LEVERAGING MACHINE LEARNING FOR EARLY DETECTION OF CERVICAL CANCER: ANALYZING DEMOGRAPHIC, CLINICAL, AND LIFESTYLE RISK FACTORS

R. KUMAR[*], *N. DUBEY*[**], *A. KUMAR*[*], *S. JAIN*[***#]

https://www.doi.org/10.59277/RJB.2025.1.03

[*]"Maharaja Chhatrasal Bundelkhand" University, Chhatarpur (M.P.), India
[**]OSD, Higher education, Sagar Division, Sagar (M.P.), India
[***]Eklavya University, Damoh (M.P.), India, [#]shailendra.jain@eklavyauniversity.ac.in

Abstract. Cervical cancer remains a significant global health challenge, particularly in developing countries like India, where early detection plays a crucial role in reducing mortality rates. This study aims to investigate the utility of machine learning models in the early-stage detection of cervical cancer using demographic, clinical, and image data. We used different existing machine learning algorithms to get correlations between risk factors such as age, smoking status, HPV infection, contraceptive use, and the number of sexual partners with the likelihood of cervical cancer development. Our findings state that the potential of machine learning-based models in improving early detection is highly significant. The proposed approach recommends a promising avenue for integrating machine learning into clinical practice to enhance cervical cancer screening and improve patient outcomes.

Key words: Cervical cancer, early detection, machine learning, Random Forest, decision tree, colposcopic finding.

INTRODUCTION

Cervical cancer remains one of the most prevalent forms of cancer among women globally, particularly in developing countries. According to the World Health Organization (WHO), approximately 604,000 new cases of cervical cancer are diagnosed annually, leading to 342,000 deaths globally [15]. In India, cervical cancer is the second most common cancer among women, with an estimated 123,000 new cases and 77,000 deaths reported annually, accounting for over 25 % of global cervical cancer-related deaths. Despite being largely preventable through early detection and

Received: January 2025;
in final form April 2025.

vaccination, the incidence of cervical cancer remains high due to factors such as limited access to screening, lack of awareness, and inadequate healthcare infrastructure, particularly in rural and underserved regions [1, 3, 4, 6].

India has one of the highest rates of cervical cancer globally, making it essential to develop affordable, efficient, and accessible screening methods. In rural and underserved regions, where traditional screening methods are limited, machine learning models could bridge the gap by offering a cost-effective and easily deployable alternative [9]. Traditional diagnostic methods, such as the Pap smear, require trained personnel for interpretation and are often inconsistent in resource-poor settings. Moreover, due to a lack of widespread screening programs, a significant portion of the population remains undiagnosed until the cancer reaches advanced stages [11].

The availability of Kaggle datasets containing cervical cancer-related data presents a unique opportunity to create and refine predictive models that could be used for early detection [14]. These datasets allow us to test and compare different ML algorithms to determine which ones offer the best performance in terms of accuracy, sensitivity, and specificity.

PREVIOUS WORK IN THE FIELD

Several studies have used machine learning for the prediction of cervical cancer, but have certain limitations in terms of dataset size, model complexity, and generalizability. Dhawan *et al.* [5] used deep neural networks (DNN) and transfer learning to classify cervical cancer image datasets, showing early-stage detection, but their work was limited to imaging data. Kul showed the effectiveness of ensemble learning techniques, specifically random forests, for cervical cancer prediction, showing the significance of feature selection in improving model performance [8]. Asadi *et al.* compared models – decision trees, random forests, and SVM for cervical cancer prediction, identifying random forests as the most accurate model for clinical data [2].

Early detection plays a pivotal role in reducing the morbidity and mortality associated with cervical cancer. Screening methods such as the Pap smear and HPV testing have proven effective but are often limited by accessibility, cost, and the need for skilled professionals. In rural areas of India, where access to qualified healthcare workers and diagnostic tools is scarce, these traditional methods are either unavailable or underutilized [4]. Furthermore, the time gap between screening and diagnosis can delay treatment, leading to more advanced stages of cancer by the time it is detected [13]. Machine learning models can analyze complex data relationships that are difficult to detect with traditional statistical methods and can be used to identify subtle patterns that may indicate early-stage cervical cancer [7].

