International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I5P109

Volume 74 Issue 5, 138-147, May 2026

Original Article

Classification of Cervical Carcinoma: Employing Traditional Tree-Based Machine Learning Methods Utilizing Feature Selection Algorithms

Proloy Kumar Mondal[1], Haewon Byeon[2]*

[1]*Department of Digital Anti-aging Healthcare, Inje University, Gimhae 50834, South Korea.*
[2]*Worker's Care & Digital Health Lab, Department of Future Technology, Korea University of Technology and Education (KOREA TECH), Cheonan 31253, South Korea.*

[*]*Corresponding Author : bhwpuma@naver.com*

Received: 11 February 2025 Revised: 22 September 2025 Accepted: 28 March 2026 Published: 30 May 2026

Abstract *- Set against the fact that Cervical Cancer (CC) is the second-highest cancer amongst women globally, the Pap smear is one of the most widely screened test systems today. The number of cases in Bangladesh is rising fast, an urgent problem. To address this problem, an increasing number of individuals and researchers are turning to Machine Learning (ML) and Deep Learning (DL) to process large amounts of data and produce insights that can be implemented. ML-powered methods for the early-stage prediction of critical illnesses such as cancer, kidney failure, and heart diseases are standard in healthcare today. The early detection of cervical cancer is a potential option for the prevention of this disease, which is a prevalent condition in females. We addressed this performance in a wrapper model with Metaheuristic Algorithms: Osprey Optimization Algorithm (OOA). Machine learning classifier for CC prediction in this paper, we used AdaBoost as the machine learning classifier for CC prediction. By keeping all the 36 features, we are getting with AdaBoost an accuracy of 96%, a precision of 97%, a recall of 98%, and an F1 score of around 98%. Also, we utilized the OOA algorithm to find a beneficial subset of features to assist in the important feature selection. An OOA method reduced the features from 36 to 7. These results demonstrate the possibility of early cervical cancer detection using a reduced feature set with retained prediction power. Critically, we produced an enumeration of essential features for cervical cancer derived from our optimization algorithm and addressed CPU time for cost-effectiveness.*

Keywords *- Cancer of the Cervical Region, Machine Learning, Metaheuristic Algorithm, AdaBoost Algorithm, Osprey Optimization Algorithm (OOA).*

1. Introduction

Vaginal cancer is a very rare cancer found almost exclusively in women, and among various types of them, Cervical Cancer(CC) is comprehensively risky for women, which greatly causes death every year in developing countries like Bangladesh, India, and Pakistan [1]. A survey of the American Cancer Society reports that approximately 13,820 women were newly diagnosed due to CC [2] throughout the United States in 2024. Moreover, almost 4,360 women died because of CC Human Papillomavirus (HPV), which is responsible for almost all CC, which is caused by an infection that persists within the body, and it can lead to cell changes in the cervix. Ultimately, these changes can develop into a precancerous lesion, and it turned into CC. Despite the effectiveness of the HPV vaccine, women are not aware of these issues, and in some cases, employed organization are gradually withdrawing their activity from countries like ours [3-6]. Now the question comes to healing, whether it is possible to identify the CC at an early stage, and hence, a series of patient data on risk factors of clinical attributes will lead to our study's further betterment process to detect this CC. As some clinical resources like biopsy, cervical Pap smear, Schiller test, and cytological screening, including involvement of the patient, are highly expensive and not affordable, this has eventually increased the death rate in many countries. Very few countries are capable of providing this health service. Therefore, by taking into consideration all limitations such as lack of proper awareness, lack of participation among the women ,greater cost of diagnosis ,test and screening, in most cases it detected late when radio therapy is more acute, perilous and less survival for patients. This problem can be overcome using Machine Learning (ML) analysis of patient data on risk parameters of clinical attributes .Hence our aim is to construct a proposed model where not only the low cost detection will be preferred ,but also the system efficiency and accuracy will be highly counted on several parameter bases. Figure 1 represents the cervical cancer symptoms.

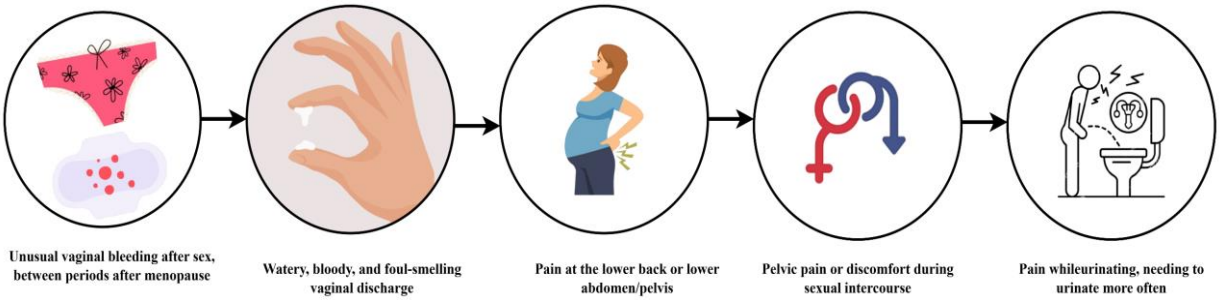

Fig. 1 Cervical cancer symptoms are shown

Moreover, an innovative way to develop a machine learning model may hold the key to viable interventions against cervical cancer and hope for a brighter future for girls and women [7]. The patients' risk factors and patient medical history based on initial screening, this study proposed a machine learning model for estimating the risk of cancer advancement among patients. In this paper, we propose a prediction model in order to make the machine learning methods easy to use for cervical cancer screening, which can bring benefits to clinical application personnel. Blood cancer, breast cancer, and prostate cancer are among other cancers that could also be diagnosed with the test. The key goal of this analysis is to classify the result of a biopsy to detect cancer in the cervix. In evaluating a model's performance, we favor sensitivity above accuracy in order to ensure that as many of the cancer patients as possible are correctly classified, so as to reduce the likelihood of a positive test reflecting an actual positive. Our contribution to this study:

