BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Cervical cancer screening uptake and its associated factor in Sub-Sharan Africa: a machine learning approach

Fetlework Gubena Arage[1,2*], Zinabu Bekele Tadese[3], Eliyas Addisu Taye[4], Tigist Kifle Tsegaw[1,2], Tsegasilassie Gebremariam Abate[5] and Eyob Akalewold Alemu[1,2]

Abstract

Introduction Cervical cancer, which includes squamous cell carcinoma and adenocarcinoma, is a leading cause of cancer-related deaths globally, particularly in low- and middle-income countries (LMICs). It is preventable through early screening, but incidence and mortality rates are significantly higher in LMICs, with 94% of deaths occurring in these regions. Poor implementation of screening programs, in addition to multiple health system barriers, leads to a high burden from cervical cancer in these countries. Projections show increasing cases and deaths due to the disease by 2030. Using machine learning instead of the usual statistical tests will incorporate the complex and non-linear relationship of factors in predicting the outcome variable.

Method The secondary data for ten Sub-Saharan African countries were utilized from the Demographic and Health Survey, DHS, to evaluate cervical cancer screening uptake among women aged 25–49 years. During cleaning missing values and outliers were removed. Class balancing by Synthetic minority oversampling techniques (SMOT) was done and tuning hyperparameters via grid search was used in the models before splitting into training and validation sets containing 89% and 20%, respectively. The following machine learning classification algorithms were used in the study: Logistic Regression, Decision Tree Classifier, Random Forest, K-Nearest Neighbor, Gradient Boosting, AdaBoost, and Extra Trees. These algorithms were employed to predict cervical cancer screening uptake. The performance of the models was evaluated using accuracy, precision, recall, and F1 score.

Result In this study, a cervical cancer screening uptake was predicted among 75,360 weighted samples of women from an African country, aged 25–49 with the final data for model formulation of 53,461, where the Extra Trees Classifier obtained an accuracy of 94.13%, a precision of 95.76%, recall of 94.12%, F1-score of 93.80%. Then followed Random Forest: accuracy = 93.87, precision = 99.18%. Health visits, proximity to health care, using contraceptives, residing in urban settings, and exposure to media were its most crucial predictors. The ensemble methods, such as Extra Trees and Random Forest, showed the best generalization, indicating that this work well on complex datasets and can help devise targeted intervention strategies.

*Correspondence:
Fetlework Gubena Arage
fetle0123@gmail.com

Full list of author information is available at the end of the article

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Conclusion This study demonstrates that the ensemble machine learning models, such as Extra Trees Classifier and Random Forest, are promising in predicting cervical cancer screening uptake among African women with accuracies of 94.13% and 93.87%, respectively. Key predictors include healthcare access, sociocultural factors, media exposure, residence in urban areas, and contraceptive use. The findings emphasize the need for a reduction in care barriers and the use of family planning visits and mass media in promoting screening. These results will be validated in different populations in order to find the clinical integration via decision support systems.

Keywords Machine learning, Cervical cancer, Sub-Sharan, Screening, Ensemble method, Feature importance

Introduction

Cervical cancer is a type of cancer that develops in a women's cervix (the entrance to the uterus from the vagina). Depending on the site there are two major histologic subtypes of Cervical cancer, namely squamous cell carcinoma and adenocarcinoma. Squamous cell carcinoma arises from squamocolumnar junction. Whereas adenocarcinoma arises from the mucus secreting glandular cells of the endocervix. The cervix goes through a transformation before the emergence of cancer. The changes in the cervical cells more likely to progress into cancer are called precancerous lesion of the cervix [1–3].

Globally one woman dies every minute due to cervical cancer. Cervical cancer is the fourth most common cancer globally and second most common cancer in countries ranked low on human development index. More than 350,000 deaths and 660,000 new cases were reported in 2022. Although it is a preventable cancer, the annual number of new cases of cervical cancer has been projected to increase to 700,000 and number of deaths to 400,000 by 2030 [3, 4].

Cervical cancer incidence and mortality rates are four and three times higher in lower and middle-income countries (LMICs) than in high income countries respectively. Nearly 94% of the deaths occurred in low- and middle-income countries (LMICs). More than 85% of those affected are young women living in the poorest country. In the absence of further intervention (no screening), over the lifetime of a cohort of 100,000 unscreened women in 78 LMICs, 1,950 cervical cancer cases and 1,456 deaths are predicted to occur [3].

Despite being preventable, cervical cancer continues to pose a significant global health burden, particularly in low- and middle-income countries. The disparity in cervical cancer incidence and mortality rates between high-income and low-income countries is striking, with rates being four and three times higher, respectively, in the latter [5].

Most of the LMICs indicated a wide range of barriers to screening. What is urgently needed is the implementation of clear policies supported by health system capacity for implementation, community-wide advocacy, information dissemination, and strengthening of policies that support women's health and gender equality [6].

Traditionally, statistical methods such as logistic regression have been used to analyze predictors of cervical cancer screening uptake. However, machine learning (ML) offers several advantages that can enhance predictive accuracy, uncover complex relationships, and improve decision-making in public health interventions.

ML models can capture nonlinear relationships and complex interactions among predictors that traditional regression models may overlook. Algorithms such as random forests, gradient boosting, and neural networks can provide better classification accuracy for screening uptake prediction. Traditional statistical models struggle with high-dimensional datasets with numerous variables, whereas ML techniques can efficiently process and rank feature importance. ML allows the inclusion of diverse data sources, such as demographic, socioeconomic, behavioral, and healthcare access variables.

Method

This study analyzed cervical cancer screening uptake using recent Demographic and Health Survey (DHS) data from ten Sub-Saharan African (SSA) countries. These countries—Kenya, Tanzania, Mozambique, Madagascar, Gabon, Benin, Burkina Faso, Côte d'Ivoire, Ghana, and Mauritania—were selected based on the inclusion of cervical cancer screening assessments in their national DHS datasets. The study included two screening methods: Visual Inspection with Acetic Acid (VIA) and cytology.

DHS surveys employ a multistage cluster sampling design. Stratified cluster sampling was used to select clusters from urban and rural areas, followed by systematic sampling to select households. Women aged 25–49 years, the primary target group for screening in most of these countries, constituted the study population. Screening services are generally integrated into healthcare systems, offered as both free and paid services.

Data access was obtained through the DHS Program website (https://dhsprogram.com/data/) following ethical guidelines to ensure participant confidentiality. Geographic coordinates were randomly displaced (5 km in rural areas and 2 km in urban areas) to protect privacy, in line with DHS protocols (http://goo.gl/ny8T6X).

Women who were uncertain about whether they had undergone screening were excluded, as were those in clusters with missing geographic data. After

preprocessing, the final weighted sample consisted of 53,461 women from an initial 75,360.

This study adhered to ethical guidelines for the secondary use of DHS data, ensuring compliance with all data protection and confidentiality protocols.

