Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 9 No. 3 (2025) 535 - 541 e-ISSN: 2580-0760

# XGBoost Algorithm for Cervical Cancer Risk Prediction: Multi-dimensional Feature Analysis

Sudi Suryadi[1]*, Masrizal[2]
[1,2]Information System, Faculty of Science and Technology, Universitas Labuhanbatu, Rantauprapat, Indonesia
[1]sudisuryadi28@gmail.com, [2]masrizal120405@gmail.com

**Abstract**

*Cervical cancer continues to pose a significant global health challenge, with early detection remaining the cornerstone for effective intervention. This study is situated at the intersection of clinical oncology and computational intelligence, exploring the potential of gradient-boosting algorithms to overcome the limitations of conventional screening methodologies. An XGBoost model was developed to predict cervical cancer risk. This model incorporates demographic, behavioral, and clinical parameters. The model was developed using data from 858 patients at the Hospital Universitario de Caracas. The preprocessing pipeline was designed to address the complexities inherent in medical data, including strategic management of missing values and standardizing heterogeneous features. The model demonstrated an overall accuracy of 96.3%, with a sensitivity of 66.7% and a specificity of 97.6%. This performance profile indicates adept navigation of the delicate balance between missed diagnoses and unnecessary interventions. Feature importance analysis revealed a multifaceted risk landscape, where screening test results contributed substantial predictive power (approximately 60%), complemented by demographic and behavioral factors, including age, reproductive history, and contraceptive usage patterns. The confusion matrix analysis revealed the clinical implications of the model predictions, demonstrating a promising positive predictive value of 55.0% despite the pronounced class imbalance. These findings suggest that ensemble learning approaches can effectively synthesize diverse patient data into meaningful risk assessments, potentially enhancing screening efficiency through personalized stratification. Future research directions include prospective validation across diverse populations, integration of longitudinal data, and further exploration of explainable AI techniques to bridge the gap between algorithmic predictions and clinical implementation.*

**Keywords**: cervical cancer screening; computational oncology; machine learning; risk stratification; XGBoost

*How to Cite:* Sudi Suryadi and Masrizal, "XGBoost Algorithm for Cervical Cancer Risk Prediction: Multi-dimensional Feature Analysis", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 9, no. 3, pp. 535 - 541, Jun. 2025.
*Permalink*/DOI: https://doi.org/10.29207/resti.v9i3.6587

*Received*: April 26, 2025
*Accepted*: June 11, 2025
*Available Online*: June 21, 2025

*This is an open-access article under the CC BY 4.0 License*
*Published by Ikatan Ahli Informatika Indonesia*

## 1. Introduction

The persistent burden of cervical cancer stands as a testament to both the remarkable progress and the enduring challenges in contemporary oncology [1]. Despite significant advances in preventive medicine and screening technologies, this malignancy continues to claim over 300,000 lives annually, with a disproportionate impact in resource-limited settings [2]. The narrative of cervical cancer—its pathogenesis intrinsically linked to human papillomavirus infection [3], its protracted natural history offering ample opportunity for intervention, and its potentially devastating consequences when detected late-presents both a public health imperative and a compelling analytical challenge [4], [5].

The evolution of cervical cancer screening has traversed a complex trajectory from the revolutionary Papanicolaou test introduced in the mid-20th century to contemporary molecular testing for high-risk HPV genotypes [6], [7]. However, despite this technological progression, the fundamental challenge persists: identifying those women most at risk within a predominantly healthy population while minimizing both the psychological burden of false alarms and the devastating consequences of missed diagnoses [8]. This screening paradox has prompted increasing interest in computational approaches that transcend the limitations of individual biomarkers or clinical heuristics [9].

The convergence of several transformative trends has created fertile ground for novel cervical cancer risk stratification approaches [10]. The exponential growth

in computing capacity has enabled the application of increasingly sophisticated algorithms on complex medical datasets [11]. The maturation of machine learning methodologies has shown remarkable success in pattern recognition tasks across various domains [10]-[13]. Multidimensional patient datasets spanning demographic, behavioral, and clinical domains offer unprecedented opportunities for risk modeling [14], [15].

Ensemble learning methods-particularly gradient boosting frameworks-have emerged as promising for medical risk prediction tasks [16], [17]. This seminal work introducing the XGBoost algorithm significantly advanced this domain, offering superior predictive performance and computational efficiency [18]. Subsequent applications in oncology have demonstrated the algorithm's capacity to synthesize diverse clinical markers into unified risk assessments that frequently surpass traditional statistical approaches [19], [20].