- Proposed an effective method for predicting severe CC cases and achieving 96% accuracy.
- Developed a robust feature selection approach that removed less important features and extracted an optimal feature set to enhance classification performance.
- We presented the potential to reduce prediction inconsistencies.

The rest of the paper is structured as follows: in Section. II, related work is discussed, and we highlight the key differences between our work and existing literature. In Section III, the methodology approach, test details, configuration, and system flowchart are provided. The results are presented in Section IV, while a discussion is given in Section V.

2. Related Work Analysis

In this section, we will describe some algorithms for predicting cervical cancer that have been suggested in recent years, which can be categorized as machine learning methods.

Jiai Lu et al. [7] introduced an original ensemble method, which can bypass the restrictions of the conventional voting-based analysis, and enhance the predictive accuracy by data correction. They also considered the possibility of gene boosting modules.

Their approach by voting resulted in a promising performance for the prediction of cervical cancer risk. In addition, Nithya et al [8] also attempted to analyze with ML in building cervical cancer risk factors. They applied several feature selection approaches to pinpoint the features necessary to accurately predict and established an effective feature selection model via iteration of model training. Akhtar et al. [9] applied three ML models, including DT, RF, and XGBOST, to predict cervical cancer risk with behavioral and trait-related data.

Their analysis has surpassed the performance of other methods with an accuracy 93.33%. Asadi et al. [10] reconfirmed the important role of ML, especially with the DT algorithm, in accurate cervical cancer prediction. Suman et al. [11] showed the power of the ML-based method in identifying cancer samples quickly by utilizing high-resolution biopsy data to speed up and refine the process of decision making in a Bayes Net method and achieved a remarkable AUC of 95% and an accuracy of 96.38% for the trained model.

Supervised ML early prediction of cervical cancer symptoms in [12], used the supervised ML method to predict early cervical cancer symptoms from a dataset from the UCI Machine Learning Repository. Performance characteristics such as precision-recall curve and F1-score, precision based predictive model and ROC-AUC, precision-recall trade-offs were comprehensively evaluated in their paper that led to early prediction of risk that could have clinical relevance.

A decision tree classifier for risk factor analysis of cervical cancer was proposed. They used RFE, Lasso and some other feature selection models to keep the features which are relevant in the cervical cancer detection.

Mehmood et al. [13] outlined a machine learning-supported cervical cancer detection model via Particle Swarm Optimization (PSO) for feature selection. They employed a real-world dataset from the UCI Machine Learning Repository that included 36 known risk factors such as demographics, habits, sexual behavior, gynecological history, and HPV

status. Algorithm PSO Used for Selection of Minimum No of Features, ML-Based Four Models: LR, SVM, RF, and ANN were developed to do Prediction for Cervical Cancer. To evaluate the performance of these models, we utilized several metrics such as accuracy, precision, recall, and F1-score compared to other approaches. Their model has the best results with 93.6% of accuracy, MSE of 0.07111, false positive rate = 6.4%, and false negative rate = 10%. Alsalatie et al. [14] also found the dataset's source in analyzing cervical cancer data from Kaggle.

The patients were classified into two groups: healthy and with cervical cancer using three Machine learning models, i.e., Decision Tree (DT), K-Nearest Neighbor (KNN), and NAIVE BAYES (NB). To assess the performance of their models, they employed 10-fold cross-validation implemented in their dataset, and the metric used was accuracy. The results in [14] show that DT has an accuracy of 97%, KNN is 95%, and NB is 93%.

3. Proposed Methodology

The complete workflow diagram of the proposed method is depicted in Figure 2. The input data is first transformed and pre-processed by filling in the missing data and encoding the data using encoding, so that the categorical features can be expressed as numerical values.

Some feature selection methods are used for the purpose of selecting relevant features. It is demonstrated that resampling is used for dataset refinement. Next, we partitioned the data into a training set (70%) and a test set (30%). We trained and tested several machine learning models for model-training data and then applied the AdaBoost classifier. The confusion matrix and the ROC curve are used to assess the model. And then we use the metaheuristic algorithm, and finally we compare the two results and choose the better model.

3.1. Dataset Description

This data was collected from the University de Caracas Hospital, Caracas, Venezuela. Feature selection: We used the risk factor dataset of the UCI Machine Learning repository [15], which is a publicly available dataset of 36 features with 858 patients.

The data set consists of age, number of sex partners, age at first sex, number of pregnancies, anyone who smokes, how many years did they smoke, and the number of packs of cigarettes smoked in how many years, the use of an IUD has ever been, whether or not someone did take the vaccine or did not, and their cine tests results, and whether someone has had DxCancer, DxCIN, DxHPV, Dx and the have Hinselmann test. If available, the Schiller test, the Cytology result, and the biopsy. However, Hinselmann, Schiller, cytology, and biopsy were considered to be target features. Biopsy was the primary study endpoint. In this dataset, the number of non-cancer and cancer patients based on biopsy is 803 and 55, respectively. The dataset attributes are presented in Table 1.