Dependent variable

Cervical cancer screening uptake (yes/no).

Independent variables

Socio-demographic variables- age, educational status, marital status, parity, wealth index, sex of household head, occupation, media exposure, and Husband's employment.

Clinical and behavioral variables- age at first sex, cigarette smoking, modern contraception utilization, HIV test, sexual transmitted infection (STI) and multiple sexual partners.

Health service-related factors- health insurance, health care visit in last year and perception of distance to health facility as a big problem.

Operational definitions

Cervical cancer screening uptake- is categorized as "yes" or "no," after excluding women who didn't know if they were screened or not. Women were asked whether they had ever tested for cervical cancer by health professionals after explaining the screening procedures and methods to recall it. So, it assesses at least one cervical screening uptake of women by any one of the methods (Cytology or Visual inspection).

Exposure to mass media- is created by combining three variables from the survey: frequency of reading a newspaper or magazine, frequency of listening to radio, and frequency of watching television. For this study, those who had exposure to one of these media at least once a week are considered to have had mass media exposure.

Sexual transmitted infection- women who had any sexually transmitted infection (STI) in the last 12 months were considered to have had STI.

Early-age sexual initiation- women who had their first sexual intercourse before they attained 15 years were categorized as "yes".

Multiple sexual partner- women who had more than one sexual partner in their lifetime were categorized as "yes".

Data preprocessing

Data preprocessing is divided into two sections, which are data cleaning and data balancing. Data preprocessing is critical since it directly impacts project success. Data impurity occurs when attributes or attribute values contain noise or outliers, and redundant or missing data. We have removed variables with more than 20% missing value and impute for variables with less than that with mean imputation for continuous variable and mode imputation for categorical variables and outliers from this dataset. The data was balanced using synthetic minority oversampling technique (SMOTe). The data was split in to two sets namely training set and validation set with 0.2 split (80% training and 20% validation) after balancing the dataset. We used 80/20 split to provide larger dataset for the training.

Model Selection (MS)

Several machine learning classification algorithms have been used in this study, namely Random Forest, decision tree classifier (DTC), logistic regression (LR), gradient boosting (GB), adaptive boosting (AB), K-nearest neighbor (KNN), and Extra tree classifier. The different algorithms were compared using evaluations metrics like accuracy, recall, precision, and F-1 score.

Logistic regression

Logistic regression is a one of the machine learning classification algorithms for analyzing a dataset in which there are one or more independent variables (IVs) that determine an outcome and also categorical dependent variable (DV). logistic regression model uses more complex cost function (known as sigmoid function or logistic function) instead of linear function. Logistic regression limits the cost function between 0 and 1 [7].

$$\sigma\ (z) = \frac{1}{1 + e^{-z}} \tag{1}$$

Sigmoid function, $\sigma\ (z)$ = output between 0 and 1 (probability estimate), z = input to the function and e = base of natural log.

Decision Tree (DT)

The DT algorithm is part of the supervised learning algorithm family, and its main objective is to construct a training model that can be used to predict the class or value of target variables through learning decision rules inferred from the training data [8]. It splits data into subsets based on the value of input features. It uses algorithms like **Gini Impurity**, **Entropy**, or other measures to evaluate splits and determine the structure of the tree [9, 10].

K-Nearest Neighbor (KNN)

The **K-Nearest Neighbors (KNN)** algorithm is a simple and intuitive classifier that assigns a class to a data point based on the majority class of its k nearest neighbors in the feature space [11, 12].

A. Distance metric

The core of KNN is determining the "nearest neighbors" using a distance metric. Common metrics include: Euclidean Distance (default in many implementations), Manhattan Distance, Minkowski Distance (generalized form).

After computing the distances from the query point to all points in the training dataset, the k nearest neighbors is selected, and the class is determined based on a voting mechanism: Majority Voting, Weighted Voting, Weighted Classification Rule:

Steps in KNN

1. **Compute Distances**: Measure distances from the query point to all points in the training set.
2. **Select Neighbors**: Identify the k nearest neighbors.
3. **Vote**: Use majority or weighted voting to assign the class.
4. **Output**: Predict the class label for the query point.

Random forest classifier

The **Random Forest Classifier** is an ensemble learning algorithm that combines multiple decision trees to make predictions. It leverages the concept of "bagging" (Bootstrap Aggregating) and introduces randomness in feature selection to enhance diversity among the trees, reducing overfitting and improving generalization [13].

Key steps in random forest

Bootstrapping (Bagging): Randomly sample n observations with replacement from the training data to create multiple subsets (bootstrap samples). Each subset is used to train a decision tree.

Random Feature Selection: At each split of a decision tree, a random subset of $m_{features}$ is selected from the total $n_{features}$. The split is made based on the best feature from this random subset, determined by a metric like Gini Impurity or Entropy.

Aggregation: Predictions from all decision trees are aggregated: For Classification: Majority voting is used. For Regression: Average of the predictions is taken.

Extra trees classifier

The **Extra Trees Classifier** (Extremely Randomized Trees Classifier) is an ensemble learning method that builds a large number of randomized decisions trees and aggregates their outputs (via majority voting) to make predictions. It is similar to a Random Forest but introduces more randomness by selecting split points randomly for each features [14].

Key characteristics of extra trees classifier

Randomness in Split Selection: Instead of finding the optimal split point (as in Decision Trees or Random Forest), Extra Trees selects split points randomly for a given feature. This increases diversity among the trees and reduces overfitting.

Feature Subset: Similar to Random Forest, Extra Trees randomly selects a subset of features for each split, promoting variance reduction.

Gradient boosting classifier

The **Gradient Boosting Classifier** is an ensemble learning method that builds a series of decision trees sequentially, where each tree corrects the errors made by the previous ones. It combines weak learners (typically shallow decision trees) into a strong learner, optimizing a specified loss function through gradient descent [15].

Key steps

Sequential Model Building: Each tree is trained to minimize the residual errors (gradients) of the previous trees. Trees are added iteratively, and their predictions are weighted to improve the overall model.

Gradient Descent Optimization: The method uses gradient descent to minimize a differentiable loss function. The residual errors are treated as the gradient of the loss function with respect to the predictions.

Adaptive boosting (AdaBoost) classifier

The AdaBoost (Adaptive Boosting) classifier combines multiple "weak" classifiers to form a "strong" classifier. It works iteratively, adjusting the weights of misclassified samples to focus on harder examples [16].

Ensemble learning

Ensemble learning is a technique in machine learning where multiple models (often called *base learners*) are combined to make predictions. The main idea is to improve the overall performance by combining the strengths of individual models, often reducing bias, variance, or both.

Here are some key mathematical formulas related to ensemble learning. The ensemble model H(x) is then defined as:

$$H_{(x)} = f(h_1(x), h_2(x), \ldots, h_m(x))$$

Where f is a function that combines the predictions from the base learners. The specific form of f depends on the type of ensemble method being used (e.g., averaging, majority voting, etc.).