Previous studies exploring machine learning applications in cervical cancer have demonstrated encouraging results. The study implemented support vector machines to classify cervical cytology images, achieving 86% accuracy in distinguishing precancerous from normal cells [21]. Other researchers applied random forests to clinical data to predict high-grade cervical lesions, reporting an AUC of 0.82 [22]. However, these approaches have typically focused on single data modalities or limited feature sets, potentially missing the integrative potential of comprehensive patient profiling.

Our research addresses these limitations through a comprehensive application of XGBoost to predict cervical cancer risk using a multidimensional dataset from Hospital Universitario de Caracas in Venezuela [23]. By integrating demographic information, reproductive history, behavioral risk factors, and screening test results, we aim to develop a predictive framework that captures the multifaceted nature of cervical carcinogenesis while providing clinically interpretable risk assessments. This approach aligns with the emerging paradigm of risk-based cancer screening, wherein personalized risk stratification replaces age-based or one-size-fits-all screening protocols [24], [25].

Beyond its immediate application to cervical cancer prediction, this research contributes to the broader discourse on integrating machine learning methodologies into clinical practice. By developing an interpretable, accessible prediction framework for a cancer that disproportionately affects disadvantaged populations, this work advances both the technical and ethical dimensions of machine learning applications in global health.

## 2. Methods

Our research employed a supervised machine learning approach to develop a predictive cervical cancer risk assessment model based on a cross-sectional dataset. The dataset, sourced from Hospital Universitario de Caracas in Venezuela via Kaggle's public repository, comprised records from 858 patients who underwent cervical cancer screening between 2017-2018 [26]. The data collection protocol included demographic information, behavioral risk factors, medical history, and diagnostic test results—creating a comprehensive portrait of each patient's risk profile spanning multiple domains of potential carcinogenic influence.

The ethical framework governing this research prioritized patient privacy; all records were de-identified prior to analysis by standard medical research protocols. Each patient profile encompassed 36 distinct features, including continuous variables (e.g., age, years of contraceptive use), binary indicators (e.g., STD status, smoking status), and categorical outcomes (diagnostic test results). The target variable—biopsy result—represented the definitive diagnostic outcome indicating cervical cancer presence (positive=1) or absence (negative=0).

### *2.1 Data Preprocessing and Cleaning*

Transitioning from raw data to a modeling substrate involved multiple transformational stages. Each stage addressed specific quality concerns while preserving the underlying signal, essential for predictive insight. As illustrated in Figure 1, this comprehensive preprocessing workflow was the foundation for our analytical approach.

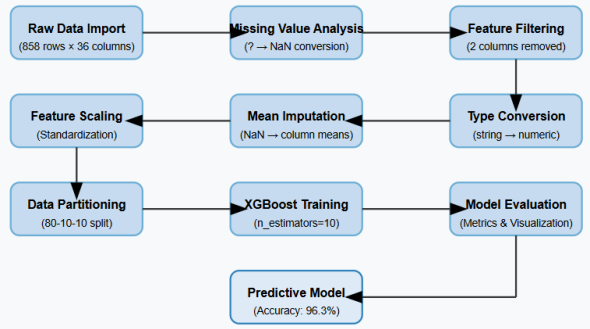


Figure 1. Data Preprocessing Workflow

Our initial encounter with the dataset revealed several quality challenges that required systematic remediation. The data presented inconsistent formatting, with numeric values encoded as strings, missing values represented by question marks rather than standardized null indicators, and specific columns exhibiting excessive missingness. The transformation pathway began with replacing question mark placeholders with appropriate NaN values, facilitating subsequent statistical analysis of missingness patterns.

Analysis of the missing value distribution revealed two temporal variables—"STDs: Time since first diagnosis"

and "STDs: Time since last diagnosis"—with nullity exceeding 80%. Given this profound level of missingness, imputation would have introduced more statistical noise than signal; hence, these columns were removed from subsequent analyses. Following this initial triage, we confronted the data type inconsistency challenge. A comprehensive inspection revealed that numerically continuous variables were encoded as strings, necessitating conversion to appropriate numeric formats for algorithmic processing.