Table 1. Description of the dataset

S. No	Feature Description	Data Type
1	Age of the patient	Integer
2	Total number of sexual partners	Integer
3	Age at first sexual activity	Integer
4	Number of pregnancies	Integer
5	Smoking habit	Boolean
6	Duration of smoking (years)	Integer
7	Smoking intensity (packs/year)	Integer
8	Use of hormonal contraceptives	Boolean
9	Duration of hormonal contraceptives use (years)	Integer
10	Use of an Intrauterine Device (IUD)	Boolean
11	Duration of IUD use (years)	Integer
12	History of any STD	Boolean
13	Number of STD infections	Integer
14	History of condylomatosis	Boolean
15	Cervical condylomatosis history	Boolean
16	Vaginal condylomatosis history	Boolean
17	Vulvo-perineal condylomatosis history	Boolean
18	History of syphilis	Boolean
19	History of pelvic inflammatory disease	Boolean
20	History of genital herpes	Boolean
21	History of molluscum contagiosum	Boolean
22	History of AIDS	Boolean
23	History of HIV	Boolean
24	History of Hepatitis B	Boolean
25	History of HPV	Boolean
26	Total number of STD diagnoses	Integer
27	Time since first STD diagnosis (years)	Integer
28	Time since last STD diagnosis (years)	Integer
29	Diagnosis: cancer	Boolean
30	Diagnosis: CIN	Boolean
31	Diagnosis: HPV infection	Boolean
32	Overall diagnosis	Boolean
33	Hinselmann test result	Boolean
34	Schiller test result	Boolean
35	Cytology test result	Boolean
36	Biopsy result (target)	Boolean

3.2. Dataset Preprocessing

These datasets require preprocessing before using machine learning models to achieve high classification accuracy [15]. Different preprocessing methods are applied to address missing values, outliers, label encoding, etc. This dataset has missing values, which must be taken care of before

we can use the machine learning model. Medical datasets are characterized by high-dimensional features with mixed types of values, and the quality of these sets can be affected by noise, outliers, missing values, duplicate observations, and unrepresentative or biased samples. It is necessary to conduct preprocessing steps to enforce accurate analysis for better raw data. These are preprocessing operations that need to be performed in this case study. Cleaning removes the records with null values and outliers, which keeps data quality in check, thereby improving overall data precision.

Then, the numerical features are standardized, which sets them to a zero mean and unit variance, which allows machine learning models to perform better and converge faster.

To fill missing values, we impute the most frequent value by preserving mode & distribution, again capitalized as frequency. At the last stage, we do data balancing not to let models be schooled by one class all the time, and also for generalization boosting.

3.3. Classification with AdaBoost

AdaBoost, short for "adaptive boosting", is an ensemble learning technique developed to improve the accuracy of models using weak classifiers [16]. It works in an iteration-based manner, where weights are assigned to the training examples at each iteration, and more weight is given to examples that are misclassified in the previous stage. A new classifier is then trained on the weighted examples.

The final model is a weighted combination of all classifiers, which helps AdaBoost focus on difficult examples and classify them correctly in subsequent iterations. Mathematically, AdaBoost is expressed by Equation (1). Figure 3 represents the working principle of the AdaBoost classifier.

$$\mathbf{H(X) = sign(\sum_t \alpha_t h_t(x)} \quad (1)$$

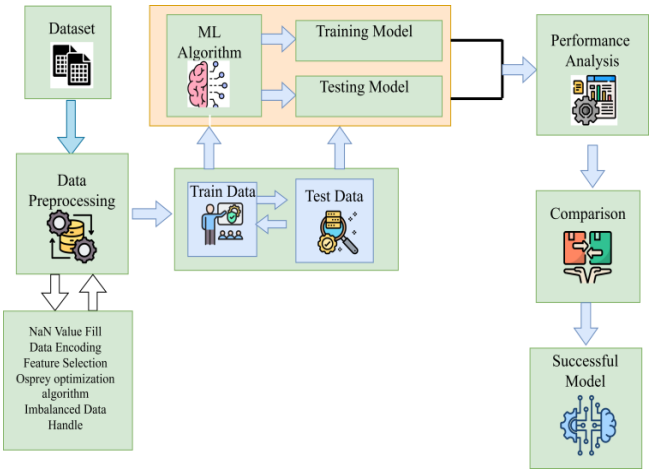

Fig. 2 The visualization of the proposed approach

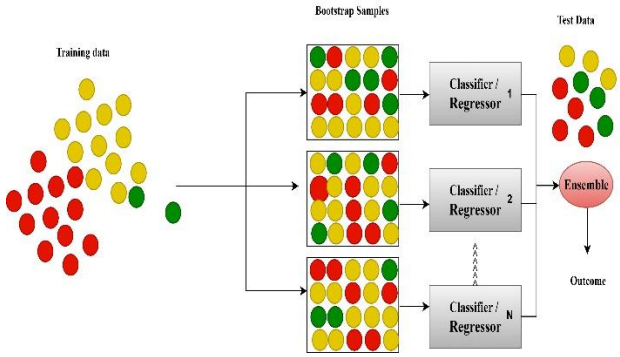

Fig. 3 Diagram of AdaBoost Classifier Algorithm [17].