- **Bagging** (Regression):

$$\widehat{y}_{ensemble} = \frac{1}{m}\sum_{i=1}^{m}(x) \tag{2}$$

- **Boosting** (Regression):

$$H(x) = \sum_{i=1}^{m}\alpha_i h_i(x) \tag{3}$$

- **Boosting** (Classification):

$$H(x) = sign\left(\sum_{i=1}^{m}\alpha_i h_i(x)\right) \tag{4}$$

- **Random Forest**: Similar to bagging but with random feature selection at each split.
- **Stacking**:

$$H(x) = g(h_1(x), h_2(x), \ldots \ldots h_m(x)) \tag{5}$$

Different algorithm was tested using grid search method and compared with their accuracy and an algorithm with a higher accuracy was further evaluated with precision, recall, F-1 score, and AUC. The hyperparameters for each model were optimized using Grid Search with 5-fold cross-validation to strike the right balance between model complexity and predictive performance.

For **Logistic Regression**, the optimal configuration included a regularization strength (C) of 0.01, a maximum iteration (max_iter) of 100, the L2 penalty for regularization, and the SAGA solver for optimization. These settings were selected to balance accuracy and generalizability.

The **Decision Tree** model was optimized with a focus on capturing underlying patterns without overfitting. The best-performing configuration included the Gini impurity for split evaluation, unrestricted tree depth, a minimum of 4 samples per leaf node, and a minimum of 2 samples required for splitting.

In the **Random Forest** model, the optimal hyperparameters included unrestricted tree depth, a minimum of 2 samples for splitting, and an ensemble of 200 trees. These settings leverage the strengths of Random Forest in reducing overfitting through ensemble learning, ensuring both model complexity and predictive accuracy.

For the **K-Nearest Neighbors** model, the best configuration used the Manhattan distance metric, considered 3 nearest neighbors for predictions, and weighted neighbors by their distance. This combination enhances the model's ability to capture local patterns while prioritizing closer data points.

The **Gradient Boosting** model was optimized with a learning rate of 0.2, a maximum tree depth of 3, and an ensemble of 200 trees. These settings balance model performance and complexity, enabling efficient learning while minimizing overfitting. The relatively high learning rate allows quick convergence, while the shallow tree depth helps prevent overly complex models.

In **AdaBoost**, the learning rate was set to 1, with an ensemble of 200 estimators. These settings ensure effective learning by adjusting the contribution of each weak learner, while the ensemble size allows the model to capture complex patterns.

Finally, the **Extra Trees Classifier** model was optimized with unrestricted tree depth, a minimum of 2 samples required for splitting, and an ensemble of 100 trees. These settings enable the model to capture intricate patterns in the data while maintaining computational efficiency. The model achieved a best cross-validated accuracy of 0.94, reflecting its strong generalization capability.

These hyperparameters were carefully selected using the grid search method to ensure a good balance between model complexity, predictive accuracy, and generalization performance.

Model evaluation

Accuracy is a common metric for overall model performance, it may not be reliable with imbalanced datasets. Hence, additional metrics like precision, recall, and AUC score are valuable for a comprehensive evaluation. Precision assesses the correctness of predictions, recall emphasizes the true positive rate, and the F-1 score combines precision and recall into a single metric for a balanced evaluation.

Classification accuracy: It indicates how frequently the model predicts the correct outcome. It can be calculated as the ratio of the classifier's correct predictions to the total number of predictions made by the classifiers. The formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

Precision: Is the number of correct outputs provided by the model or how many of the positive classes predicted correctly by the model were actually true.

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

Recall: is the percentage of positive classes predicted correctly by the model out of a total of positive classes.

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

F-score: is a metric used to evaluate the performance of classification models, particularly in cases of imbalanced datasets. It is the harmonic mean of precision and recall, giving a single measure that balances the two.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{9}$$

Result

Out of the entire population 75,360 (weighted) samples of women aged 25–49 from the chosen African nation were included. Variables were chosen from previous literature. Table 1 shows count and percentage of the categories under each variable.

Different countries have different cervical screening uptake which can be seen from Fig. 1. In most countries, a significant majority of women fall under the **not tested** category (purple). Countries like **Gabon** (17.4%) and **Mozambique** (14.9%) have comparatively higher proportions of women who have been screened for cervical cancer, though these percentages are still relatively small. In countries such as **Benin** (0.8% tested) and **Mauritania** (0.7% tested), almost no women have been screened for cervical cancer.

However, for our final model we dropped missing values for variables with greater tha 20% and the final data is 53,461 in total with 38,907 women were screened for cervical cancer, while 14,554 women were not screened for cervical cancer with any test. As the data was imbalance, we performed class balancing by oversampling the minority classes. The final number was 77,814 with equal number of women in both classes that can be seen in Fig. 2.

After class balancing, we encode the variable based on the variable type. We used one hot encoding for ordinal variables and label encoding for nominal variables. The variables selected from different variables were then tested for their effect on whether a women will be screened for cervical cancer or not. We used random forest to explore the effect of the variables.

Figure 3 presents the top 10 features of influence in predicting the likelihood of cervical cancer screening uptake, with emphases on access to healthcare, awareness, and sociocultural factors. Major predictors include history of health testing, health visits, and access to media such as radio and TV, probably increasing awareness. Other important ones are proximity to healthcare facilities, use of contraceptives, urban residence, and employment status; these reflect accessibility and prioritization of care. Other sociocultural variables include partner education, religious affiliation, and age-related factors such as age at first sexual activity. Improvement in health access, awareness campaigns, and cultural considerations can improve the uptake of screening.

Using the top features different algorithms were tested(Table 2). Among the tested algorithms, Extra Trees Classifier had the best accuracy of 94.13% and a very balanced performance in all metrics, closely followed by Random Forest and Adaptive Boosting. K-Nearest Neighbor and Decision Tree models were also pretty strong, with their accuracies being above 90%. These results emphasize how well the ensemble methods, such as Extra Trees Classifier and Random Forest, can be used for robust and accurate classification tasks on the given dataset.

Figure 4 visualizes the result of grid search optimization for Extra Trees Classifier on accuracy, precision, recall, and F1-score, where most of the combinations have a score of over 0.90 in all of them, meaning that the model is robust and suitable for the prediction of cervical cancer screening uptake. The noticeable performance dip around parameter indices 13–16 shows sensitivity to certain hyperparameters-probably related to tree splits, estimators, or depth-scores recover and afterward stabilize, reflecting the model's robustness when its parameters are well tuned. This again highlights the importance of hyperparameter optimization to sustain a balanced performance and prevent overfitting or underfitting of the model.