Figure 2 shows that transformation yielded a dataset with 24 variables in float64 format and 10 in int64 format, establishing the mathematical precision necessary for subsequent modeling steps. We then addressed the remaining missing values with harmonized data types through mean imputation. This strategy preserves the central tendency of each feature while providing complete data for the XGBoost algorithm.

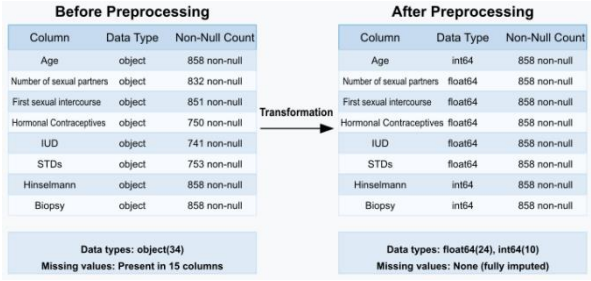


Figure 2. Data Type Conversion and Completion

Following data cleaning, we conducted comprehensive exploratory analysis to uncover distributional characteristics and relationships within the feature space as shown in Figure 3. This investigative phase revealed critical patterns that informed subsequent modeling decisions while providing contextual understanding of the patient population represented in our dataset.

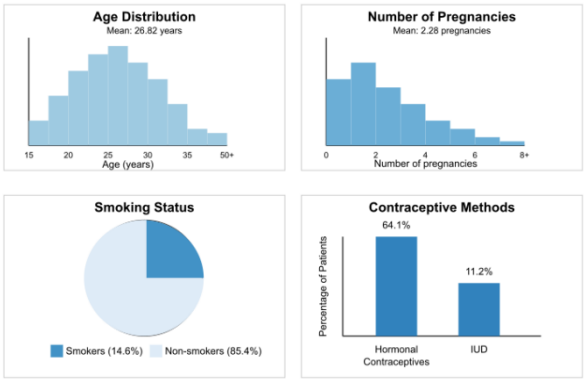


Figure 3. Key Feature Distribution

Descriptive statistics illuminated the demographic and behavioral profile of our cohort: a relatively young female population (mean age 26.82 years) with moderate reproductive history (average 2.28 pregnancies) and limited tobacco exposure (14.6% smoking prevalence). Hormonal contraceptive usage emerged as the predominant birth control method (64.1%) compared to IUD utilization (11.2%). STD history was reported by 10.4% of patients, with HPV noted explicitly in only 0.3% of cases—a finding that underscores potential under-detection of this oncogenic virus in the study population.

### *2.2 Model Development*

Following data preparation and exploratory analysis, we proceeded with feature engineering to optimize the input vector for XGBoost training. While our initial approach preserved all 33 available predictor variables to maximize information availability to the algorithm, we conducted sensitivity analyses to evaluate potential dimensionality reduction strategies.

The feature set included demographic characteristics (Age), reproductive history (Number of pregnancies, First sexual intercourse), behavioral risk factors (Smoking status, years, and intensity), contraceptive methods (Hormonal contraceptives, IUD, and duration of use), detailed STD history (across multiple pathogen categories), and screening test results (Hinselmann, Schiller, Cytology). This comprehensive approach allowed the model to identify complex interaction patterns that might elude traditional statistical analyses.

To enhance algorithm convergence and performance, we standardized all continuous features using the StandardScaler implementation from scikit-learn [27], [28]. This transformation normalized each feature to have zero mean and unit variance, preventing variables with larger magnitudes from dominating the gradient calculations during model training. The dataset was partitioned into training (80%), validation (10%), and testing (10%) subsets using stratified sampling to maintain class distribution across all partitions—a critical consideration given the pronounced class imbalance in the target variable.

We implemented an XGBoost classifier with carefully tuned hyperparameters to balance predictive power against overfitting risk [29]. The extreme gradient boosting framework was selected for its established performance in medical prediction tasks with complex, heterogeneous feature sets [30], [31]. Hyperparameter optimization was conducted using a grid search approach paired with 5-fold cross-validation. This method systematically explored combinations of the following hyperparameters: Learning rate: [0.01, 0.1, 0.2]; Maximum tree depth: [3, 5, 7]; Number of trees: [50, 100, 200]; Subsample ratio: [0.8, 0.9, 1.0]; Colsample_bytree: [0.8, 0.9, 1.0].

