (left) after SMOTE(right)

SMOTE: Synthetic minority over-sampling technique

Table 1. Comparative analysis of classifier efficacy in cervical cancer prediction

Model	Accuracy	ROC-AUC	Time taken
XGBClassifier	0.96	0.96	1.06
RandomForestClassifier	0.97	0.97	0.74
DecisionTreeClassifier	0.96	0.96	0.02
BaggingClassifier	0.96	0.96	0.06
ExtraTreesClassifier	0.96	0.96	0.18
LightGBMClassifier	0.97	0.97	0.33
LabelPropagation	0.95	0.95	0.23
LabelSpreading	0.95	0.95	0.36
ExtraTreeClassifier	0.95	0.95	0.02
AdaBoostClassifier	0.87	0.87	0.17
K-NeighborsClassifier	0.91	0.91	0.06
BernoulliNB	0.86	0.87	0.02
GaussianNB	0.69	0.75	0.02

LightGBM: Light Gradient Boosting Machine, XGBClassifier: Extreme gradient boosting, BernoulliNB: Bernoulli Naïve Bayes

Table 2. Classification performance metrics of the cervical cancer prediction model

	Precision	Recall	f1-score
Healthy	0.96	0.95	0.94
Cancer	1.00	0.80	0.84
CIN	0.88	0.84	0.81
HPV	0.78	0.93	0.94

CIN: Cervical intraepithelial neoplasia, HPV: Human papillomavirus

screening strategies significantly enhances early detection of precancerous changes.

Finally, the balance between CIN classification metrics (precision: 0.88, recall: 0.84) indicates that the model can reasonably detect intermediate lesion stages, which are crucial for preventive interventions before progression to invasive cancer.

Overall, the study highlights that a careful combination of data preprocessing, feature selection, and ensemble learning can produce predictive models with both high accuracy and practical applicability. These results contribute to ongoing efforts to optimize automated screening tools and provide clinicians with evidence-based decision support in cervical cancer prevention and management.

Study Limitations

The potentially limited and non-diverse sample size may affect the generalizability of the findings. Data quality issues, such as missing information, could introduce bias, and additionally, important risk factors may have been overlooked in feature selection. Additionally, the complexity of the models may hinder interpretability, and the lack of external validation on independent datasets limits the applicability of the results in real-world settings. Despite these limitations, our findings indicate the integration of ML into clinical diagnostics to enhance early detection and treatment of cervical cancer, while recognizing the need for further research to address these limitations.

Conclusion

The study highlights the multifactorial nature of cervical cancer risk, revealing significant correlations among variables through a heatmap analysis. While addressing class imbalance with SMOTE improved model performance, particularly with ensemble classifiers like random forest and LightGBM, the lower recall for cancer detection emphasizes the need for further investigation to avoid missing true cases.

Ethics

Ethics Committee Approval: Ethics committee approval was not required for this study, as the data used does not contain public, anonymous, or personally identifiable information. Therefore, ethics committee approval was not obtained.

Informed Consent: This study was designed as a retrospective study.

Footnotes

Authorship Contributions

Concept: M.K., Design: T.M.O., M.K., Data Collection or Processing: T.M.O., M.K., Analysis or Interpretation: I.Y., M.K., Literature Search: T.M.O., M.K., Writing: T.M.O., I.Y., M.K.

Conflict of Interest: No conflicts of interest were declared by the authors.

Financial Disclosure: This study received no financial support.

References

1. Zhou L, Li Y, Wang H, Qin R, Han Z, Li R. Global cervical cancer elimination: quantifying the status, progress, and gaps. BMC Med. 2025;23:67.
2. Reza S, Anjum R, Khandoker RZ, Khan SR, Islam MR, Dewan SMR. Public health concern-driven insights and response of low- and middle-income nations to the World Health Organization call for cervical cancer risk eradication. Gynecol Oncol Rep. 2024;54:101460.
3. Cervical cancer. Avaible from: https://www.who.int/health-topics/cervical-cancer (accessed 17 Apr2025).
4. Alemany L, Felsher M, Giuliano AR, et al. Oral human papillomavirus (HPV) prevalence and genotyping among healthy adult populations in the United States and Europe: results from the PROGRESS (PRevalence of Oral hpv infection, a Global aSSessment) study. EClinicalMedicine. 2025;79:103018.

