Jurnal Ilmiah Permas: Jurnal Ilmiah STIKES Kendal
Volume 15 Nomor 4, Oktober 2025
e-ISSN 2549-8134; p-ISSN 2089-0834
http://journal.stikeskendal.ac.id/index.php/PSKM

CERVICAL CANCER RISK PREDICTION WITH MACHINE LEARNING: ANALYSIS OF CERVICAL CANCER RISK CLASSIFICATION DATASET

Yoga Paripurna*, Irwan Budiono
Department of Public Health, Universitas Negeri Semarang, Sekaran, Gn. Pati, Semarang, Jawa Tengah 50229
Indonesia
*yogaparipurna0@gmail.com

ABSTRACT

Cervical cancer remains one of the leading causes of cancer death in women, especially in developing countries. Early detection through screening is essential to reduce morbidity and mortality, but the main challenge is to identify individuals at high risk efficiently. This study aims to build a machine learning prediction model to classify cervical cancer biopsy results based on available risk factors. Objectives: This study aims to build a cervical cancer risk prediction model using a machine learning algorithm based on available risk factors. The public dataset "Cervical Cancer Risk Classification" includes demographic data, sexual behavior, contraceptive use, and medical test results. Three machine learning algorithms are applied: Logistic Regression, Decision Tree, and Support Vector Machine (SVM). Model evaluation uses accuracy, precision, recall, F1 score, and Matthews Correlation Coefficient (MCC). The Decision Tree model performed best with an F1 Score of 0.956 and MCC of 0.639. Significant contributing risk factors are age, age at first sexual intercourse, Schiller test results, cytology, and number of pregnancies. Machine learning has great potential to improve the effectiveness of cervical cancer screening. Data balancing techniques and ensemble methods are recommended to increase accuracy in detecting positive cases.

Keywords: cervical cancer; early detection; machine learning; decision tree; SVM; risk prediction; screening; schiller test

INTRODUCTION

Cervical cancer is a significant public health issue globally, particularly in developing countries, where it remains the second leading cause of cancer-related deaths among women after breast cancer (Vu et al., 2018; Sun et al., 2019). Despite the availability of screening methods such as the Pap smear, Human Papillomavirus (HPV) testing, and biopsy procedures, early detection continues to pose a major challenge (Barquet-Muñoz et al., 2024). Early diagnosis is crucial as it significantly lowers both morbidity and mortality associated with the disease (Tobore, 2019). Nevertheless, participation in cervical cancer screening remains low in many regions. Factors such as limited healthcare access, low awareness levels, and disparities in medical resources contribute to this issue (Greenley et al., 2023; Israel, 2022). Consequently, many cases are identified at advanced stages when prognosis worsens. To improve early identification, innovative strategies are necessary, including the integration of technological advancements like data-driven analytics (Sreelatha & Shivashetty, 2023).

In the digital era, machine learning (ML) offers a promising solution to enhance early disease detection, including cervical cancer (Shetty & Shah, 2018). ML can analyze large datasets to uncover complex patterns that may elude traditional analysis (MacEachern & Forkert, 2021). By integrating diverse risk factors such as sexual behavior, contraceptive use, smoking habits, and clinical findings, ML can enable more accurate and rapid prediction of cervical cancer risk (Yadav et al., 2025).This study utilizes the Cervical Cancer Risk Classification dataset to explore the application of three machine learning algorithms: Logistic Regression, Decision Tree, and Support Vector Machine (SVM), chosen for their suitability in binary classification tasks, interpretability, and predictive performance (Uddin et al., 2025; Gimeno et al., 2023). The objective is to gain a deeper understanding of the contributing risk factors and demonstrate how ML can enhance screening effectiveness. Ultimately, these findings may support the development of innovative clinical decision support systems to expedite medical interventions, improve patient outcomes, and reduce cervical cancer mortality.

METHOD

Research Design

This study is a quantitative study with a descriptive-analytical approach. The purpose of the study is to build a classification model to predict cervical cancer biopsy results based on available risk factor data, and to compare the performance of several machine learning algorithms.