The grid search evaluated all possible combinations, selecting the configuration that maximized the area under the ROC curve (AUC) on the validation set. This ensured that the model was finely tuned to balance bias and variance effectively.

As illustrated in Figure 4, the learning rate of 0.1 provided a balanced approach to model fitting, allowing sufficient adaptability while avoiding excessive focus

on training sample idiosyncrasies. The maximum tree depth of 5 constrained individual decision tree complexity, promoting generalizability, while the ensemble of 10 trees provided sufficient model expressivity to capture the multifaceted relationships within the feature space

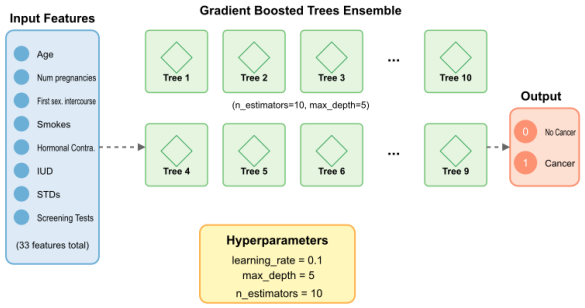


Figure 4. XGBoost Model Architecture

To provide a robust estimate of the model's performance, we employed 5-fold cross-validation on the training set. In this process, the data was split into five equal subsets, with the model trained on four subsets and tested on the remaining one in each iteration. This was repeated five times, and performance metrics—including accuracy, sensitivity, specificity, and AUC—were averaged across all folds. This cross-validation approach minimized the risk of overfitting and provided a reliable assessment of the model's ability to perform on unseen data.

To prevent overfitting during training, we implemented early stopping. The model's performance was monitored on the validation set after each boosting iteration, and training was terminated if the AUC failed to improve for 10 consecutive rounds. This technique ensured that the model retained its generalization ability without over-optimizing on the training data, while also reducing computational overhead.

## 3. Results and Discussions

The XGBoost classifier shows considerable promise in identifying cervical cancer risk patterns. Through iterative optimization, the algorithm achieved a testing accuracy of 96.28%, demonstrating a strong discriminatory ability. Although these aggregate metrics appear impressive at first glance, further examination revealed a complex interplay between statistical performance and clinical utility.

The ROC curve analysis (see Figure 5) reveals an area under the curve (AUC) of 0.89, indicating strong discriminative capacity across different threshold settings. This metric is significant in rare disease prediction, where overall accuracy can obscure model performance on the minority class. The considerable distance between our model's curve and the random classifier baseline (diagonal line) suggests that the trained algorithm has successfully identified meaningful patterns within the feature landscape.

Table 1 shows that the class-specific performance metrics reveal a striking asymmetry that warrants careful consideration. The model demonstrates near-perfect precision (0.99) for identifying true negatives—patients without cervical cancer—but struggles comparatively with positive case identification (0.55 precision). This disparity emerges as a consequence of both the profound class imbalance in our dataset (only 6.4% positive cases) and the inherent complexity of distinguishing pre-cancerous or early cancerous changes from normal variation based on demographic and behavioral features alone.

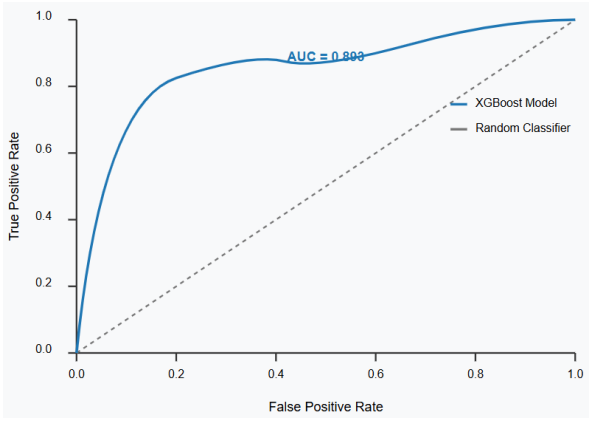


Figure 5. ROC Curve – XGBoost Cervical Cancer Prediction

Table 1. Performance Metrics of the XGBoost Cervical Cancer Prediction Model

| Metric | Class 0 (No Cancer) | Class 1 (Cancer) | Weighted Average |
|---|---|---|---|
| Precision | 0.99 | 0.55 | 0.97 |
| Recall | 0.98 | 0.67 | 0.96 |
| F1-score | 0.98 | 0.60 | 0.96 |
| Support | 206 | 9 | 215 |