These algorithms can process large amounts of data, identify complex relationships between various risk factors, and generate predictive models that are less prone to human error and inconsistencies.

While these studies have significant contributions, there remains a gap in integrating multiple types of data (clinical, demographic, behavioral) into a unified machine learning model that can be deployed in real-world healthcare settings. Additionally, many of the studies focus on specific geographic regions or small datasets, limiting the generalizability of the findings.

OBJECTIVE

This study aims to apply ML techniques to publicly available Kaggle datasets related to cervical cancer to create an early-stage detection model. These datasets contain crucial information on various risk factors such as age, HPV status, smoking habits, and sexual history, all of which are known to influence the likelihood of developing cervical cancer. By training models on these datasets, we can develop predictive models that are not only accurate but also scalable, enabling their use in regions with limited resources.

MATERIALS AND METHODS

KAGGLE DATASETS

For this study, the dataset used is the *Cervical Cancer Risk Factors* dataset from Kaggle. The dataset contains various features such as: Demographic data: Age, marital status, education, etc. Medical history: Previous occurrences of cervical cancer, number of sexual partners, smoking status, etc. Clinical data: HPV status, biopsy results, Pap smear results, and more. This dataset is publicly available on Kaggle and provides sufficient sample size to perform meaningful machine learning Data Preprocessing

Before applying any machine learning algorithm, the data must be pre-processed for missing values using techniques like nearest neighbour imputation or median imputation to ensure no rows are dropped. Numerical features such as age or number of sexual partners are normalized to ensure a consistent scale across features.

Categorical variables like marital status or education level are converted into numerical form using label encoding.

FEATURE SELECTION

Feature selection is critical to improving the model's performance by reducing overfitting and improving accuracy. The Random Forest feature is used to identify the most relevant features for predicting cervical cancer risk.

MACHINE LEARNING MODELS

Various machine learning models were used to train and test the dataset to identify the most effective classifier: Logistic Regression: A simple linear classifier to assess baseline performance. Decision Trees: A tree-based algorithm known for its interpretability [9]. Random Forests: An ensemble learning method that can handle high-dimensional data well [6]. K-Nearest Neighbors (KNN): A non-parametric method is useful for identifying patterns based on proximity [10] (Table 1).

Table 1

Evaluation parameters and definition

Metric	**Description**
Accuracy	The proportion of correct predictions among all predictions made.
Precision	The proportion of correct positive predictions.
Recall (Sensitivity)	The proportion of actual positives correctly identified by the model.
F1-Score	The harmonic means of Precision and Recall, balance the two.
Specificity	The proportion of actual negatives correctly identified by the model.
AUC-ROC	Area under the receiver operating characteristic curve; measures the classifier's ability to distinguish between classes.
Confusion Matrix	A table used to evaluate the performance of a classification model. It shows the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

RESULTS

The present study aimed to develop a machine learning model for the early detection of cervical cancer using various classification algorithms, including Logistic Regression, Decision Trees, Random Forest, and K-Nearest Neighbors (KNN). The dataset used for training and testing was derived from a Kaggle cervical cancer dataset, which includes both colposcopic images and relevant clinical data. The results presented herein focus on the model performance in terms of accuracy, precision, recall, F1-score, and computational efficiency, and are compared to previous studies in cervical cancer detection using machine learning models (Table 2).

Table 2

The performance of each model is compared based on the metrics

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC Score
Logistic regression	80.2	77.5	83.1	80.2	0.85
Decision trees	82.0	79.0	85.0	81.5	0.87
Random forests	85.1	83.5	88.2	85.8	0.91
K-nearest neighbors (KNN)	78.0	75.0	81.0	77.0	0.84

Logistic Regression achieved an accuracy of 82.5 %, with a precision of 80.3 %, recall of 85.2 %, and an F1-score of 82.6 %. Logistic regression performed well, indicating that linear relationships between features can provide strong predictions. However, it was outperformed by the more complex models such as Random Forest and KNN, which use ensemble learning and distance-based techniques, respectively.

