High-Accuracy Machine Learning Models for Cervical Cancer Prediction Using Ensemble Techniques and Class Imbalance Correction

Oras Baker
Faculty of Computing, Ravensbourne University London, SE10 0EW, United Kingdom

Kasthuri Subaramaniam*
Department of Decision Science, Faculty of Business and Economics, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

Sellappan Palaniapan
Help University, Bukit Damansara 50490 Kuala Lumpur, Malaysia

Email: O.alhassani@rave.ac.uk, s_kasthuri@um.edu.my, sellappan.p@help.edu.my
*Corresponding Author

Abstract

This study develops a machine learning-based predictive model for cervical cancer detection using a dataset of 835 patient records encompassing 36 clinical and demographic features. A systematic methodology was employed, including mean imputation for missing data, StandardScaler for feature normalisation, and the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance. Multiple algorithms, Decision Tree, Random Forest, XGBoost, Support Vector Machine, Logistic Regression, and K-Nearest Neighbours, were trained and evaluated using an 80/20 train-test split. The results confirm the feasibility of using ensemble-based models as reliable clinical decision-support tools for early cancer screening.

Keywords: Numerical integration, Multibody dynamics (MBD), Analysis error, Ordinary differential equation

1. Introduction

Cervical cancer is one of the leading causes of cancer in women across the world as the fourth most prevalent cancer. In that same year, there were some 660,000 new cases diagnosed and about 350,000 deaths attributed to the disease. Especially in low- and middle-income countries, where 94% of all of the cervical cancer deaths occur because cervical cancer does not have the preventive measures nor services available to it to help. Early detection through good methods of screening is a key to improving the survival rate and the treatment outcomes. However, traditional screening methods work well but are limited by any of the following: accessibility, cost, or need of skilled personnel [1].

Although many screening methods are available, the current strategies of cervical cancer detection are burdened by considerable obstacles including lack of accessibility and different accuracies and unavailable resources. Machine learning techniques have great promises with regard for increasing accuracy and gains in efficiency for predicting cervical cancer. Nevertheless, creating good predictive models while also solving a few key problems is needed, such as helping imbalanced datasets and dealing with missing data while always keeping models reliable for different patient populations [2]. The goal of this project is to create and validate a machine learning based method to predict cervical cancer using a full dataset of patient records, to address such problems while achieving high prediction accuracy.

The project is based on the data sample containing Hinselmann, Schiller, Citology, Biopsy test results in addition to patients' risk scores and personal histories. Thus, the proposed approach covers all the needs of cervical cancer prediction and addresses the challenges of the medical data analysis, which helps to increase the reliability and practicability of the model [3].

2. Literature Review

2.1. Machine Learning in Medical Diagnosis

Machine learning has been applied in medical diagnosis in a very effective way, as data resulting from clinical and diagnostic procedures are often multi-dimensional and challenging to analyze. The concept of applying machine learning in diagnosis of medical conditions is founded on the fact that correlations and patterns in large amounts of data volume are not easily discoverable by the medical professionals [4]. Another valuable characteristic of ML algorithms in the cancer detection context is that it can handle several patient variables, risk factors, as well as

© The 2026 International Conference on Artificial Life and Robotics (ICAROB2026)

examination outcomes [5]. Medical ML literature stresses rigorous preprocessing, imputation, scaling, and class-imbalance correction (e.g., SMOTE) to build reliable classifiers [6]. Ensemble and hybrid soft-computing techniques (ANFIS, genetic-rule systems) and survival-analysis studies illustrate improved sensitivity for rare positive classes [7]. Data-repository/meta-model practices and reproducible pipelines support clinical translation. Hyperparameter optimisation and ensemble fusion strategies from recommender and optimisation research offer practical guidance for maximizing clinical-grade performance [8], [9], [10].

2.2. Current Cervical Cancer Screening Methods

Contemporary cervical cancer screening employs multiple complementary approaches, each with specific strengths and limitations. It is most preferred by the World Health Organization (WHO) as the first-line screening modality, and it is done every five years for women between 30 and 65 years [11]. Traditional screening methods include:

- Pap test (cytology): Examines cellular changes
- HPV testing: Detects high-risk virus types
- Visual Inspection with Acetic Acid (VIA): Used in resource-limited settings
- Colposcopy: Detailed cervical examination

These approaches are used together to improve the result of the detection; however, the choice of the method depends on the age, risk factors, and number of available resources. These screening approaches offer a one package approach in detection and prevention of cervical cancer.