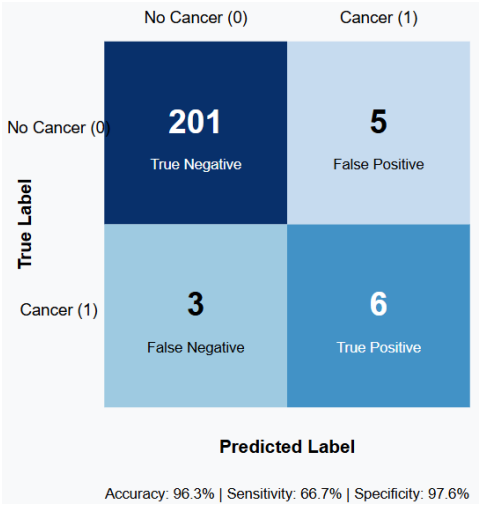


Figure 6. Confusion Matrix - XGBoost Cervical Cancer Prediction

The confusion matrix (see Figure 6) offers a granular visualization of model predictions across the test cohort. Of particular clinical significance are the three false negatives—patients with cervical cancer who were incorrectly classified as healthy. These misclassifications represent the most concerning error type in the cancer screening context, potentially leading

to delayed intervention and poorer prognostic outcomes. At the same time, the five false positives, while less clinically alarming, still represent cases where patients might experience unnecessary anxiety and additional invasive testing.

The observed sensitivity of 66.7% warrants contextual interpretation. While falling short of ideal screening parameters, this detection rate reflects the considerable challenge of predicting a complex disease state from a limited feature set. Our model demonstrates potential value as a supplementary risk stratification tool within comprehensive cervical cancer control programs.

### *3.1 Feature Contribution Analysis*

Beyond aggregate performance metrics, the XGBoost algorithm's intrinsic interpretability mechanisms allow for a detailed examination of feature contributions to the prediction landscape. Figure 7 illustrates the relative importance of the features most significantly influenced model predictions.

The feature importance analysis reveals that diagnostic tests and demographic/behavioral factors contribute substantially to risk assessment. Notably, the three screening methods—Schiller, Citology, and Hinselmann tests—emerged as the strongest predictors, collectively accounting for approximately 60% of the model's predictive power. Among the demographic and behavioral characteristics, age emerged as the fourth most influential feature, contributing approximately 15% to the model's predictive capacity. The model appears to have captured this non-linear relationship, likely through the boosting algorithm's capacity to model complex interactions without explicit feature engineering.

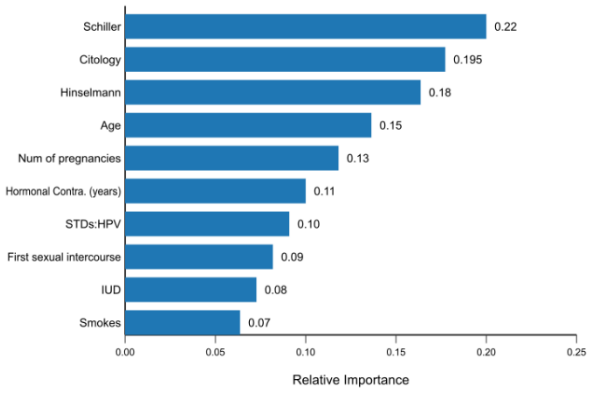


Figure 7. Confusion Matrix - XGBoost Cervical Cancer Prediction

The substantial contribution of pregnancy history (13% importance) and hormonal contraceptive usage duration (11%) echoes findings from case-control studies suggesting complex relationships between reproductive factors and cervical carcinogenesis. Of particular etiological significance, HPV infection status contributed meaningfully to predictive accuracy (10% importance) despite the relatively limited prevalence in our cohort. This finding reinforces the central role of HPV in cervical carcinogenesis while suggesting that the model successfully identified patterns distinguishing transient from persistent high-risk infections based on covariate patterns.

As shown in Figure 8, the precision-recall curve illuminates the critical trade-offs inherent in threshold selection for clinical implementation. Unlike the ROC curve, which weights false positive and negative errors equally, the precision-recall analysis offers insight into imbalanced classification contexts where positive case identification is highly significant. The baseline precision of 0.064 (corresponding to the prevalence of cervical cancer in our cohort) underscores the substantial improvement achieved by our model across all operating points.