Discussion

This paper has demonstrated the potential of machine learning models in predicting cervical cancer screening uptake among women in an African setting. These results have shown that the ensemble methods, especially the Extra Trees Classifier, are very efficient in performing the classification task at hand with the highest accuracy of 94.13% and the best balance between the precision, recall, and F-score. Its strong generalization ability makes this model appropriate for the identification of critical predictors of cervical cancer screening uptake and targeted intervention strategies. This aligns with findings who highlighted the strong performance of ensemble methods in cervical cancer diagnosis [17]. Similarly, some studies emphasized the importance of feature selection in improving model accuracy [18]. also found that tree-based models outperform other classifiers when properly trained [19]. These comparisons affirm the reliability of ensemble models in predictive healthcare analytics.

A comparative study of machine learning and statistical survival models for cervical cancer prognosis found that machine learning approaches, particularly Random Survival Forests, outperformed traditional methods in predictive accuracy and feature selection. These findings emphasize the advantages of machine learning in handling intricate patterns within healthcare datasets [20]. Traditional statistical models, such as logistic regression, have been widely utilized in healthcare due to their

Table 1 Descriptive counts for different variable with cervical cancer screening uptake

Variable	Category	Count	Percentage	Cervical cancer screened	
				Yes	No
Residence	Rural	42,192	58.53	3440	38,752
	Urban	29,893	41.47	4385	25,508
Radio	No	35,511	49.26	3066	32,445
	Yes	35,360	49.05	4606	30,754
	Not a resident	1214	1.68	153	1061
Tv	No	38,503	53.41	2709	35,794
	Yes	32,368	44.90	4963	27,405
	Not a resident	1214	1.68	153	1061
Religion	Muslim	30,961	42.98	3231	27,730
	Catholic	10,895	15.11	1729	9166
	Protestant	7733	10.73	1156	6577
	Traditional	7508	10.41	979	6529
	No religion	5665	7.86	303	5362
	Invalid	8942	12.4	381	8561
	Other	381	0.53	46	335
Sex of the House hold	Male	51,261	71.11	5243	46,018
	Female	20,824	28.89	2582	18,242
Contraceptive type	Modern	207,828	28.83	3197	17,585
	Traditional	2780	3.86	358	2422
	Folkloric	259	0.36	26	233
	No method	48,264	66.95	4244	44,020
Health visit	Yes	42,031	58.31	5677	36,354
	No	30,054	41.69	2148	27,906
Cigarette smoker	No	71,106	98.64	7736	63,370
	Yes	979	1.36	89	890
Distance to hospital	Not a big problem	46,713	64.80	5732	40,981
	Big problem	25,372	35.20	2093	23,279
Health insurance	No	58,892	81.70	6320	52,572
	Yes	13,193	18.30	1505	11,688
Marital status	Married	42,017	58.29	4297	37,720
	Living with partner	15,569	21.60	1719	13,850
	Never in union	5097	7.07	639	4459
	Separated	4496	6.24	645	3851
	Widowed	2554	3.54	341	2213
	Divorced	2352	3.26	184	2168
Partners occupation	Agricultural self employed	11,794	20.34	763	11,031
	Skilled manual	9375	16.17	1200	8175
	Agricultural employee	7051	12.16	66	642
	Professional/technical/managerial	6761	11.66	1355	5406
	Services	6033	10.40	453	5580
	Sales	5000	8.62	629	4371
	Unskilled manual	4267	7.36	434	3833
	Clerical	754	1.30	113	641
	Domestic	346	0.60	50	296
	Did not work	5904	10.18	433	5471
	Do not know	708	1.22	66	642

Table 1 (continued)

Variable	Category	Count	Percentage	Cervical cancer screened	
				Yes	No
Occupation	Sales	12,382	17.20	1589	10,793
	Agriculture self employed	10,320	14.34	641	9679
	Services	7082	9.84	566	6516
	Agriculture employee	5807	8.07	580	5227
	Professional/technical/managerial	5393	7.50	1421	3972
	Skilled manual	4076	5.66	382	3694
	Unskilled manual	3759	5.22	504	3255
	Domestic	1397	1.94	208	1189
	Clerical	711	0.99	173	538
	Not working	20,904	29.04	1700	19,204
	Do not know	158	0.22	30	128
STI	No	67,981	94.31	7183	60,798
	Yes	3995	5.54	634	3361
	Do not know	109	0.15	8	101
HIV test	Yes	47,311	65.63	718	24,056
	No	24,774	34.37	7107	40,204

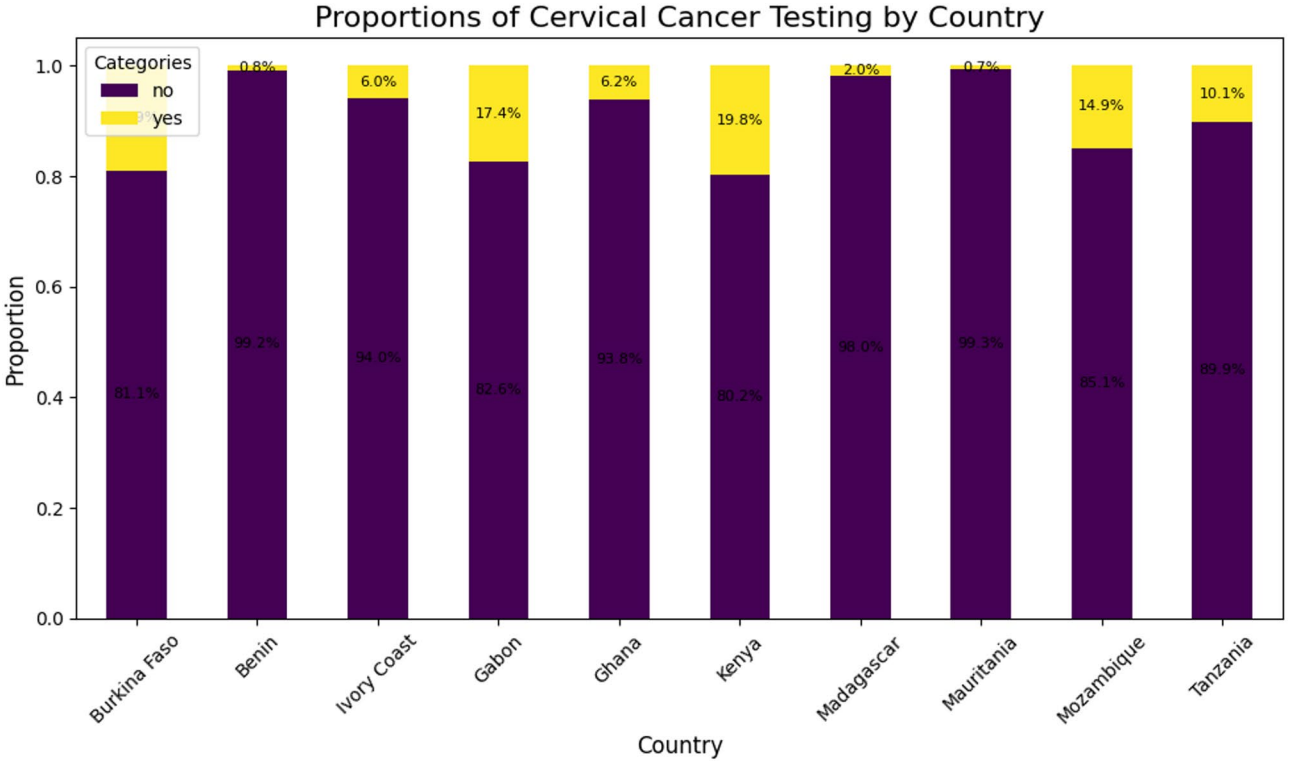


Fig. 1 Cervical cancer screening uptake in different African countries

interpretability and theoretical robustness. However, they rely on linear assumptions and may not adequately capture complex relationships among predictive variables. In contrast, machine learning models, especially ensemble techniques like the Extra Trees Classifier and Random Forest, excel in modeling non-linear interactions and handling high-dimensional data [21].

