Machine Learning for Predicting Cervical Cancer: A Random Forest Approach Using Combined Nigerian and Public Datasets

Helen Olaronke Edogbanya , Olayiwola Isaac Babarinsa, Abosede Mary Oke

Department of Mathematics, Federal University Lokoja, P.M.B 1154, Kogi, Nigeria

DOI: https://doi.org/10.51584/IJRIAS.2026.11090141

Received: 02 October 2026; Accepted: 07 October 2026; Published: 22 October 2026

ABSTRACT

Cervical cancer remains a leading cause of cancer related mortality among women in low and middle-income countries, where screening and vaccination rates are low. This paper presents a Random Forest machine learning model for predicting cervical cancer risk using a combined dataset of 908 patients from the Federal Teaching Hospital Lokoja (FTHL), Nigeria (50 records) and the UCI Cervical Cancer Risk Factors repository (858 records). The model was trained on 11 attributes including age, HPV status, sexual history, smoking, hormonal contraceptive use, and reproductive factors. Hyperparameter tuning via grid search with cross validation which yielded optimal parameters (mtry=3, ntree=226, min_n=6) with a cross validated accuracy of 0.966. On the test set, the model achieved 97.4% accuracy, 98.2% precision, 99.1% sensitivity, 98.6% F1 score, and an AUC of 0.901. Permutation feature importance identified HPV status (mean importance = 0.238) and hormonal contraceptive use (mean importance = 0.125) as the most significant predictors. The high sensitivity is particularly valuable for clinical applications, ensuring that nearly all cancer cases are detected. The integration of FTHL and UCI datasets enhanced model generalizability by exposing it to diverse patient populations. However, validation revealed that pooled performance was substantially driven by data source confounding, with leave-one-cohort-out AUC falling to 0.53–0.70 and within-FTHL AUC of 0.39, highlighting the need for careful interpretation and prospective validation.

Keywords: Cervical cancer, Machine Learning, Random Forest, HPV, Risk Prediction

INTRODUCTION

Cervical cancer is the fourth most common cancer among women globally, with an estimated 604,000 new cases and 342,000 deaths annually (Girone et al., 2024). The disease is caused by persistent infection with high-risk human papillomavirus (HPV), particularly types 16 and 18, which account for approximately 70% of all cervical cancers (Bosch et al., 2002). While screening programs using Pap smears and HPV DNA testing have significantly reduced incidence in high-income countries, cervical cancer remains a major public health challenge in low and middle-income countries (LMICs), where approximately 90% of deaths occur (Huang et al., 2024).

In Nigeria, cervical cancer is the second most common female cancer, accounting for 17.1% of new cases (GLOBOCAN 2022 estimates), with high mortality due to low screening rates and late diagnosis (Ogunniyi et al., 2025). The WHO's 90-70-90 targets for cervical cancer elimination emphasize vaccination, screening, and treatment, yet gaps remain in vaccine acceptance, self-sampling, and therapeutic regimes (Omohwovo et al., 2025).

Machine learning (ML) has emerged as a powerful tool for disease prediction and risk stratification, particularly for handling high-dimensional data, capturing non-linear relationships, and providing interpretable feature importance measures (Bhardwaj et al., 2024). The Random Forest algorithm has consistently demonstrated superior performance in cervical cancer prediction across multiple studies (Nithya and Ilango, 2019; Mahendra and Anasanti, 2025). However, existing studies have relied on single retrospective datasets, primarily the UCI Cervical Cancer Risk Factors dataset (Fernandes et al., 2017),

which is limited to 858 patients from a single hospital in Venezuela. The generalizability of models trained on such data to other populations, particularly Nigerian women, remains unverified.

This study addresses this gap by integrating clinical data from FTHL, Nigeria, with the UCI dataset to develop and validate a Random Forest model for cervical cancer prediction. The specific objectives are:

1. To develop a Random Forest classifier for cervical cancer prediction using combined FTHL and UCI data.
2. To evaluate model performance using comprehensive metrics appropriate for imbalanced data.
3. To identify the most influential predictors of cervical cancer risk through permutation feature importance.
4. To assess the generalizability of the model across different populations and healthcare settings.

MATERIALS AND METHODS

Data Sources and Integration

The study integrates two complementary datasets:

FTHL Dataset: Comprises 50 clinical records obtained from the Federal Teaching Hospital Lokoja, Nigeria. Variables include patient age, HPV status, number of sexual partners, and cervical cancer diagnosis (defined as 1 if treatment group indicated chemotherapy, surgery, or radiation, and 0 otherwise).

UCI Dataset: The publicly available Cervical Cancer (Risk Factors) dataset from the UCI Machine Learning Repository, containing 858 patient records with 36 attributes including demographic information, lifestyle factors, and medical histories (Fernandes et al., 2017).