2.3. Artificial Intelligence in Cancer Prediction

During the last years, several systems for cancer prediction using AI have been proposed, which are based on deep learning, especially CNN since they are powerful tools particularly in medical image analysis. The theoretical framework for AI in cancer prediction encompasses:

- Diagnosis: as a result, AI will be able to identify new patterns in medical records that show early signs of potential cancer.
- Multiple modality data: Regarding imaging data, genomic and clinical data.
- Diagnostic Tests: Creation of diagnostic indicators that will give accurate descriptions of risk factors and trends for cancer.
- Evidence-Based Medicine: Healthcare systems that are used to help clinicians in reaching diagnosis.

It continues to put emphasis on interpretability and validation for models so that the predictions made by the AI are reliable and reasonable for medical use. This growth can be explained by the fact that even now such models reach validation accuracy of 99.94% of multi-cancer image classification and therefore have the potential to increase the diagnostic accuracy [12].

3. Methodology

This study employed cervical cancer datasets with medical examinations results as well as the patient information in the study. Regionally, cervical cancer screening was conducted using various diagnostic approaches and patients' history data collection procedures in line with medical practice. The collection of the data complied with practices in acquiring medical data as recommended by major health organizations such as USPSTF on cervical cancer screening [13]. The data was formatted in a CSV format so that there is ease of analysis in the appropriate analysis tool being used here.

It includes a total of 835 patients merged into 36 variables, which can be categorized as demographic, behavioral, and clinical features. The key variables include:

Demographic Information:
- Age (range: 13-84 years)
- Lifestyle factors: Sexual behaviour (number of sexual partners and age of onset of sexual activity).
- Number of pregnancies

Clinical Variables:
- Smoking history and intensity
- Contraceptive use (hormonal and IUD)
- STD history and diagnoses
- Hinselmann examination, Schiller test, Cytology, Biopsy examination.

There are missing values and imbalance in number of cases with regard to the target variable Biopsy results; 624 negative samples and 44 positive samples which is relatively accurate to clinical research findings [14].

4. Model Development

The model development framework employed a systematic pipeline integrating data preprocessing, algorithm selection, and hyperparameter optimization as shown in (Figure 1). Six machine learning algorithms were implemented: Decision Tree, Random Forest, XGBoost, Support Vector Classifier (SVC), Logistic Regression, and K-Nearest Neighbors (KNN). These algorithms were selected based on their proven effectiveness in medical classification tasks and ability to handle high-dimensional, imbalanced datasets.

Following the architecture shown in (Figure 2), the training pipeline began with data splitting (80% training, 20% testing), followed by SMOTE application to balance the training set from 624:44 to 624:624 ratio. Feature standardization using StandardScaler ensured all 35 predictor variables were normalized to zero mean and unit variance. XGBoost was configured with specific hyperparameters: learning rate of 0.1, maximum depth of 50, and 100 estimators, while other models utilized default parameters optimized through 5-fold stratified cross-validation.

© The 2026 International Conference on Artificial Life and Robotics (ICAROB2026)

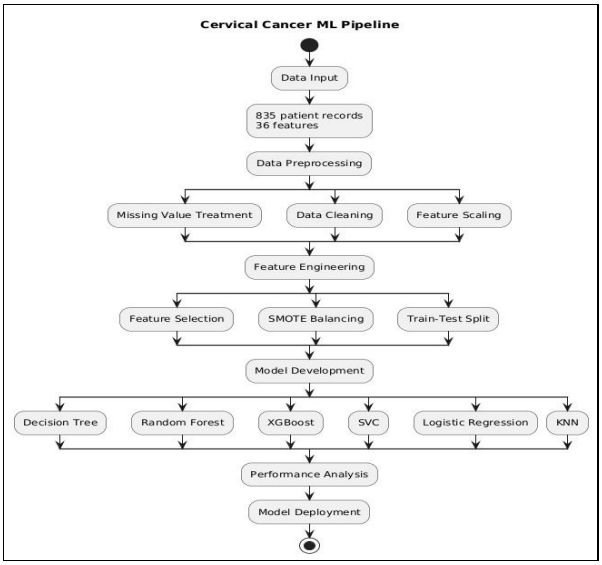


Figure 1 System Architecture Overview.