3.4. Feature Selection Algorithm

The Feature selection methods are used to identify the most informative subset of features from a dataset based on certain characteristics established by a model [18]. Some features in datasets have varying importance to the classification task, and not all of them are equally helpful for model performance. Feature selection methodologies are classified into Filter, Wrapper, and Embedded methods. Metaheuristic Algorithms (MHAs): MHAs are classified into the Wrapper method, which can compensate for the shortcomings of deterministic methods due to its global search. MHAs are classified into four businesses: Evolutionary Algorithms, Swarm Intelligence-based strategies, Physics-inspired algorithms, and Human Behavior-based Processes. Swarm-based MHAs have received greater attention in recent years as they help solve complex, nonlinear, and high-dimensional optimization problems. MHAs perform well within a reasonable time cost and low computational cost to achieve near-optimal solutions compared to the conventional methods.

3.5. Osprey Optimization Algorithm

Osprey Optimization Algorithm (OOA) is a Bio-Inspired metaheuristic algorithm that imitates the hunting mechanism of the osprey. It is made for engineering optimization [19]. It takes inspiration from their hunting method and comes with two phases: exploration and exploitation. This is mathematically formulated by the hunting scenario, in which the OOA generates poor solutions (hunting), locates the best one (attack), and transports them (transport). The OOA algorithm is shown in Figure 4.

The Osprey Optimization Algorithm has shown remarkable performance compared to other metaheuristic algorithms. Compared to evolutionary methods such as Genetic Algorithms (GA), which often require time-consuming convergence in the case of large parameter settings and high-dimensional search cases, OOA provides a simpler hunting-based search strategy with considerably fewer parameters, leading to quicker and more effective convergence. When swarm intelligence algorithms such as the Particle Swarm Optimization (PSO) are to be used, they tend

to reach premature convergence in multimodal landscapes; this means that RFA does a lot of searching before it speeds up its exploitation phase, while OOA maintains longer suboptimal solutions, avoiding local optima. While traditional physics-based methods, such as Simulated Annealing (SA), avoid getting stuck in local optima at the cost of high computation time, OOA provides similar or better results at lower computational expense and higher flexibility.

Alternatively, one or more optimization methods like the Whale Optimization Algorithm (WOA) excel in global search while being less effective for exploitation. OOA naturally couples exploration and harvest to yield fast convergence while maintaining accuracy. Moreover, this is validated on 29 benchmark functions from the CEC 2017 and 22 real-world constrained optimization problems from the CEC 2011. The statistical test also showed the advantages of its stability and competitiveness over existing methods.

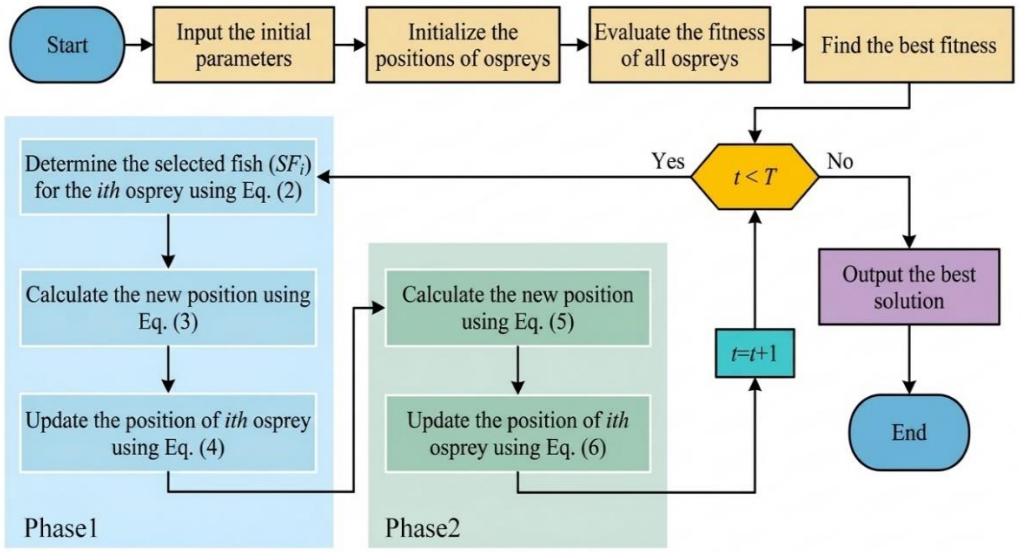

Fig. 4 The Flowchart of OOA

3.6. Model Evaluation

When we are not only measuring the correctness of a model. The matrix consisting of TP, TN, FP, and FN is calculated using the confusion matrix. The following is a table design for showing the model performance. A True Positive (TP) is a cancer patient who is correctly predicted to have cancer. A TN is an individual without cancer and whose forecast is that they would remain cancer-free.

False positive (FP) states the high-risk individual without cankering being identified as having cancer, and the False Negative (FN) is the reverse, that the malignant individual is evaluated as cancer-free. FN is the dominant feature and should be reduced as much as possible.

We consider these metrics to evaluate the performance of our model in this paper. Below is the definition of these metrics in terms of their naive representations.