5. Chidebe RCW, Osayi A, Torode JS. The global fund, cervical cancer, and HPV infections: what can low- and middle-income countries do to accelerate progress by 2030? EClinicalMedicine. 2025;81:103127.
6. Hamid MKI, Hasneen S, Lima AK, Shawon SR, Shahriar M, Anjum R. Cervical cancer trends, HPV vaccine utilization, and screening in low- and lower-middle-income countries: an updated review. Ther Adv Vaccines Immunother. 2025;13:25151355251356646.
7. Hong MK, Ding DC. Early diagnosis of ovarian cancer: a comprehensive review of the advances, challenges, and future directions. Diagnostics (Basel). 2025;15:406.
8. Alhudhud M, Maqsood S, Hussein ME, et al. Cervical cancer screening: a comparative study of TruScreen vs. Pap Smear. BMC Womens Health. 2025;25:198.
9. Fashedemi O, Ozoemena OC, Peteni S, et al. Advances in human papillomavirus detection for cervical cancer screening and diagnosis: challenges of conventional methods and opportunities for emergent tools. Anal Methods. 2025;17:1428-50.
10. Kartika Ririe A, Khan M, Abbas Bangash S, Mahmoud Younes Elsherbiny RM, Umer Ali Ayub M, Louis E, et al. Advancements in artificial intelligence and machine learning for early cardiovascular risk prediction and diagnosis. J Neonatal Surg. 2025;14:6170-80.
11. Fernandes K, Cardoso J, Fernandes J. Cervical Cancer (Risk Factors) [Dataset]. UCI Machine Learning Repository; 2017. Available from: https://doi.org/10.24432/C5Z310
12. Wang AX, Le VT, Trung HN, Nguyen BP. Addressing imbalance in health data: synthetic minority oversampling using deep learning. Comput Biol Med. 2025;188:109830.
13. Poslavskaya E, Korolev A. Encoding categorical data: is there yet anything 'hotter' than one-hot encoding? 2023.
14. Salehi AR, Khedmati M. A cluster-based SMOTE both-sampling (CSBBoost) ensemble algorithm for classifying imbalanced data. Sci Rep. 2024;14:5152.
15. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002;16:321-57.
16. Roth JP, Bajorath J. Machine learning models with distinct Shapley value explanations decouple feature attribution and interpretation for chemical compound predictions. Cell Rep Phys Sci. 2024;5:102110.
17. GitHub - shankarpandala/lazypredict: Lazy Predict help build a lot of basic models without much code and helps understand which models works better without any parameter tuning. https://github.com/shankarpandala/lazypredict (accessed 30 June2025).
18. Nelli F. Python data analytics: data analysis and science using pandas, matplotlib and the python programming language. Apress, 2015.
19. Huang JC, Tsai YC, Wu PY, et al. Predictive modeling of blood pressure during hemodialysis: a comparison of linear model, random forest, support vector regression, XGBoost, LASSO regression and ensemble method. Comput Methods Programs Biomed. 2020;195:105536.
20. Cabot JH, Ross EG. Evaluating prediction model performance. Surgery. 2023;174:723-6.
21. International Collaboration of Epidemiological Studies of Cervical Cancer. Cervical carcinoma and reproductive factors: collaborative reanalysis of individual data on 16,563 women with cervical carcinoma and 33,542 women without cervical carcinoma from 25 epidemiological studies. Int J Cancer. 2006;119:1108-24.
22. Plummer M, Peto J, Franceschi S; International Collaboration of Epidemiological Studies of Cervical Cancer. Time since first sexual intercourse and the risk of cervical cancer. Int J Cancer. 2012;130:2638-44.
23. Wong OGW, Ng IFY, Tsun OKL, Pang HH, Ip PPC, Cheung ANY. Machine learning interpretation of extended human papillomavirus genotyping by onclarity in an Asian cervical cancer screening population. J Clin Microbiol. 2019;57:e00997-19.
24. Zhu J, Pu S, He J, et al. Processing imbalanced medical data at the data level with assisted-reproduction data as an example. BioData Min. 2024;17:29.
25. Schiffman M, Castle PE, Jeronimo J, Rodriguez AC, Wacholder S. Human papillomavirus and cervical cancer. Lancet. 2007;370:890-907.