The analysis indicated that cervical cancer screening uptake was influenced by different factors. Among these, access to healthcare variables such as health visits, proximity to healthcare facilities, and having health insurance emerged as top predictors. These findings are consistent with previous studies that indicate reducing barriers to healthcare access is crucial in improving the

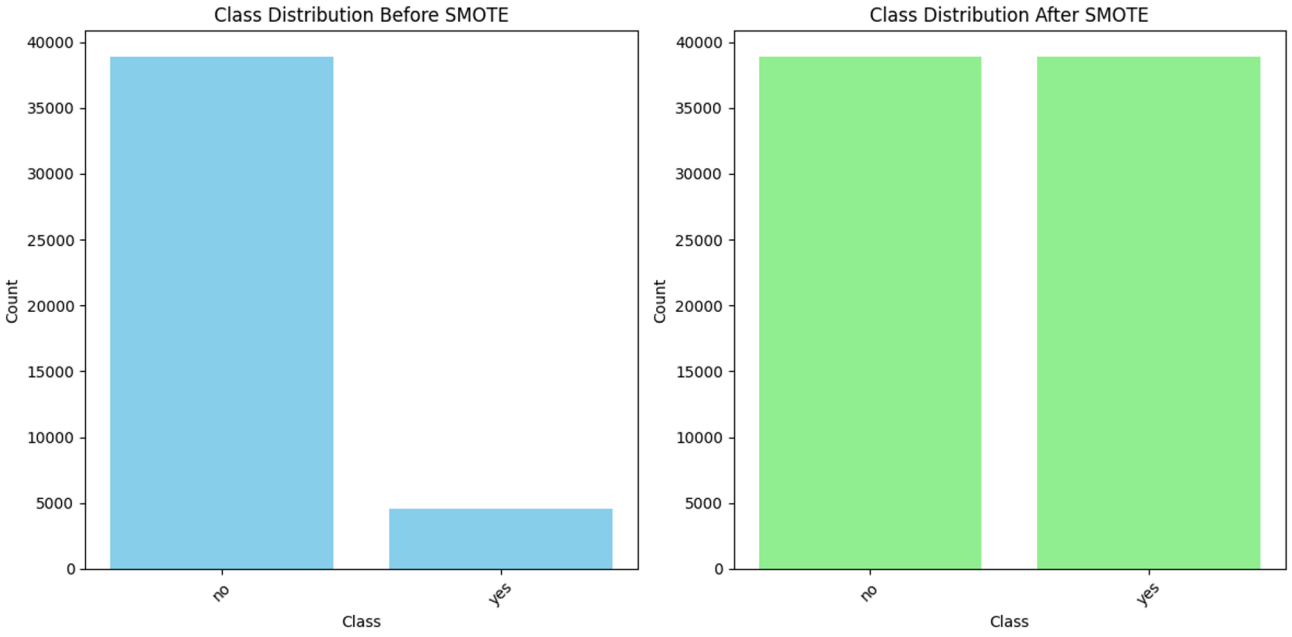


Fig. 2 Number of datasets under each class before and after class balancing

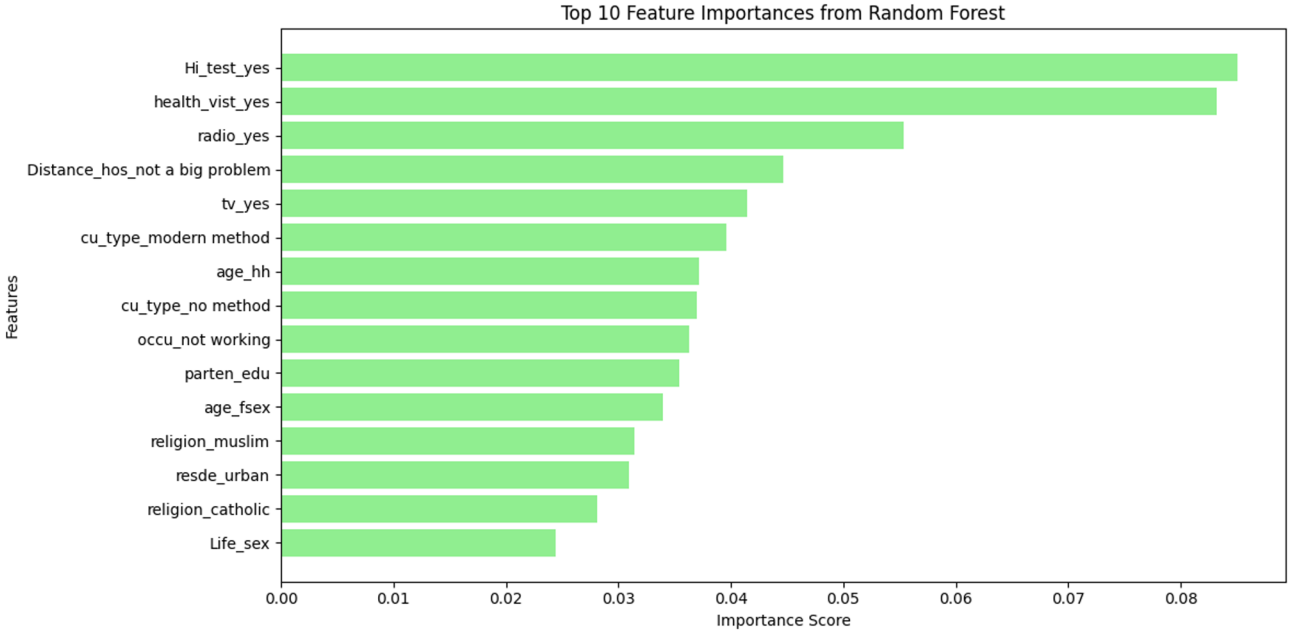


Fig. 3 Feature importance using random forest for exploring the effect of variables on the cervical cancer screening uptake of a women