- Accuracy: It is the proportion of accurate predictions made by the model, from the total predictions, and it is expressed with a certain formula.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

- Sensitivity: It is about the model's capability to correctly identify who has and doesn't have cervical cancer. 100% sensitivity was defined as the point at which the predictive model correctly assigns all cervical cancer cases and was determined for a specific equation.

$$precision = \frac{TP}{TP+FP} \quad (3)$$

- Specificity: It estimates how good our model is at recognizing those who will never get cervical cancer and is calculated with a certain formula.

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

- F-Measure: It is defined in terms of a particular formula, which is formed as a combination of the sensitivity and precision of the model, called the harmonic mean.

$$F1\ Score = 2 \times \frac{precision \times recall}{precision+recall} \quad (5)$$

4. Results and Discussion

4.1. Machine Learning Approach

The study used a methodology to evaluate the performance of the AdaBoost model for CC classification.

The dataset was split, and after training, the model was tested to determine its accuracy, precision, recall, and F-measure. The results provided insight into the consistency and overall effectiveness of the model. The performance metrics for the five iterations of cross-validation are listed in Table 2.

Table 2. Model evaluation of adaboost classifier algorithm

Class	Accuracy	Precision	Recall	F1-Score
Not CC	96%	97%	98%	98%
CC		67%	62%	65%

The table presents the performance metrics of a classification model differentiating between the two classes "Not CC" (Cervical Cancer Not) and "CC" (Cervical Cancer). For the “not CC” class, the model shows a high accuracy of 96%, indicating that it is 96% correct in predicting individuals without cervical cancer. The accuracy for this class is 97%, meaning that when the model identifies a patient as "NOT CC", it is correct 97% of the time, indicating strong reliability in its predictions.

On the other hand, the metrics for the "CC" class present some challenges: its accuracy is 67%, indicating that 67% of patients are detected as cervical cancer, resulting in a significant number of false positives. Moreover, the recall for the “CC” class is only 62%, which means that the model can correctly detect 62% of the actual cervical cancer cases, indicating the need for further improvement to detect this condition effectively. Overall, although the model performs well in identifying patients without cervical cancer, it faces problems in correctly classifying those with cervical cancer.

Figure 5 illustrates the performance of the AdaBoost classification model in distinguishing the two classes “Not CC” (Cervical Cancer Not) and “CC” (Cervical Cancer). This matrix is arranged in a 2x2 grid, where the rows show the actual labels (true labels) and the columns show the predicted labels.

In the upper left cell, the value 230 indicates the True Negative (TN), i.e., the model correctly identified 230 patients as cancer-free. A value of 5 in the upper right cell indicates a False Positive (FP), meaning that the model incorrectly identified 5 cancer-free patients as having cancer.

In the lower left cell, a value of 6 indicates a False Negative (FN), where the model misclassified 6 cancer patients as cancer-free. In the lower right cell, a value of 10 indicates a True Positive (TP), indicating that the model correctly identified 10 patients as having cancer. Overall, this confusion matrix provides valuable insight into the classification accuracy of the model, identifying its strengths and areas for improvement in cervical cancer detection.

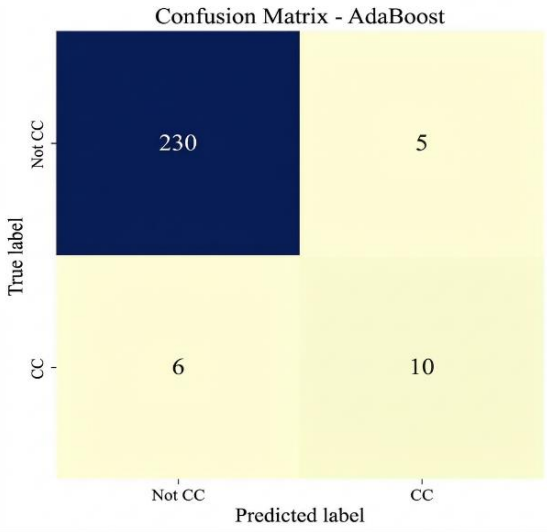

Fig. 5 Confusion matrix analysis of adaboost classifier algorithm

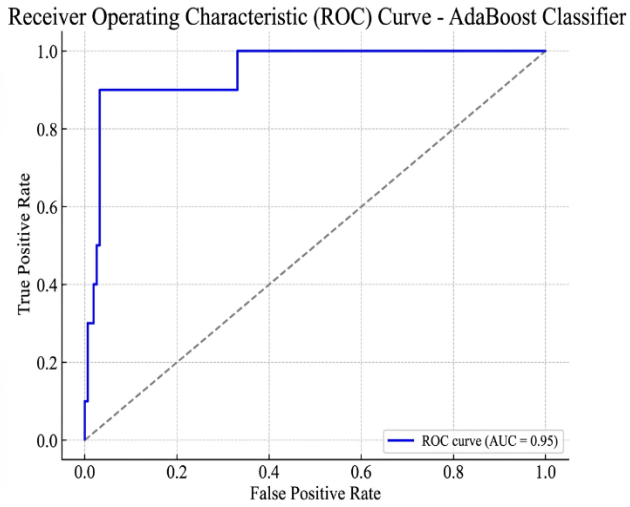

Fig. 6 ROC curve and AUC (AdaBoost Classifier)

The weighted ROC curve for an AdaBoost classifier is presented in Figure 6, providing insight into the discrimination of positive and negative cases of the model.