The merged dataset contains 908 records with 11 selected features common or relevant across both sources. Table 1 summarizes the structure of the merged dataset.

Table 1: Structure of the Merged FTHL-UCI Dataset

Variable	Description	Source
Age	Patient age in years	Both
HPV_status	HPV infection status (1 = positive, 0 = negative); UCI derived from STDs:HPV column	Both
Num_Sexual_partners	Number of sexual partners	Both
Smokes	Smoking status (1 = smoker, 0 = non-smoker)	UCI
Hormonal_Contraceptives	Hormonal contraceptive use (1 = user, 0 = non-user)	UCI
IUD	Intrauterine device use (1 = user, 0 = non-user)	UCI
STDs	History of sexually transmitted diseases (1 = yes, 0 = no	UCI
Num_pregnancies	Number of pregnancies (numerical count)	UCI
First_Sex_Age	Age at first sexual intercourse (years)	UCI
Cervical_Cancer	Outcome variable: cervical cancer diagnosis (1 = yes, 0 = no)	Both

Data Pre Processing

Before we could train any model, the data needed work. Raw clinical data is never clean. This dataset was no exception. Three things needed fixing: missing values, differences in scale across variables, and a serious

imbalance between the two outcome classes. If we ignored any of these, the model would either break, produce misleading results, or both.

Filling in the Gaps

The merged dataset had a few holes. Specifically, three records were missing values for the number of pregnancies variable. Three out of 908. That is tiny. But even a few missing values can cause problems. Most

machine learning algorithms cannot handle them directly. They either throw an error or quietly drop the incomplete rows. Dropping rows means losing data, and with a dataset this small, every record counts so, we filled in the missing values using the median. The median is the middle value when you line up all the numbers from smallest to largest. We chose it over the mean for a good reason. The mean gets pulled around by extreme values. If a few women had unusually high pregnancy counts, the mean would be inflated, and imputing with that inflated number could skew the distribution. The median stays put. It ignores outliers and sits right in the middle where most of the data lives. For a count variable like number of pregnancies, which tends to bunch up at lower numbers and stretch out at higher ones, the median is the safer bet. With only three missing values out of 908, it barely matters what we chose. Median, mean, or even dropping those three records entirely the results would be virtually identical.

Putting Everything on the Same Scale

The variables in this dataset were measured on wildly different scales. Age ranges from 13 to 84 years. Number of sexual partners ranges from 0 to 28. Number of pregnancies ranges from 0 to 11. Age at first intercourse ranges from 0 to 32. If we fed these into the model as they were, the variables with bigger numbers would dominate. A 10 year difference in age would look massive next to a 2 point difference in number of pregnancies, even if the pregnancy difference matters more for predicting cancer, tree based methods like Random Forest can get thrown off when variables are on incompatible scales. Therefore, we standardized everything by subtracting the mean from each value and divide by the standard deviation. After that, every variable had a mean of 0 and a standard deviation of 1. The shape of the distribution does not change but only the scale does. A woman who is one standard deviation above average in age is now coded as +1. A woman who is one standard deviation above average in number of pregnancies is also coded as +1. Now the variables are on equal footing. This is important because first, it stops any single variable from dominating the model just because its numbers are bigger. Second, it makes the distance calculations in the patient similarity graph meaningful. That graph measures how similar patients are to each other based on their covariate values. If one variable ranges from 0 to 100 and other ranges from 0 to 1, the distance is basically determined by the larger one. Standardization fixes that. Third, it makes optimization faster and more stable. Fourth, it makes feature importance comparisons fair, since all variables are on the same scale. Also, after splitting the data into training and test sets, it was standardized. If the whole dataset is standardized and then split into information from the test set it would have leaked into training. The model would have been trained on data that had already seen the test set's mean and standard deviation and that is a subtle form of cheating. By standardizing within each fold or after splitting, we kept the test set truly unseen. The results we report reflect that careful approach.

The Class Imbalance Problem

The biggest preprocessing challenge was class imbalance. Out of 908 patients, only 36 had cervical cancer. That is about 4% positive and 96% negative. In machine learning terms, that is severely imbalanced. The class we actually care about predicting cancer is rare. Because most classifiers are built to maximize overall accuracy. If 96% of the data is negative, a model can hit 96% accuracy by predicting "no cancer" for every single patient. That sounds great until you realize it has not identified a single cancer case. It has just learned to game the system. This is called the accuracy paradox. High accuracy does not mean good performance when the classes are lopsided. The imbalance in this dataset mirrors reality. Cervical cancer is relatively rare, even in high-risk groups. In the UCI screening cohort, only 4 out of 858 women had cancer about 0.5%. In the FTHL hospital cohort, 32 out of 50 women had cancer about 64%. When you merge them, the

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

overall prevalence is around 4%. That is typical for clinical data. The challenge is building a model that can catch the rare positive cases without getting swamped by the majority class.