Table 2 Evaluation metrics for different algorithms

Algorithm	Accuracy (100%)	Precision (100%)	Recall (100%)	F-score (100%)
Logistic Regression	70.91	69.08	75.65	72.20
K-Nearest Neighbor	90.44	91.95	90.44	90.35
Random Forest	93.87	99.18	88.56	91.60
Gradient Boosting	89.32	86.00	89.00	85.00
Adaptive Boosting	92.65	98.25	87.07	90.14
Extra trees classifier	94.13	95.76	94.12	93.80
Decision tree	90.35	92.90	87.90	88.42
Ensemble	86.00	84.00	86.00	85.00

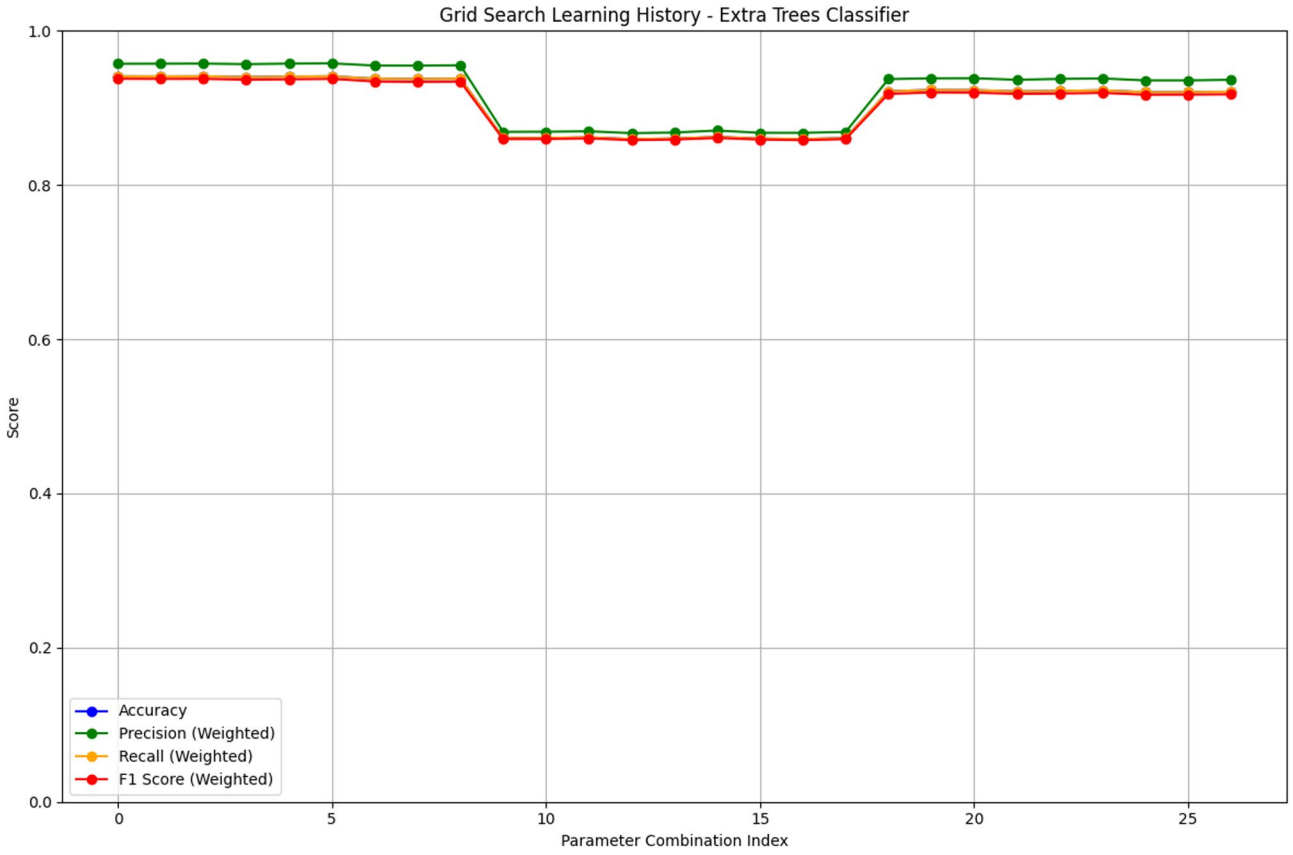


Fig. 4 Grid search learning history for the highest achieving algorithm

rates of screening for this disease [22, 23]. Regular health visits have been consistently linked to increased cervical cancer screening rates. Studies show that routine interactions with healthcare providers provide opportunities for education, awareness, and encouragement to undergo screening. A study in Nigeria found that contact with healthcare professionals significantly increased the likelihood of women getting screened, as healthcare visits allow for discussions about the importance of early detection and preventive care [24]. Similarly, research from Nicaragua highlighted that women who engaged with healthcare services were more likely to participate in cervical cancer screening programs due to increased awareness and accessibility of services [25]. These findings align with global patterns, where regular engagement with healthcare services has been identified as a strong predictor of cervical cancer screening, particularly in low- and middle-income countries where awareness and accessibility are major barrier.

Sociocultural factors such as marital status, partner's occupation, and religious affiliation also played a significant role. Married women and those in cohabiting relationships had higher rates of screening, possibly reflecting greater support from partners in making health care decisions [26]. Various studies indicate that marriage is generally associated with better health outcomes, including increased likelihood of preventive care utilization. For instance, research suggests that married individuals tend to engage more in health-promoting behaviors due to spousal encouragement and shared decision-making regarding medical care [27].

The occupation of the partner influenced screening, with women whose partners were professionals or employed in skilled manual labor more likely to be screened. This trend may reflect the socioeconomic advantages associated with these occupations, which facilitate access to healthcare services. A study conducted in Central Uganda found that women whose partners provided emotional encouragement had significantly higher odds of undergoing screening [28]. In Nigeria, research demonstrated that male involvement, particularly among medical staff, positively impacted their partners' uptake of Pap smear screenings. Partners of medical staff had higher screening rates compared to those of paramedics and non-medical workers [29]. This suggests that a partner's occupation, especially in healthcare, may enhance awareness and support for cervical cancer screening.

Media access, particularly radio and TV exposure, strongly predicted screening uptake, corroborating

findings from other studies that highlight the role of mass media in raising awareness about cervical cancer and its prevention [30]. Studies have demonstrated that media campaigns can effectively raise awareness and encourage women to participate in screening programs [31]. Similarly, research in the UK examined the impact of media coverage surrounding a celebrity's cervical cancer diagnosis. The study found a significant, albeit temporary, rise in cervical screening uptake, particularly among women aged 25–44 years, following the media reports. This underscores the potential of media reporting to influence health behaviors and increase screening attendance [32].