This ROC curve is based on plotting the true positive rate (sensitivity) on the y-axis and the false positive rate (1-specificity) on the x-axis for a range of threshold values. The blue line shows the performance of the classifier, and the diagonal dashed line represents the baseline (performance level of a random classifier).

The curve shows that the classifier can yield high sensitivity and a low false positive rate, demonstrating its good performance. The AUC is 0.95, meaning that the classifier distinguishes between the two classes.

A value of 1.0 means a perfect classification. The closer the ROC curve is to the combining point of the upper and left

coordinate axes, the better the predictive performance of the model. Particularly, we observed that the AdaBoost classifier did well with the AUC being close to 1, and it indicated the strong classification capacity in this setting.

4.2. Feature Selection Techniques OOA Approach

The initial step of our dataset analysis was to apply the AdaBoost algorithm on the entire dataset, which included a total of 35 features. Then, using the metaheuristic algorithm Osprey Optimization Algorithm (OOA) in the feature selection process, we were able to determine the 7 most effective features. These selected characteristics are: Age, first sexual intercourse, Num of pregnancies, Smokes (packs/year), Hormonal Contraceptives (years), Dx, and Schiller. These features played an important role in identifying our target features. For more clarity, these features are illustrated in Figure 7 below.

Table 3 summarizes the performance metrics of the OOA, a metaheuristic algorithm, after application to a classification task. 97% accuracy indicates that the OOA algorithm correctly classifies 97% of the cases, indicating its high predictability. 80% accuracy means that 80% of instances identified as positive were correct (likely to be cervical cancer), indicating the model's ability to reduce false positives. However, the 72.73% recall rate implies that the algorithm detected 72.73% of the true positive cases, i.e., missed some positive cases that have room for further improvement. The F1-score of 76.19%, which balances precision and recall, indicates that the algorithm handles both false positives and false negatives efficiently, although it is slightly biased towards precision.

The confusion matrix of the AdaBoost classifier on the test data set (Figure 8) provides information for the classification of the test samples based on the features selected by OOA. This matrix presents the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The large TP and TN cell values show that the model recognized both classes correctly, obtaining 97.01% accuracy. Meanwhile, small values for the FP and FN cells show that the model rarely confuses positive and negative classes, which demonstrates confident prediction of the model. This finding indicates that OOA is good at choosing useful features for the AdaBoost classifier, and the feature subset obtained by OOA is useful for improving the accuracy of the AdaBoost and decreasing the misclassifications.