Data Source

The data used in this study is the publicly available Cervical Cancer Risk Classification dataset. This dataset contains information on various risk factors related to cervical cancer, including demographic data, sexual behavior, history of contraceptive use, smoking habits, and clinical test results such as the Schiller, Hinselmann, and cytology tests. Dataset was sourced from https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification/data

Research Procedure

The research procedure consisted of several key steps. First, data pre-processing was conducted, which involved cleaning the dataset to address missing values and inconsistencies, encoding data into numeric formats for machine learning compatibility, removing irrelevant features, and performing normalization when necessary. In the data exploration phase, descriptive analysis was used to examine variable distributions and understand dataset characteristics, while correlation analysis identified factors most associated with biopsy results. During machine learning model development, the dataset was split into training and test sets with an 80:20 ratio, and three algorithms—Logistic Regression, Decision Tree, and Support Vector Machine (SVM)—were applied. Each model underwent hyperparameter tuning to optimize performance. In the model evaluation stage, performance was assessed using accuracy, precision, recall, F1 score, and Matthews Correlation Coefficient (MCC), followed by a comparative analysis of the models. All analyses were performed using Python 3.10, utilizing the scikit-learn, pandas, and matplotlib libraries for data analysis, model development, and visualization.

Research Ethics

The data used are publicly available, so no additional ethics approval was required for this study.

RESULT

Data Pre-Processing

The Cervical Cancer Risk Classification dataset has 858 samples with 36 features. As many as 11.7% of the data have missing values. Pre-processing is done by imputing missing data, encoding categorical features into numeric ones, selecting relevant features, and normalizing data if necessary.

Data Exploration

The distribution of biopsy results shows that most samples have negative results.

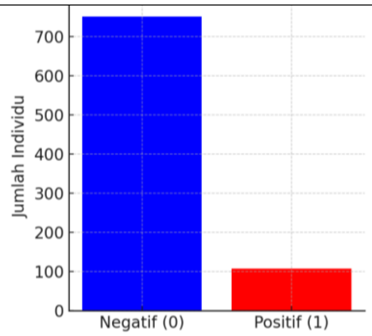

Figure 1. Distribution of Biopsy Results.

Machine Learning Model Development

The data is divided into 80% training data and 20% test data. The three algorithms used are Logistic Regression, Decision Tree, and Support Vector Machine (SVM). Hyperparameter tuning is performed for each model.

Model Evaluation

Table 1.
Machine Learning Model Evaluation Results.

Model	AUC	CA (Accuracy)	F1 Score	Precision	Recall	MCC
Logistic Regression	0.900	0.948	0.947	0.947	0.948	0.566
Decision Tree	0.695	0.959	0.956	0.956	0.959	0.639
SVM	0.936	0.935	0.927	0.928	0.935	0.367

Feature Importance - Decision Tree

The most influential features in predicting biopsy results were Schiller (most influential), Age at first sexual intercourse, patient age, cytology, number of pregnancies, and history of sexually transmitted infections (HIV).