Women residing in urban areas were more likely to be screened than their rural counterparts, indicating disparities in healthcare accessibility and awareness between urban and rural settings. A study analyzing India's National Family Health Survey (2019-21) found that cervical cancer screening coverage was higher in urban areas versus in rural regions. The study attributed these disparities primarily to differences in resources such as wealth, knowledge, and social networks, which are more prevalent in urban settings Addressing these disparities through mobile clinics or community-based interventions may improve rural screening rates. For example, a community-based mobile cervical cancer screening program in rural India successfully utilized peer education and screening in existing community spaces. This approach demonstrated the feasibility of reaching underserved populations through mobile clinics [33, 34].

Contraceptive use, specifically modern methods, was another strong predictor, which points to the opportunity for healthcare providers to advocate for cervical cancer screening during family planning visits. A study conducted in Guinea demonstrated the feasibility of combining family planning counseling with mass cervical cancer screening campaigns [35]. Similarly, research in Kenya assessed the integration of cervical cancer screening within family planning clinics. The study found that such integration is beneficial for clients, as the clinics were well-prepared to provide high-quality screening services [36]. Furthermore, a review focusing on Sub-Saharan Africa highlighted that integrating cervical cancer prevention services into existing family planning or HIV/AIDS service delivery platforms can rapidly expand "screen and treat" programs [37]. These findings are supported by other studies where family planning visits were used as a venue for other preventive health services as well [38]. However, it also noted that family planning services reach only a small percentage of women most at risk, indicating the need for strategies to extend services to a broader population.

Conclusion

This study indicates the potential of machine learning models, particularly the ensemble methods like the Extra Trees Classifier, in predicting cervical cancer screening uptake for women within African contexts. With a high accuracy rate of 94.13%, these models have excellent generalizability, and therefore they are valuable tools in identifying major predictors and informing targeted intervention strategies.

The primary determinants of uptake in screening include health access determinants such as high health visitation, facility access to health facilities, and insurance cover from health insurance. The findings confirm other international studies that have identified a decline in healthcare access barriers significantly improving the screening rates. Sociocultural determinants such as marriage status, occupation of partner, and access to mass media also influence screening. Married women with partner support or more access to information related to health via mass media have better opportunities to enroll for screening.

Additionally, urban living was associated with higher screening rates than rural living, emphasizing disparities in access. Mobile clinics and community interventions would be able to bridge this gap. The study implies the merit of integrating cervical cancer screening with family planning care to maximize preventive care utilization.

In general, these findings revisit the importance of multi-faceted interventions targeting healthcare access, sociocultural processes, and health communication for cervical cancer screening promotion. Utilizing machine learning models and integrating screening programs into routine healthcare services, policymakers and healthcare providers are in a better position to design more impactful interventions that contribute to higher uptake of screening and reductions in cervical cancer burden among disadvantaged populations.

These findings carry important implications for medical informatics and public health. Machine learning models can be leveraged in support of high-risk group identification and the optimization of resource allocation across cervical cancer screening programs. Policy makers and providers will use these insights to help frame data-driven intervention design targeting structural and socio-cultural barriers to screening. Future studies will be related to the validation of findings across populations with diverse distributions and integration within clinical decision support systems to yield scalable impact.

Strength and limitation

Strengths

High Model Performance: The use of ensemble machine learning models, particularly the Extra Trees Classifier, resulted in high accuracy (94.13%) and balanced

precision, recall, and F-score. This demonstrates the robustness of the model in predicting cervical cancer screening uptake.

Identification of Key Predictors: The study effectively identified critical predictors of screening uptake, including healthcare access (health visits, proximity to facilities, and insurance coverage), sociocultural factors (marital status, partner's occupation), media exposure, and contraceptive use. These findings provide actionable insights for targeted intervention strategies.

Integration of Sociocultural and Healthcare Access Factors: Unlike many studies that focus solely on clinical risk factors, this research incorporated sociocultural variables, improving the model's real-world applicability in African settings where social determinants significantly influence healthcare behaviors.

Evidence-Based Comparisons: The study findings were compared with previous literature, reinforcing the reliability of the results. The alignment with global and regional studies enhances the study's credibility and contextual relevance.

Potential for Policy and Intervention Design: The findings can inform public health policies by highlighting the need for healthcare access improvements, community-based interventions in rural areas, and media campaigns to enhance screening awareness. The study also supports the integration of cervical cancer screening with family planning services to improve screening uptake.

Limitations

Data Balancing and Generalizability: While balancing the dataset improved model performance, it may limit real-world application since actual screening uptake is often imbalanced, with more women not undergoing screening. This could result in an overestimation of screening likelihood in deployment settings.

Limited Scope of Predictors: Although the study included key predictors, other factors such as individual perceptions of screening, cultural beliefs, and healthcare provider influence were not explicitly modeled, which could further refine the predictive accuracy.

Cross-Sectional Nature of the Study: The study relies on cross-sectional data, meaning it captures a snapshot in time rather than assessing changes in screening behavior over time. Longitudinal studies would be needed to establish causal relationships.

Limited Geographic and Population Scope: The findings are based on an African setting and may not be fully generalizable to other regions with different healthcare systems, screening policies, and cultural influences. Expanding the study to include diverse populations could strengthen its applicability.

Barriers to Implementation of Model Findings: Although the study suggests policy interventions such as media campaigns and mobile clinics, implementing these recommendations requires financial and infrastructural support that may not always be feasible in resource-limited settings.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-03039-y.

Supplementary Material 1

Acknowledgements
We would like to give our gratitude to DHS for making the data available and for letting us use it for our research.

Author contributions
FGA: Major role in analysis and model selection with participation in writing. EAA: Major role in data extraction and cleaning with participation in writing. EAT: Major role in writing result and discussion. TKT: Participated in the writing and proof reading. ZBT: Participated analysis and model selection. TGG: Participated in the writing and proof reading.

Funding
This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Data availability
All the relevant datasets and materiales was accesed on DHS website https://www.dhsprogram.com/data/available-datasets.cfm and https://www.dhsprogram.com/Data/Guide-to-DHS-Statistics/index.cfm.

Declarations

Ethics approval and consent to participate
This study was conducted in accordance with the ethical guidelines outlined in the Declaration of Helsinki. The DHS Program obtained ethical approval for the data collection processes from the relevant institutional review boards of each participating country and ensured that participants provided informed consent. This study is based on publicly available data from the Demographic and Health Surveys (DHS) Program. The DHS Program ensures that all individual participants provide informed consent at the time of data collection.

Consent for publication
DHS allows for the publishing of findings gathered from the dataset.