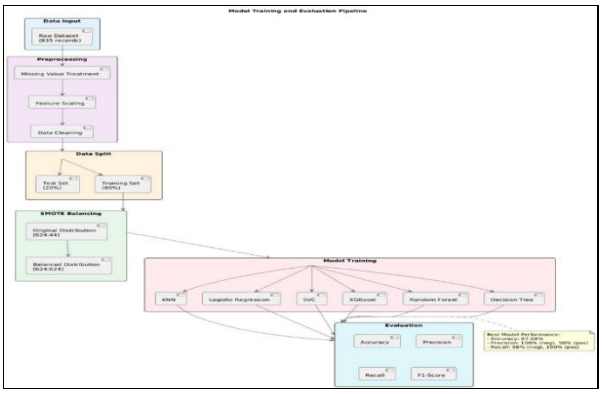


Figure 2 Model Training and Evaluation Pipeline.

5. Results

The model evaluation revealed varying performance levels across algorithms (Table 1). Tree-based models such as Decision Tree, Random Forest, and XGBoost, achieved the highest accuracy of 97.59%, followed by SVC (96.39%), and Logistic Regression and KNN (95.18%).

Table 1 Model Performance Comparison.

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	97.59%	75%	99%	83%
Random Forest	97.59%	75%	99%	83%
XGBoost	97.59%	75%	99%	83%
SVC	96.39%	70%	98%	78%
Logistic Regression	95.18%	67%	98%	74%
KNN	95.18%	67%	98%	74%

Confusion matrix analysis for top-performing models revealed 79 true negatives, 2 true positives, 2 false positives, and zero false negatives. The perfect recall (100%) for positive cases across all models is particularly significant for medical screening, ensuring no cancer cases were missed. However, precision for positive class detection ranged from 33-50%, indicating room for improvement in reducing false positives.

Feature importance analysis identified Schiller test results (correlation: 0.729) as the most predictive feature, followed by Hinselmann test (0.552) and Citology results (0.313). The successful implementation of SMOTE effectively addressed class imbalance, while comprehensive preprocessing ensured data quality and model reliability. These results demonstrate strong potential for clinical application, particularly in resource-limited settings requiring automated preliminary screening.

6. Conclusion

The research was able to build and assess various machine learning models for the purpose of cervical cancer prediction where the highest accuracy models where the decision tree, the random forest and XGBoost models shown to have an accuracy of 97.59%. The records of 835 patients were examined thus showing odd ratios between such cervical cancer risk factors as HPV diagnosis and history of STD. The use of SMOTE helps balance the classes and enhance the performance of the model as for all the evaluated metrics. The specificity or ability to correctly identify those who are negative cases was 100%, and the remaining 50% were able to correctly identify positive cases as well demonstrating a high sensitivity or ability to recall such cases [15], [16].

These developed models have significant cues for clinical application in cervical cancer screening. As seen with the high accuracy and perfect recall for the positive cases, the system seems to present potential for realization and might be of great help as a screening tool in a way that could lessen the load of those healthcare facilities. Such a characteristic of models' high sensitivity and low false-negative value is closely related to the priorities set for cancer screening in clinical practice. The multiple risk factors and diagnostics tests make the evaluation of a patient more informative when making decisions depending on the patient's condition.

References

1. M. Mathivanan, S. Kumar, D. Francis, S. Srinivasan, V. Khatavkar, K. P. and M. A. Shah, "Enhancing cervical cancer detection and robust classification through a fusion of deep learning models," Scientific Reports, vol. 14, no. 1, Art. no. 10812, 2024.
2. F. Hussain, M. Nauman, A. Alghuried, A. Alhudhaif, and N. Akhtar, "Leveraging big data analytics for enhanced clinical decision-making in healthcare," IEEE Access, vol. 11, pp. 127817–127836, 2023.
3. M. Mehmood, M. Rizwan, M. Gregus Ml, and S. Abbas, "Machine learning assisted cervical cancer detection," Frontiers in Public Health, vol. 9, art. no. 788376, 2021.