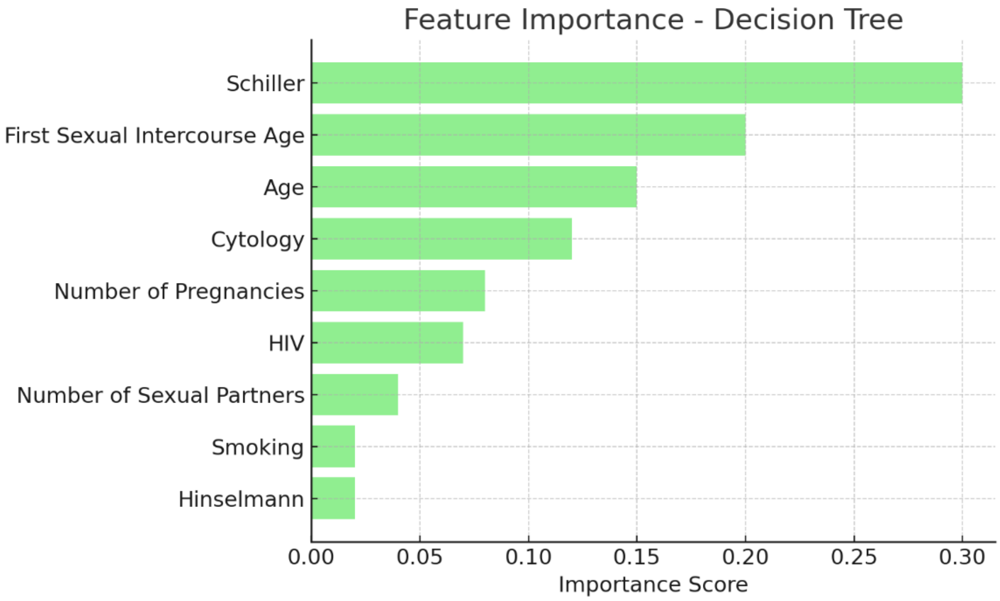

Figure 2. Feature Importance Graph in the Decision Tree Model.

The best model for cervical cancer risk prediction based on this dataset is Decision Tree, with the highest F1 Score (0.956) and the highest MCC (0.639). Decision Tree is able to provide the best balance between specificity and sensitivity in classification.

DISCUSSION

The results of this study indicate that variables such as age, number of pregnancies, Schiller test outcomes, and cytology results play a significant role in predicting the risk of cervical cancer. These findings are consistent with those reported by Ashar et al. (2020), who identified that early age at first sexual intercourse and high parity were significantly associated with increased risk of precancerous cervical lesions. They also emphasized that prolonged HPV infections, particularly in younger individuals, tend to elevate the risk of cervical cancer. Similar observations were made by Piyathilake et al. (2023), who utilized a machine learning approach and found that reproductive history and cytology results were critical factors in predictive modeling for cervical cancer.This study also emphasizes the importance of the Schiller test, which was shown to be a strong predictor of cervical cancer risk. This aligns with Ashar et al. (2020), who reported the high sensitivity of the Schiller test in detecting abnormal cervical changes, underscoring its utility as an early screening tool. Nevertheless, some discrepancies were noted between this study and previous research regarding the relative importance of predictive features. For instance, in the current study, age emerged as a more dominant factor than cytology, which might be attributed to differences in data collection methods or variations in the populations studied.

An important aspect highlighted in this study concerns class imbalance within the dataset. The Decision Tree model demonstrated better performance in classifying negative biopsy results but struggled with detecting false negatives. This observation is consistent with the findings of Kuruvilla and Jayanthi (2022), who noted similar limitations in machine learning models when faced with imbalanced datasets in cervical cancer classification. They suggested applying data balancing techniques such as SMOTE to reduce misclassification of minority class samples and enhance model sensitivity.The success of the Decision Tree algorithm in this study can be attributed to its capability to handle complex features and non-linear relationships among variables. This supports findings by Battista et al. (2023), who reported that decision tree models are particularly effective in identifying intricate interactions in medical datasets. However, the ongoing difficulty in accurately classifying positive cervical cancer cases remains a significant challenge. As such, the application of ensemble learning methods—such as Random Forest or Gradient Boosting—may offer improved sensitivity and accuracy in detecting true positive cases.

In summary, the results of this study reaffirm that while machine learning models offer substantial potential in supporting cervical cancer screening, major challenges remain in managing data imbalance and improving minority class predictions. Variables like age, number of pregnancies, and Schiller test results have shown predictive strength and should be considered key factors in model development. Future work should prioritize the use of balancing techniques and ensemble methods to enhance model performance in clinical applications.

CONCLUSION

This study successfully identified the main risk factors influencing the likelihood of biopsy results in detecting cervical cancer. Factors such as age, number of pregnancies, Schiller test results, and cytology were shown to have significant contributions in predicting biopsy results, which aligns with previous findings in the literature. The Decision Tree model showed the best performance in classifying biopsy results with a high F1 Score and significant Matthews Correlation Coefficient (MCC). However, there were still challenges in classifying cervical cancer-positive cases (false negatives). Overall, the findings of this study provide important insights for early detection of cervical cancer, emphasizing the importance of measurable risk factors for better prediction. This study also contributes to the development of machine learning-based prediction models that can be used to improve the efficiency of medical decision-making and reduce unnecessary biopsies.