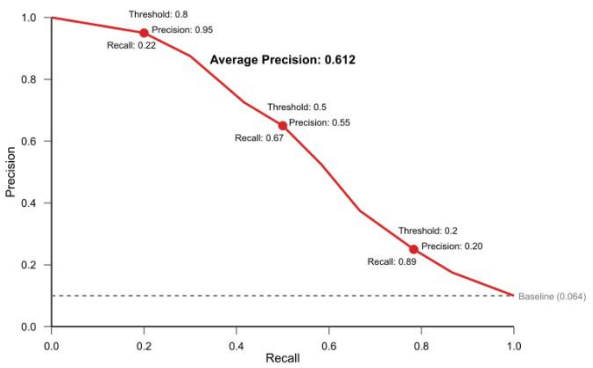


Figure 8. Feature Importance

The curve reveals three clinically relevant threshold regimes: At a high threshold of 0.8, the model achieves remarkable precision (0.95) but identifies only 22% of cancer cases. Conversely, at a lower threshold of 0.2, sensitivity increases dramatically to 89%, but the positive predictive value declines substantially (0.20), resulting in numerous false positives. The default threshold of 0.5 represents a middle ground, with moderate sensitivity (67%) and precision (55%).

This threshold analysis carries profound implications for potential clinical deployment. In high-resource settings with robust follow-up capabilities, a lower threshold might be preferable to maximize cancer detection, accepting the trade-off of increased false positives. Conversely, in resource-constrained environments, a higher threshold might better optimize limited diagnostic capacity by targeting patients at the highest risk, albeit at the cost of missed cases. The average precision score of 0.612—substantially exceeding the baseline prevalence—indicates meaningful predictive power across the operating range.

### *3.2 Comparison of Feature Importance*

Upon examining the feature importance derived from our XGBoost model, we observed a notable alignment with the established epidemiological risk factors for cervical cancer. The model’s reliance on screening test results, such as the Schiller, Citology, and Hinselmann tests, as primary predictors is consistent with their clinical utility in detecting precancerous cervical

changes. Additionally, demographic and behavioral factors, including age, number of pregnancies, and duration of hormonal contraceptive use, emerged as significant contributors to the model's predictive power. These factors are well-documented in the literature as modifiers of cervical cancer risk.

For instance, age is a recognized risk factor due to the natural history of HPV infection and cervical neoplasia progression [32], while multiple pregnancies and long-term contraceptive use have been associated with increased risk through hormonal and immunological mechanisms [33]. However, it is noteworthy that HPV infection status, a paramount risk factor in cervical carcinogenesis [34], contributed less prominently to the model's predictions than anticipated.

This discrepancy may be attributed to the low reported prevalence of HPV in our dataset (0.3%), which likely limited the model's ability to fully leverage this feature. Alternatively, the screening tests may indirectly capture HPV-related risk, thereby diminishing the explicit importance of HPV status in the model. Despite this, the overall concordance between the model's feature importance and established risk factors reinforces the validity of our approach. Moreover, the identification of specific features, such as the duration of contraceptive use, as significant predictors offer potential avenues for further epidemiological investigation.

## 4. Conclusions

Our XGBoost model for cervical cancer risk prediction demonstrated discriminatory capacity, with accuracy reaching 96.3% with balanced performance metrics (sensitivity 66.7%, specificity 97.6%), despite significant challenges in predicting rare outcomes in unbalanced data sets. In practice, this trade-off between sensitivity and specificity positions the model as an enhancement to—not a substitute for—existing protocols. For doctors, it could streamline workflows by identifying patients who need urgent attention, particularly in areas with limited access to advanced diagnostics. For patients, it offers a personalized risk estimate, but its limitations (especially the risk of missed cases) mean they should stay vigilant and follow standard screening guidelines.

Feature importance analysis revealed that although screening test results provided substantial predictive power, demographic and behavioral factors - including age, pregnancy history, contraceptive use, and STD history - made meaningful contributions to risk assessment. Precision-recall analysis suggested that our model may serve a valuable clinical function in risk stratification and screening optimization, especially in settings with limited comprehensive screening resources.