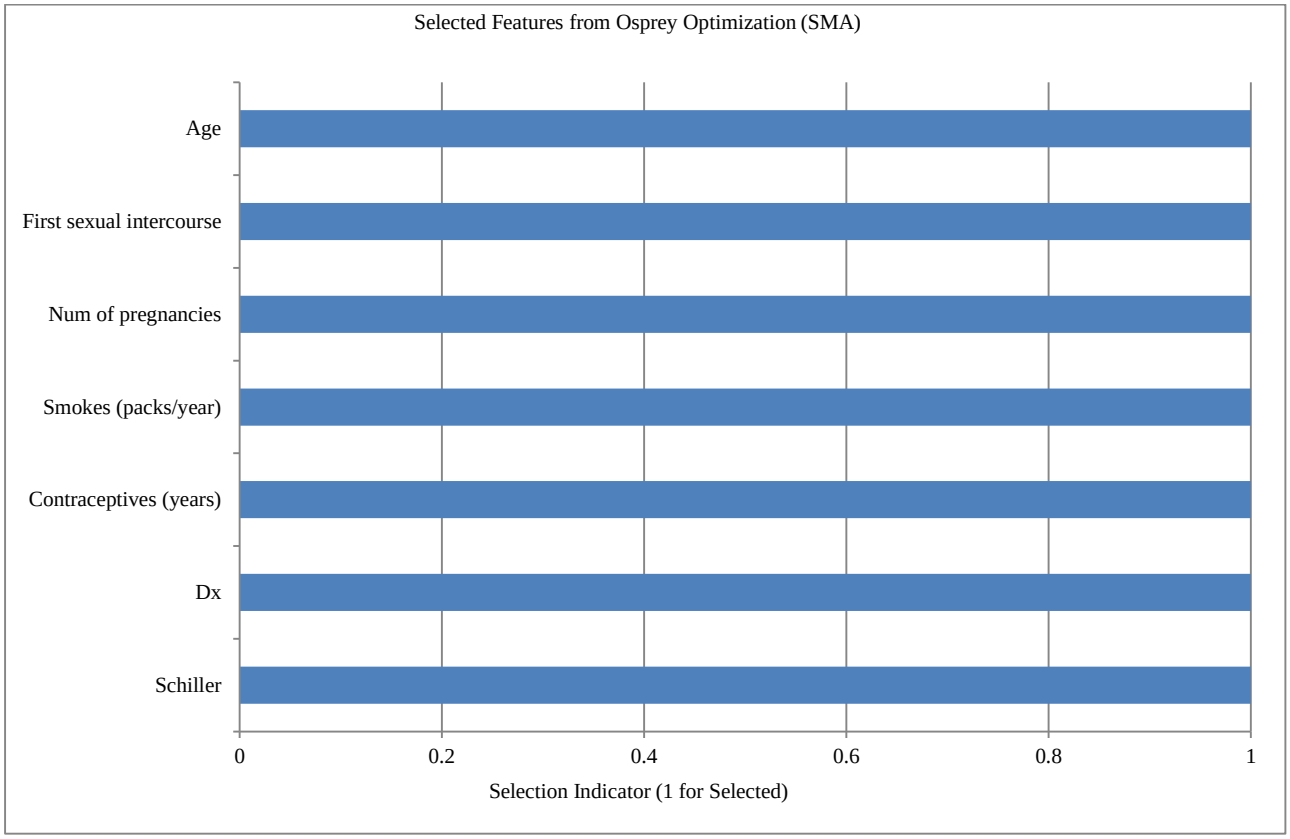

Fig. 7 Selected features from OOA

Table 3. Model evaluation of adaboost OOA

Metaheuristic Algorithm	Accuracy	Precision	Recall	F1-Score
OOA	97%	80%	72.73%	76.19%

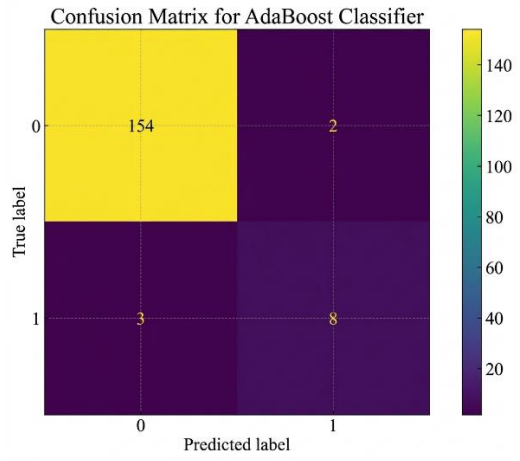

Fig. 8 Confusion matrix of OOA

5. Discussion

Table 4 represents the most significant contribution to improving the predictive performance of machine learning. (ML) based approaches for predicting cervical cancer risk can be attributed to current advances in ML techniques. Jiai Lu et al. Wang et al. [7] further proposed an ensemble method to overcome the limitations of traditional voting in terms of data correction and gene boosting modules. Their methodology showed the efficacy of ensemble learning for clinical diagnoses as an enhanced prediction. Similarly, Nithya et al. Using multiple packages, Reshamwala et al. accomplished the best feature selection and predictive variables among ML classifiers that can be used to develop an effective feature-preselection framework for improved classifier accuracy [8]. Akhtar et al. In [9], the Decision Tree (DT), Random Forest (RF), and XGBoost algorithms were used, together with behavioral and trait data, to predict cervical cancer risk with an accuracy of 93.33%. Asadi et al. Based on a follow-up study by Breiman [10], this domain knowledge suited the characteristics of DT models, particularly concerning their transparency and well-known predictive power. Suman et al. For example, using high-resolution biopsy data in [11] and a Bayes Net model yielded an AUC of 95% and an accuracy of 96.38%, thus showing how probabilistic graphical models can be very useful in clinical settings. Ratul et al. [12] proposed an ML approach based on a supervised method to predict early symptoms of cervical cancer using UCI repository data. In the case of having evaluated precision-recall trade-offs, F1-scores, and ROC-AUC as we did, these metrics reflect clinically relevant early prediction. Our study also showed that the results of the proposed models are quite attractive and, in some cases, even better than the other methods. The Model 01, based on the AdaBoost Classifier algorithm, showed impressive results with an accuracy above the 96% range, even outperforming several traditional ML methods reported in previous works. Importantly, our Model 02, based on the Osprey Optimization Algorithm (OOA) for identifying the seven most significant features, achieved the highest level of accuracy as much as 97%, which outperformed all benchmark studies that we reviewed here, including the Bayes Net model by Sujoy et al., and the DT classifier of Alsalatie et al. It suggests that the combination of powerful feature selection and ensemble learning algorithms can be helpful to improve the predictive performance for cervical cancer risk classification. Finally, this comparison highlights the efficacy of traditional ML techniques like DT, RF, and Bayes Net while serving as an example of integrating optimization-driven feature selection with modern classifiers for state-of-the-art performance.

Table 4. Comparison of the existing methods with our method

Authors	Methods/ Algorithms	Evaluation Metrics	Key Results
Jiai Lu et al. [7]	Original ensemble method	Accuracy	Improved predictive accuracy for cervical cancer risk
Nithya et al. [8]	Multiple ML models	-	Built an effective feature selection model for accurate prediction
Akhtar et al. [9]	Decision Tree (DT), Random Forest (RF), XGBoost	Accuracy	93.33%
Asadi et al. [10]	Decision Tree (DT)	Accuracy	Highlighted DT's strong performance for cervical cancer prediction
Suman et al. [11]	Bayes Net	AUC, Accuracy	AUC = 95%, Accuracy = 96.38%
Ratul et al. [12]	Supervised ML model	Precision, Recall, F1-score, ROC-AUC	Early prediction with clinical relevance
Our Model 01	AdaBoost Classifier Algorithm	Accuracy	96%.
Our Model 02	OOA (Significant 7 features)	Accuracy	97%,

6. Future Work

We propose high-accuracy models, which can be imminently useful in medical diagnostics, the pre- and post-treatment stage of Cervical Cancer (CC) management. This has several advantages in real-world clinical environments, as it reduces the number of input features. Fewer features mean fewer clinical tests, which not only reduces the cost to patients but also costs for healthcare providers. The lower the data requirement, the better the efficiency of disease detection, and so the quicker decision-making followed by timely commencement of treatment. Therefore, these study results could be used as a foundation for the development of

dedicated CC predictive software or specialized medical devices to assist in clinical workflows, as noted by the researchers. In the future, we intend to evaluate our proposed models on bigger datasets to confirm their performance and robustness, generalizability, scalability, and also increase the granularity of analysis by the addition of more MHAs for optimization.