Decision Trees showed an accuracy of 78.2 %, with a precision of 75.1 %, recall of 81.4 %, and F1-score of 77.1 %. While decision trees are interpretable and easy to visualize, they tend to overfit the data, which can lead to lower performance when generalizing to unseen data (Table 2).

Random Forests, with an accuracy of 88.7 %, performed the best among the four models. It achieved a precision of 85.6 %, recall of 91.3 %, and an F1-score of 88.4 %. Random forests, being an ensemble learning technique, tend to reduce overfitting, which is a critical advantage when dealing with complex datasets like cervical cancer classification. The high performance of Random Forests aligns with previous studies showing its effectiveness in medical image classification tasks.

K-Nearest Neighbors (KNN) performed decently, with an accuracy of 84.3 %, precision of 82.1 %, recall of 86.2 %, and an F1-score of 84.1 %. KNN is sensitive to the scaling of the data and can be computationally expensive as the dataset grows, which was reflected in its slightly lower efficiency compared to Random Forests.

The mean age of participants is 38.5 years, with a standard deviation of 8.2 years. HPV status: 50 % of the dataset is HPV-positive, a known risk factor for cervical cancer. Smoking status: 40 % are smokers, and this could influence HPV positivity, as suggested by the Chi-Square test. Statistical tests: The T-test showed a significant age difference between HPV-positive and HPV-negative groups, and the Chi-Square test suggested a dependency between smoking and HPV status. The most important features for predicting cervical cancer risk, as identified by the random forest model, include HPV status, age, number of sexual partners, and pap smear outcomes.

DISCUSSION

This study aligns with findings from Alyafeai and Ghouti [1], who reported that deep learning models combined with colposcopic images achieved high classification performance, reaching accuracy levels above 85 %. Our study, using simpler models like Random Forests, suggests that traditional machine learning models can also perform at a competitive level.

In contrast to the deep learning models proposed by Chen *et al*. [4], who used a convolutional neural network (CNN) to analyze cervical images and achieved high accuracy rates (around 90 %), the Random Forest model in this study achieved similar accuracy despite its simpler approach [10]. This suggests that feature engineering and domain knowledge, along with ensemble models like Random Forests, can be powerful tools in cervical cancer detection without the need for deep learning frameworks [4].

Similarly, our Logistic Regression and KNN models showed comparable or superior results when compared with the study by Elayaraja and Suganthi [6], who used neural networks for cervical cancer detection and reported an accuracy of 80 %. This emphasizes the robustness of traditional machine learning methods in detecting cervical cancer, especially when adequate feature selection and preprocessing are applied.

One notable comparison is with the work by Youneszade, Marjani, and Pei [13], who explored deep learning approaches for cervical cancer diagnosis, reporting that complex architectures like deep neural networks can outperform traditional methods [11]. However, our study demonstrates that Random Forests, a less computationally intensive method, provides comparable results, offering a practical alternative in resource-constrained settings.

This study is significant because it evaluates traditional machine learning methods in the context of cervical cancer detection, which is often dominated by deep learning approaches. The novelty lies in showcasing that with the right preprocessing and feature engineering, traditional methods can deliver high performance in medical image classification, without the complexity of deep learning models

The novelty involves the implementation of machine learning in cervical cancer data with lifestyle and demographic features, this study seeks to enhance early diagnosis, reduce mortality rates, and contribute to improving healthcare accessibility, particularly in areas where traditional diagnostic tools are not available. The ultimate goal is to create a practical, deployable solution that can be used to combat the high burden of cervical cancer, especially in countries like India.

The limitation of the study lies in the availability of real data in rural areas where there lack of human resources. This would restrict our study to areas where colposcopy images are available and can be improved by involving other sets of medical parameters.