REFERENCES

Ashar, H., Kusrini, I., Musoddaq, A., & Asturiningtyas, I. P. (2020). First sexual intercourse and high parity are the most influential factors of precancerous cervical lesion. Majalah Obstetri & Ginekologi, 28(3), 113–118. https://doi.org/10.20473/mog.v28i32020.113-118

Barquet-Muñoz, S. A., Arteaga-Gómez, C., Díaz-López, E., Rodríguez-Trejo, A., Marquez-Acosta, J., & Aranda-Flores, C. (2024). Current status and challenges in timely detection of cervical cancer in Mexico: Expert consensus. Frontiers in Oncology, 14, 1383105. https://doi.org/10.3389/fonc.2024.1383105

Battista, K., Diao, L., Patte, K. A., Dubin, J. A., & Leatherdale, S. T. (2023). Examining the use of decision trees in population health surveillance research: An application to youth mental health survey data in the COMPASS study. Health Promotion and Chronic Disease Prevention in Canada, 43(2), 73–86. https://doi.org/10.24095/hpcdp.43.2.03

Gimeno, M., Sada Del Real, K., & Rubio, A. (2023). Precision oncology: A review to assess interpretability in several explainable methods. Briefings in Bioinformatics, 24(4), bbad200. https://doi.org/10.1093/bib/bbad200

Greenley, R., Bell, S., Rigby, S., Legood, R., Kirkby, V., McKee, M., & CBIG-SCREEN Consortium. (2023). Factors influencing the participation of groups identified as underserved in cervical cancer screening in Europe: A scoping review of the literature. Frontiers in Public Health, 11, 1144674. https://doi.org/10.3389/fpubh.2023.1144674

Israel, A. (2022). Partnering to strengthen health systems and improve access to quality cancer care. JCO Global Oncology, 8, e2200148. https://doi.org/10.1200/GO.22.00148

Kuruvilla, A., & Jayanthi, B. (2022). Analysis and review on feature selection and classification methods on cervical cancer. ICTACT Journal on Soft Computing, 12(2), 2551–2558. https://doi.org/10.21917/ijsc.2022.0365

MacEachern, S. J., & Forkert, N. D. (2021). Machine learning for precision medicine. Genome, 64(4), 416–425. https://doi.org/10.1139/gen-2020-0131

Piyathilake, C. J., Badiga, S., & Jolly, P. E. (2023). Potential effects of age-based changes in screening guidelines on the identification of women at risk for developing cervical cancer. Cancer Prevention Research, 16(2), 99–108. https://doi.org/10.1158/1940-6207.CAPR-22-0426

Shetty, A., & Shah, V. (2018). Survey of cervical cancer prediction using machine learning: A comparative approach. In 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1–6). IEEE. https://doi.org/10.1109/ICCCNT.2018.8494169

Sreelatha, S., & Shivashetty, V. (2023). Proactive cervical cancer risk assessment using data-driven analytics. International Journal of Artificial Intelligence, 13(4), 4301–4311. https://doi.org/10.11591/ijai.v13.i4.pp4301-4311

Sun, W., Shen, N.-M., & Fu, S.-L. (2019). Involvement of lncRNA-mediated signaling pathway in the development of cervical cancer. European Review for Medical and

Pharmacological Sciences, 23(9), 3672–3687. https://doi.org/10.26355/eurrev_201905_17791

Tobore, O. (2019). On the need for the development of a cancer early detection, diagnostic, prognosis, and treatment response system. Future Science OA, 6(2), FSO439. https://doi.org/10.2144/fsoa-2019-0028

Uddin, K. M. M., Sikder, I. A., & Hasan, M. N. (2025). A comparative study on machine learning classifiers for cervical cancer prediction: A predictive analytic approach. EAI Endorsed Transactions on Internet of Things, 11, e6223. https://doi.org/10.4108/eetiot.6223

Vu, M., Yu, J., Awolude, O. A., & Chuang, L. (2018). Cervical cancer worldwide. Current Problems in Cancer, 42(5), 457–465. https://doi.org/10.1016/j.currproblcancer.2018.06.003

Yadav, U., Bondre, V. D., Bondre, S. V., Thakre, B., Agrawal, P., & Thakur, S. (2025). Intelligent cervical cancer detection: Empowering healthcare with machine learning algorithms. International Journal of Artificial Intelligence, 14(1), 298–306. https://doi.org/10.11591/ijai.v14.i1.pp298-306