7. Conclusion

Now cervical cancer is known as one of the most common female murderers. Source: University of Wisconsin "Cervical Cancer Fact Sheet." The World Health Organization (WHO) on cervical cancer outside the US says: More than 85% of all cervical cancer deaths occur in the developing world." We have used the machine learning models to find the factors that increase women's chances of having this respiratory disease. As the base classifier, we used an AdaBoost classifier as a machine learning algorithm, and an OOA-based swarm meta-heuristic algorithm for feature selection with the prediction of cervical cancer, by means of patient data on risk factors.

We observe that class balancing led to a significant improvement in prediction accuracy. In reference [15], classification can further benefit from random oversampling. Our proposed model has great potential in testing (the classification in testing is 96%). Further, we considered a feature selection method for feature reduction and identified seven predominant features that contribute to the detection of cancer. Our future work is focused on acquiring enhanced performance, which will be improved by gathering a large dataset. We will also study ensemble learning classifiers like boosting and bagging to extend the class balancing methods for further improvements and online screening.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF-RS-2023-00237287).

References

[1] Naif Al Mudawi, and Abdulwahab Alazeb, "A Model for Predicting Cervical Cancer using Machine Learning Algorithms," *Sensors*, vol. 22, no. 11, pp. 1-19, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Cervical Cancer, World Health Organization, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cervical-cancer

[3] Riham Alsmariy, Graham Healy, and Hoda Abdelhafez, "Predicting Cervical Cancer using Machine Learning Methods," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 7, pp. 173-184, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] Karl Ulrich Petry, "HPV and Cervical Cancer," *Scandinavian Journal of Clinical and Laboratory Investigation*, vol. 74, no. sup244, pp. 59-62, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[5] Guglielmo Ronco et al., "Efficacy of HPV-based Screening for Prevention of Invasive Cervical Cancer: Follow-Up of Four European Randomised Controlled Trials," *The Lancet*, vol. 383, no. 9916, pp. 524-532, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[6] Talha Mahboob Alam et al., "Cervical Cancer Prediction Through Different Screening Methods using Data Mining," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 2, pp. 388-396, 2019. [Google Scholar] [Publisher Link]

[7] Jiayi Lu, "Machine Learning for Assisting Cervical Cancer Diagnosis: An Ensemble Approach," *Future Generation Computer Systems*, vol. 106, pp. 199-205, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[8] B. Nithya, and V. Ilango, "Evaluation of Machine Learning based Optimized Feature Selection Approaches and Classification Methods for Cervical Cancer Prediction," *SN Applied Sciences*, vol. 1, no. 6, pp. 1-16, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[9] Laboni Akter et al., "Prediction of Cervical Cancer from Behavior Risk using Machine Learning Techniques," *SN Computer Science*, vol. 2, no. 3, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] F. Asadi, C. Salehnasab, and L. Ajori, "Supervised Algorithms of Machine Learning for the Prediction of Cervical Cancer," *Journal of Biomedical Physics and Engineering*, vol. 10, no. 4, pp. 513-522, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[11] Sujay Kumar Suman, and Nishtha Hooda, "Predicting Risk of Cervical Cancer: A Case Study of Machine Learning," *Journal of Statistics and Management Systems*, vol. 22, no. 4, pp. 689-696, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[12] Ishrak Jahan Ratul et al., "Early Risk Prediction of Cervical Cancer: A Machine Learning Approach," *2022 19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Prachuap Khiri Khan, Thailand, pp. 1-4, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] Mavra Mehmood et al., "Machine Learning Assisted Cervical Cancer Detection," *Frontiers in Public Health*, vol. 9, pp. 1-14, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[14] Mohammed Alsalatie et al., "A New Weighted Deep Learning Feature using Particle Swarm and Ant Lion Optimization for Cervical Cancer Diagnosis on Pap Smear Images," *Diagnostics*, vol. 13, no. 17, pp. 1-19, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[15] Kelwin Fernandes, Jaime S. Cardoso, and Jessica C. Fernandes, Cervical Cancer (Risk Factors), UCI Machine Learning Repository, 2017. [Online]. Available: https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors

[16] Ruihu Wang, "Adaboost for Feature Selection, Classification and its Relation with SVM, a Review," *Physics Procedia*, vol. 25, pp. 800-807, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[17] Sancho Salcedo-Sanz et al., "Analysis, Characterization, Prediction and Attribution of Extreme Atmospheric Events with Machine Learning: A Review," *arXiv preprint*, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Huan Liu, and Lei Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491-502, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[19] Mohammad Dehghani, and Pavel Trojovský, "Osprey Optimization Algorithm: A New Bio-Inspired Metaheuristic Algorithm for Solving Engineering Optimization Problems," *Frontiers in Mechanical Engineering*, vol. 8, pp. 1-43, 2023. [CrossRef] [Google Scholar] [Publisher Link]