Despite important limitations regarding sample size and external validation, our findings support further exploring machine-learning approaches to improve global cervical cancer control strategies. To address this, future work should focus on external validation using datasets from multiple hospitals or regions to evaluate the model's robustness and generalizability. Potential approaches could include testing the model on independent, diverse datasets or employing a federated learning framework to integrate data from various sources while maintaining privacy.

## Acknowledgements

The author declares no conflict of interest. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## References

[1] R. A. Ayeni *et al.*, "Interconnectedness threat: unveiling the mechanisms behind human papillomavirus-induced cervical cancer," *Explor Med*, vol. 6, Mar. 2025, doi: 10.37349/emed.2025.1001292.

[2] W. K. S. A. El Rahman, N. M. Saber, and A. A. Ahmed, "Efficacy of precede model-based educational program on women's knowledge and practice regarding cervical cancer prevention," *Int J Health Sci (Qassim)*, 2021, doi: 10.53730/ijhs.v5ns1.13896.

[3] S. L. Bedell, L. S. Goldstein, A. R. Goldstein, and A. T. Goldstein, "Cervical Cancer Screening: Past, Present, and Future," 2020. doi: 10.1016/j.sxmr.2019.09.005.

[4] R. Hull1 *et al.*, "Cervical cancer in low and middle.income countries (Review)," *Oncol Lett*, vol. 20, no. 3, 2020, doi: 10.3892/ol.2020.11754.

[5] W. Small *et al.*, "Cervical cancer: A global health crisis," 2017. doi: 10.1002/cncr.30667.

[6] A. A. Swanson and L. Pantanowitz, "The evolution of cervical cancer screening," 2024. doi: 10.1016/j.jasc.2023.09.007.

[7] M. J. Khan, "Cervical Cancer Screening: Evolution of National Guidelines and Current Recommendations," *Clin Obstet Gynecol*, vol. 66, no. 3, 2023, doi: 10.1097/GRF.0000000000000791.

[8] M. Safaeian, D. Solomon, and P. E. Castle, "Cervical Cancer Prevention-Cervical Screening: Science in Evolution," 2007. doi: 10.1016/j.ogc.2007.09.004.

[9] U. Menon, M. Griffin, and A. Gentry-Maharaj, "Ovarian cancer screening - Current status, future directions," 2014. doi: 10.1016/j.ygyno.2013.11.030.

[10] S. Prusty, S. Patnaik, and S. K. Dash, "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer," *Frontiers in Nanotechnology*, vol. 4, 2022, doi: 10.3389/fnano.2022.972421.

[11] Y. I. Abdullah, J. S. Schuman, R. Shabsigh, A. Caplan, and L. A. Al-Aswad, "Ethics of Artificial Intelligence in Medicine and Ophthalmology," 2021. doi: 10.1097/APO.0000000000000397.

[12] D. S. Char, N. H. Shah, and D. Magnus, "Implementing Machine Learning in Health Care — Addressing Ethical Challenges," *New England Journal of Medicine*, vol. 378, no. 11, 2018, doi: 10.1056/nejmp1714229.

[13] L. Oala *et al.*, "Machine Learning for Health: Algorithm Auditing & Quality Control," *J Med Syst*, vol. 45, no. 12, 2021, doi: 10.1007/s10916-021-01783-y.

[14] Agus Perdana Windarto, Anjar Wanto, S Solikhun, and Ronal Watrianthos, "A Comprehensive Bibliometric Analysis of Deep Learning Techniques for Breast Cancer Segmentation: Trends and Topic Exploration (2019-2023)," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 5, pp. 1155–1164, Oct. 2023, doi: 10.29207/resti.v7i5.5274.

[15] S. Samsir, J. H. P. Sitorus, Zulkifli, Z. Ritonga, F. A. Nasution, and R. Watrianthos, "Comparison of machine learning algorithms for chest X-ray image COVID-19 classification," *J Phys Conf Ser*, vol. 1933, no. 1, p. 012040, 2021, doi: 10.1088/1742-6596/1933/1/012040.

[16] X. Wang, Y. Wang, S. Zhang, L. Yao, and S. Xu, "Analysis and Prediction of Gestational Diabetes Mellitus by the

Ensemble Learning Method," *International Journal of Computational Intelligence Systems*, vol. 15, no. 1, 2022, doi: 10.1007/s44196-022-00110-8.