Relevant guidelines and regulations
The relevant guidelines and regulations can be found at https://www.dhsprogram.com/Privacy-Policy.cfm.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Public Health, Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia
[2]Department of Epidemiology and Biostatistics, Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia
[3]Department of Health Informatics, School of Public Health, College of Medicine and Health Sciences, Samara University, Semera, Ethiopia
[4]Department of Health Informatics, Institute of Public Health, College of Medicine and Health Sciences, University of Gondar, Gondar, Ethiopia
[5]Department of Maternal and Child Health, Lemi Kura Subcity Health Office, Addis Ababa City Administration Health Bureau, Addis Ababa, Ethiopia

Received: 23 January 2025 / Accepted: 19 May 2025

Published online: 26 May 2025

References

1. Cheng S, Liu S, Yu J, Rao G, Xiao Y, Han W, et al. Robust whole slide image analysis for cervical cancer screening using deep learning. Nat Commun. 2021;12(1):5639.
2. Torgovnik J. Cervical Cancer. Worled Health Organization(WHO); 2023.
3. Hull R, Mbele M, Makhafola T, Hicks C, Wang SM, Reis RM, et al. Cervical cancer in low and middle–income countries. Oncol Lett. 2020;20(3):2058–74.
4. WHO. Cervical cancer 2023. Available from: https://www.who.int/health-topics/cervical-cancer#tab=tab_1
5. Singh D, Vignat J, Lorenzoni V, Eslahi M, Ginsburg O, Lauby-Secretan B, et al. Global estimates of incidence and mortality of cervical cancer in 2020: a baseline analysis of the WHO global cervical Cancer elimination initiative. Lancet Global Health. 2023;11(2):e197–206.
6. Petersen Z, Jaca A, Ginindza T, Maseko G, Takatshana S, Ndlovu P, et al. Barriers to uptake of cervical cancer screening services in low-and-middle-income countries: a systematic review. BMC Womens Health. 2022;22(1):486.
7. Nishadi AT. Predicting heart diseases in logistic regression of machine learning algorithms by Python jupyterlab. Int J Adv Res Publications. 2019;3(8):1–6.
8. Charbuty B, Abdulazeez A. Classification based on decision tree algorithm for machine learning. J Appl Sci Technol Trends. 2021;2(01):20–8.
9. Quinlan JR. Improved use of continuous attributes in C4. 5. J Artif Intell Res. 1996;4:77–90.
10. Breiman L. Classification and regression trees. Routledge; 2017.
11. Cover T. Nearest neighbor pattern classification. IEEE Trans Inform Theory. 1968;4(5):515–6.
12. Dudani S. The distance-weighted k-nearest neighbor rule. IEEE Trans Syst Man Cybernet. 1978;8(4):311–3.
13. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
14. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63:3–42.
15. Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of statistics. 2001:1189–232.
16. Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. Math Intelligencer. 2005;27(2):83–5.
17. Lu J, Song E, Ghoneim A, Alrashoud M. Machine learning for assisting cervical cancer diagnosis: an ensemble approach. Future Generation Comput Syst. 2020;106:199–205.
18. Nithya B, Ilango V. Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction. SN Appl Sci. 2019;1:1–6.
19. Ding D, Lang T, Zou D, Tan J, Chen J, Zhou L, Wang D, Li R, Li Y, Liu J, Ma C. Machine learning-based prediction of survival prognosis in cervical cancer. BMC Bioinformatics. 2021;22(1):331.
20. Elmi AH, Abdullahi A, Ali Bare M. A comparative analysis of cervical cancer diagnosis using machine learning techniques. Indonesian J Electr Eng Comput Sci. 2024;34(2):1010.
21. Kolasseri AE. Comparative study of machine learning and statistical survival models for enhancing cervical cancer prognosis and risk factor assessment using SEER data. Sci Rep. 2024;14(1):22203.
22. Black E, Hyslop F, Richmond R. Barriers and facilitators to uptake of cervical cancer screening among women in Uganda: a systematic review. BMC Womens Health. 2019;19(1):1–12.
23. Anorlu R, Obiako R, Oyeneyin L. Cervical cancer screening in low-resource settings: challenges and opportunities. Int J Women's Health. 2022;14:15–28.
24. Williams-Brennan L, Gastaldo D, Cole DC, Paszat L. Social determinants of health associated with cervical cancer screening among women living in developing countries: a scoping review. Arch Gynecol Obstet. 2012;286:1487–505.
25. Claeys P, Gonzalez C, Gonzalez M, Page H, Bello RE, Temmerman M. Determinants of cervical cancer screening in a poor area: results of a population-based survey in Rivas, Nicaragua. Tropical Med Int Health. 2002;7(11):935–41.
26. Vu M, Yu J, Awolude OA, Chuang L. Cervical cancer worldwide. Curr Probl Cancer. 2018;42(5):457–65.
27. Liu H, Umberson DJ. The times they are a Changin': marital status and health differentials from 1972 to 2003. J Health Soc Behav. 2008;49(3):239–53.
28. Isabirye A. Individual and intimate-partner factors associated with cervical cancer screening in central Uganda. PLoS ONE. 2022;17(9):e0274602.
29. Olowokere IE, Roberts OA. Male workers' influence on partners uptake of pap smear screening in a teaching hospital in Nigeria. 2013.
30. Mutyaba T, Faxelid E, Mirembe F, Weiderpass E. Influences on uptake of reproductive health services in Nsangi community of Uganda and their implications for cervical cancer screening. Reproductive Health. 2007;4:1–9.
31. Anderson JO, Mullins RM, Siahpush M, Spittal MJ, Wakefield M. Mass media campaign improves cervical screening across all socio-economic groups. Health Educ Res. 2009;24(5):867–75.
32. MacArthur GJ, Wright M, Beer H, Paranjothy S. Impact of media reporting of cervical cancer in a UK celebrity on a population-based cervical screening programme. J Med Screen. 2011;18(4):204–9.
33. Garg P, Krishnamoorthy Y, Halder P, Rajaa S, Verma M, Kankaria A, Goel A, Kakkar R. Urban-rural disparities in cervical cancer screening among Indian women between 30–49 years: a Geospatial and decomposition analysis using a nationally representative survey. BMC Cancer. 2025;25(1):67.
34. Srinivas V, De Cortina SH, Nishimura H, Krupp K, Jayakrishna P, Ravi K, Khan A, Madhunapantula SV, Madhivanan P. Community-based mobile cervical cancer screening program in rural India: successes and challenges for implementation. Asian Pac J Cancer Prevention: APJCP. 2021;22(5):1393.
35. Leno DW, Diallo FD, Delamou A, Komano FD, Magassouba M, Niamy D, Tolno J, Keita N. Integration of family planning counselling to mass screening campaign for cervical cancer: experience from Guinea. Obstet Gynecol Int. 2018;2018(1):3712948.
36. Claeys P, De Vuyst H, Mzenge G, Sande J, Dhondt VE, Temmerman M. Integration of cervical screening in family planning clinics. Int J Gynecol Obstet. 2003;81(1):103–8.
37. White HL, Meglioli A, Chowdhury R, Nuccio O. Integrating cervical cancer screening and preventive treatment with family planning and HIV-related services. Int J Gynecol Obstet. 2017;138:41–6.
38. Teguete I, Muwonge R, Traore C, Dolo A, Bayo S, Sankaranarayanan R. Can visual cervical screening be sustained in routine health services? Experience from Mali, Africa. BJOG: Int J Obstet Gynecol. 2012;119(2):220–6.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.