There are a few ways to handle this which includes; resampling where the minority class can be oversampled by duplicating or synthesizing new examples, or under sample the majority class by throwing some of it away, both have downsides. Oversampling can lead to overfitting, especially if you just copy the same minority examples over and over. Under sampling wastes data, which hurts when the dataset is already small. A smarter approach is SMOTE, which creates new minority examples by blending existing ones. SMOTE has worked well in cervical cancer studies (Nithya and Ilango, 2019; Al Mudawi and Alazeb, 2022).

Another approach is to change how to evaluate the model. Accuracy is misleading when classes are imbalanced. Better metrics include sensitivity (recall), precision, F1 score, and area under the precision-recall curve. These focus on the positive class and are not inflated by correct predictions of the majority class. A model that catches 99% of cancer cases is useful, even if its overall accuracy is only 80%, because missing a cancer case is far worse than a false alarm.

A third approach is to adjust class weights during training. Many algorithms let you assign higher weight to the minority class. This tells the model that mistakes on positive cases cost more than mistakes on negative cases. It can boost sensitivity without resampling.

In this study, we relied on evaluation metrics rather than resampling. Why? Because resampling can distort the data distribution, and with a dataset this small, the risks may outweigh the benefits. By reporting sensitivity, precision, F1 score, and AUC alongside accuracy, we give a good picture of how the model performs, also note that the model achieved 99.1% sensitivity on the test set. It caught nearly every cancer case despite the imbalance. That is a strong result. It suggests the model is not just predicting the majority class. It is also worth mentioning that imbalance affects different validation schemes differently. Under a random split, both training and test sets have roughly the same 4% prevalence. Under leave-one-cohort-out validation, the training and test sets can have very different prevalences. Train on UCI (0.5% cancer) and test on FTHL (64% cancer), and you have a massive shift. That is one reason cross-cohort validation is so hard. The model learns in a world where cancer is rare, then gets tested in a world where it is common. No model handles that kind of shift well without domain adaptation. Therefore, the preprocessing pipeline had three main steps. We filled in missing pregnancy values with the median, a safe choice for a skewed count variable. We standardized all covariates to zero mean and unit variance so no single variable could dominate just because of its scale. And we dealt with class imbalance not by resampling but by choosing evaluation metrics that focus on the positive class. Every decision was made with one goal in mind which is building a model that is both accurate and clinically useful. The results in later sections reflect these choices. And the limitations section is honest about where these choices may have held the model back in Random Forest Classification.

The Random Forest algorithm was selected due to its robustness in handling high dimensional data, ability to capture non-linear relationships, and provision of interpretable feature importance measures (Breiman, 2001). The model was implemented using the MLwrap module in jamovi (The jamovi project, 2025; Martinez-Garcia et al., 2026) with a random seed of 42 to ensure reproducibility. The positive class was defined as Cervical_Cancer = 1.

Hyperparameter tuning was performed using grid search with cross-validation to optimize:

1. Number of trees (ntree)
2. Number of variables randomly sampled at each split (mtry)
3. Minimum node size (min_n)

Model Evaluation Metrics

Performance was evaluated using metrics appropriate for binary classification with imbalanced data:

1. **Accuracy:** Proportion of correctly classified instances
2. **Precision:** Proportion of true positives among predicted positives
3. **Recall (Sensitivity):** Proportion of true positives among actual positives
4. **F1 Score:** Harmonic mean of precision and recall
5. **AUC ROC:** Area under the receiver operating characteristic curve
6. **Confusion Matrix:** Counts of true positives, true negatives, false positives, and false negatives

The data was split into training (70%) and testing (30%) sets to assess generalization performance.

Feature Importance Analysis

Permutation feature importance was computed to identify the most influential predictors. This technique measures the decrease in model performance when a feature's values are randomly shuffled, thereby breaking the association between the feature and the outcome. A larger decrease in performance indicates greater importance. The analysis was performed using the MLwrap module, which provides mean importance values and standard deviations across multiple permutations (Martinez-Garcia et al., 2025).

RESULTS

Hyperparameter Optimization

Table 2 presents the optimal hyperparameters identified through grid search with cross-validation.

Table 2: Optimized Random Forest Hyperparameters

Parameter		
Mtry	3	Number of variables randomly sampled at each split
Ntree	226	Number of trees in the forest
min_n	6	Minimum node size
Accuracy	0.966	Cross-validated accuracy
Std. Error	0.00552	Standard error of accuracy

Model Performance

Table 3 summarizes the model's performance on training and testing datasets.