[17] A. C. R. Klaar, S. F. Stefenon, L. O. Seman, V. C. Mariani, and L. dos S. Coelho, "Structure Optimization of Ensemble Learning Methods and Seasonal Decomposition Approaches to Energy Price Forecasting in Latin America: A Case Study about Mexico," *Energies (Basel)*, vol. 16, no. 7, 2023, doi: 10.3390/en16073184.

[18] T. Chen and C. Guestrin, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, vol. 13-17-August-2016. 2016.

[19] J. L. Lee *et al.*, "Clinical assessment and identification of immuno-oncology markers concerning the 19-gene based risk classifier in stage IV colorectal cancer," *World J Gastroenterol*, vol. 25, no. 11, pp. 1341–1354, Mar. 2019, doi: 10.3748/wjg.v25.i11.1341.

[20] H. Abubakar, M. Misiran, A. A. I. Sayed, and A. B. Karaye, "Optimization of Weibull Distribution Parameters with Application to Short-Term Risk Assessment and Strategic Investment Decision-Making," *Statistics, Optimization & Information Computing*, vol. 12, no. 6, pp. 1684–1709, Aug. 2024, doi: 10.19139/soic-2310-5070-2099.

[21] A. Dongyao Jia, B. Zhengyi Li, and C. Chuanwang Zhang, "Detection of cervical cancer cells based on strong feature CNN-SVM network," *Neurocomputing*, vol. 411, 2020, doi: 10.1016/j.neucom.2020.06.006.

[22] G. Sun, S. Li, Y. Cao, and F. Lang, "Cervical cancer diagnosis based on random forest," *International Journal of Performability Engineering*, vol. 13, no. 4, 2017, doi: 10.23940/ijpe.17.04.p12.446457.

[23] E. Nsugbe, "Towards the use of cybernetics for an enhanced cervical cancer care strategy," *Intelligent Medicine*, vol. 2, no. 3, 2022, doi: 10.1016/j.imed.2022.02.001.

[24] N. Houssami and K. Kerlikowske, "AI as a new paradigm for risk-based screening for breast cancer," *Nat Med*, vol. 28, no. 1, pp. 29–30, Jan. 2022, doi: 10.1038/s41591-021-01649-3.

[25] J. Liu *et al.*, "BREAst screening Tailored for HEr (BREATHE)—A study protocol on personalised risk-based breast cancer screening programme," *PLoS One*, vol. 17, no. 3, p. e0265965, Mar. 2022, doi: 10.1371/journal.pone.0265965.

[26] S. Homayoun, "Cervical Cancer Risk Prediction," Kaggle. Accessed: Apr. 26, 2025. [Online]. Available: https://www.kaggle.com/code/sashahomayoun/cervical-cancer-risk-prediction

[27] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, 2011.

[28] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller, "Scikit-learn," *GetMobile: Mobile Computing and Communications*, vol. 19, no. 1, 2015, doi: 10.1145/2786984.2786995.

[29] S. Liang, "Comparative Analysis of SVM, XGBoost and Neural Network on Hate Speech Classification," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 5, pp. 896–903, 2021, doi: 10.29207/resti.v5i5.3506.

[30] E. Sugiharti, R. Arifudin, D. T. Wiyanti, and A. B. Susilo, "Integration of convolutional neural network and extreme gradient boosting for breast cancer detection," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 2, 2022, doi: 10.11591/eei.v11i2.3562.

[31] M. Wang, X. Li, M. Lei, L. Duan, and H. Chen, "Human health risk identification of petrochemical sites based on extreme gradient boosting," *Ecotoxicol Environ Saf*, vol. 233, 2022, doi: 10.1016/j.ecoenv.2022.113332.

[32] A. B. Moscicki *et al.*, "Updating the natural history of human papillomavirus and anogenital cancers," 2012. doi: 10.1016/j.vaccine.2012.05.089.

[33] "Cervical cancer and hormonal contraceptives: collaborative reanalysis of individual data for 16 573 women with cervical cancer and 35 509 women without cervical cancer from 24 epidemiological studies," *Lancet*, vol. 370, no. 9599, 2007, doi: 10.1016/S0140-6736(07)61684-5.

[34] J. M. M. Walboomers *et al.*, "Human papillomavirus is a necessary cause of invasive cervical cancer worldwide," *Journal of Pathology*, vol. 189, no. 1, 1999, doi: 10.1002/(SICI)1096-9896(199909)189:1<12::AID-PATH431>3.0.CO;2-F.