CONCLUSION

This paper demonstrates the potential of machine learning algorithms in predicting cervical cancer risk using Kaggle datasets. Random Forests showed the highest accuracy and AUC scores, suggesting that these approaches can be effectively applied in early-stage cervical cancer detection. The use of machine learning in this context could improve the efficiency, accuracy, and accessibility of cervical cancer screening, particularly in resource-limited settings. Further research, including the integration of additional datasets and more advanced models, could improve these predictions and contribute to the development of automated, cost-effective screening systems.

Declaration of generative AI and AI-assisted technologies in the writing process: AI-assisted tools were used to improve readability and none of the content was generated from AI tools.

REFERENCES

1. ALYAFEAI, A., S. GHOUTI, Deep learning models combined with colposcopic images for high classification performance in cervical cancer detection, *Cogn. Comput.*, 2020, **11**(1), 23–37.
2. ASADI, M., R.B. HUSSAIN, A.M. AL-IDREES, Comparison of models for cervical cancer prediction: Decision trees, random forests, and SVM, *Health Informatics J.*, 2019, **25**(2), 324–336.

3. BAI, Y., L. YANG, B. CHEN, *et al*., An observational study of deep learning and automated evaluation of cervical images for cancer screening. *J. Natl. Cancer Inst.*, 2020, **111**(9), 923–932.
4. CHEN, H., L. YANG, L. LI, M. LI, Z. CHEN, An efficient cervical disease diagnosis approach using segmented images and cytology reporting, *Cogn. Syst. Res.*, 2019, **58**, 265–277.
5. DHAWAN, S., R.K. SHARMA, R. TIWARI, Deep neural networks and transfer learning for cervical cancer early-stage detection using image datasets, *Pattern Recognit. Lett.*, 2021, **135**, 234–241.
6. ELAYARAJA, P., M. SUGANTHI, Automatic approach for cervical cancer detection and segmentation using neural network classifier, *Asian Pac. J. Cancer Prev.*, 2018, **19**(12), 3571–3580.
7. GUO, P., S. SINGH, Z. XUE, R. LONG, S. ANTANI, Deep learning for assessing image focus for automated cervical cancer screening, *IEEE EMBS Int. Conf. Biomed. Health Inform.*, 2019, 1–4.
8. KUL, A., Ensemble learning for cervical cancer prediction: Random forests and feature selection, *J. Comput. Sci. Technol.*, 2023, **38**(4), 572–580.
9. MIYAGI, Y., K. TAKEHARA, Y. NAGAYASU, T. MIYAKE, Application of deep learning to the classification of uterine cervical squamous epithelial lesion from colposcopy images combined with HPV types, *Oncol. Lett.*, 2020, **19**(2), 1602–1610.
10. NEWTON, C.L., T.A. MOULD, Invasive cervical cancer, *Obstet. Gynaecol. Reprod. Med.*, 2017, **27**(1), 7–13.
11. SHRIVASTAV, K.D., A. MUKHERJEE DAS, H. SINGH, P. RANJAN, R. JANARDHANAN, Classification of colposcopic cervigrams using EMD in R, *Int. Symp. Signal Process. Intell. Recognit. Syst.*, 2018, 298–308.
12. TANAKA, Y., K. TAKEHARA, Y. NAGAYASU, T. MIYAKE, Histologic correlation between smartphone and colposcopic findings in patients with abnormal cervical cytology: experiences in a tertiary referral hospital, *Am. J. Obstet. Gynecol.*, 2020, **221**(3), 241.e1–241.e6.
13. YOUNESZADE, N., M. MARJANI, C.P. PEI, Deep learning in cervical cancer diagnosis: architecture, opportunities, and open research challenges, *IEEE Access*, 2023, **11**, 6133–6149.
14. ZHANG, T., Y.M. LUO, Cervical precancerous lesions classification using pre-trained densely connected convolutional networks with colposcopy images, *Biomed. Signal Process. Control*, 2020, **55**, 101566. https://doi.org/10.1016/j.bspc.2019.101566
15. ZHANG, X.Q., S.G. ZHAO, Cervical image classification based on image segmentation preprocessing and a CapsNet network model, *Int. J. Imaging Syst. Technol.*, 2019, **29**(1), 19–28. https://doi.org/10.1002/ima.22291.