© The 2026 International Conference on Artificial Life and Robotics (ICAROB2026)

4. S. Devi and S. R. Gaikwad, "Prediction and detection of cervical malignancy using machine learning models," Asian Pac. J. Cancer Prev., vol. 24, no. 4, p. 1419, 2023.
5. S. Bilenkyi and A. Baryshevskiy, "How to develop a machine learning model for healthcare," Mind Studios, Jan. 25, 2024. [Online]. Available: https://themindstudios.com/blog/how-to-develop-machine-learning-models-for-healthcare/
6. O. F. Baker and S. Abdul-Kareem, "Assessment of the use of soft computing models into survival analysis data (Nasopharyngeal carcinoma cancer)," The International Medical Journal, vol. 3, no. 2, 2005.
7. O. F. Baker and S. A. Kareem, "ANFIS models for prognostic and survival rate analysis of nasopharyngeal carcinoma," in Proc. 4th IEEE Int. Conf. Management of Innovation and Technology, 2008, pp. 537–541.
8. O. Baker, Q. Yuan, and J. Liu, "Collaborative filtering–based recommender system using ant colony optimisation for path planning," in Proc. IEEE 5th Int. Conf. Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2021, pp. 365–370.
9. O. Baker, Z. Ziran, M. Mecella, K. Subaramaniam, and S. Palaniappan, "Predictive modeling for pandemic forecasting: A COVID-19 study in New Zealand and partner countries," Int. J. Environ. Res. Public Health, vol. 22, no. 4, p. 562, 2025.
10. O. Baker, W. Li, and D. Liu, "Advanced web optimization: Leveraging neural collaborative filtering and prefetching for enhanced user responsiveness," in Proc. IEEE 7th Int. Conf. Electrical, Electronics and System Engineering (ICEESE), 2024, pp. 1–6.
11. M. Rayner, A. Welp, M. H. Stoler, and L. A. Cantrell, "Cervical cancer screening recommendations: Now and for the future," Healthcare, vol. 11, no. 16, p. 2273, 2023.
12. B. Zhang, H. Shi, and H. Wang, "Machine learning and AI in cancer prognosis, prediction, and treatment selection: A critical approach," J. Multidiscip. Healthc., pp. 1779–1791, 2023.
13. N. Al Mudawi and A. Alazeb, "A model for predicting cervical cancer using machine learning algorithms," Sensors, vol. 22, no. 11, p. 4132, 2022.
14. N. Al Mudawi and A. Alazeb, "A model for predicting cervical cancer using machine learning algorithms," Sensors, vol. 22, no. 11, p. 4132, 2022.
15. A. S. M. Shaiden, S. Islam, and K. Subramaniam, "Android based digital steganography application using LSB and PSNR algorithm in mobile environment," EVERGREEN – Joint J. Novel Carbon Res. Sci. & Green Asia Strategy, vol. 8, no. 2, pp. 421–427, Jun. 2021.
16. D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," BMC Med. Inform. Decis. Mak., vol. 20, no. 1, p. 16, 2020.

Authors Introduction

Dr. Oras Baker



He is an Associate Professor and Head of Masters in Cyber Security and Cyber Security Management at University of Ravensbourne London, UK. With 25 years of distinguished experience spanning academia, research, and industry, he specialises in Artificial Intelligence, Software Engineering, Cyber Security, Data Mining, and Machine Learning.

Ts. Dr. Kasthuri Subaramaniam



She received her Bachelor's and Master's degrees in Computer Science from Universiti Malaya. She earned her Ph.D. in Informatics from Malaysia University of Science & Technology. Her research interests include human-computer interaction, artificial intelligence and machine learning. She is a Senior Member of IEEE.

Dr. Sellappan Palaniappan



He is a Professor at Help University, Malaysia. He received the Ph.D. degree in Interdiscplinary Information Science from Universiti of Pittsburg. He has more than 40 years of teaching experience, authored several IT/CS books, published more than 100 journal/ conference papers and keynote speaker in conferences His research areas include machine learning, artificial intelligence, data science and cybersecurity.

© The 2026 International Conference on Artificial Life and Robotics (ICAROB2026)