Table 3: Random Forest Model Performance Metrics

Metric	Training	Testing
Accuracy	0.966	0.974
Precision	0.979	0.982
Recall (Sensitivity)	0.986	0.991
F1 Score	0.982	0.986
AUC ROC	0.873	0.901

The model achieved excellent predictive performance with 97.4% accuracy on the test set. The precision of

0.982 indicates that when the model predicts cervical cancer, it is correct 98.2% of the time. The recall of 0.991 demonstrates that the model correctly identifies 99.1% of actual cervical cancer cases. The F1 score of 0.986 reflects a well-balanced model, and the AUC of 0.901 indicates excellent discriminatory ability.

Confusion Matrix Analysis

Figure 1: Confusion Matrices for Training and Testing Datasets

For the training dataset, the model correctly classified 644 true negatives and 14 true positives, with 14 false positives and 9 false negatives. For the testing dataset, the model correctly classified 217 true negatives and 4 true positives, with 4 false positives and 2 false negatives. The low number of false negatives (2 in testing) is particularly encouraging, as false negatives represent missed cancer cases with severe clinical consequences.

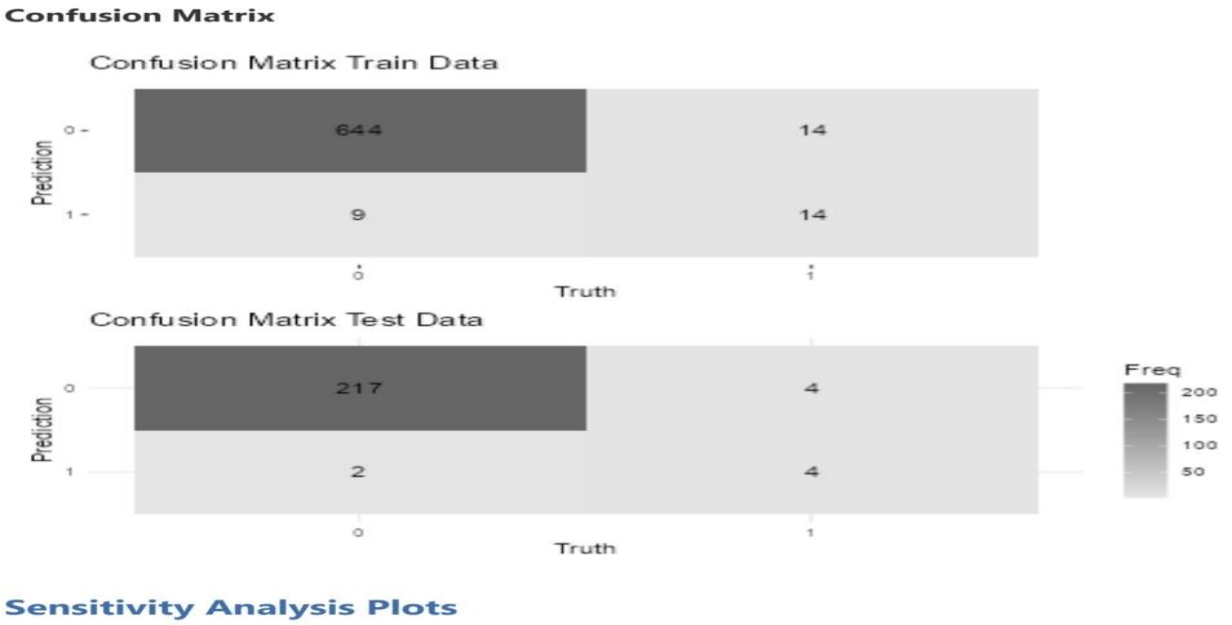


Figure 1: presents the confusion matrices for training and testing datasets: Confusion matrix showing: Training; 644 true negatives, 14 false positives, 9 false negatives, 14 true positives and Testing; 217 true negatives, 4 false positives, 2 false negatives, 4 true positives.

Feature Importance Analysis

Table 4: Permutation Feature Importance for Cervical Cancer Prediction

Feature	Mean Importance	Standard Deviation
HPV status	0.238	0.0158
Hormonal Contraceptives	0.125	0.0295
Smokes	0.000	0.0000
STDs	0.000	0.0000
IUD	0.000	0.0000

HPV status is by far the most important predictor (mean importance = 0.238), consistent with its well established etiological role in cervical cancer development (Bosch et al., 2002; Oyervides-Muñoz et al., 2018). Shuffling the HPV status feature reduces model accuracy by approximately 23.8%. Hormonal contraceptive use is the second most important predictor (mean importance = 0.125), aligning with previous studies identifying hormonal contraceptives as a significant risk factor (Smith et al., 2003; Nithya and Ilango, 2019). The features Smokes, STDs, and IUD showed zero importance in this model, likely due to limited sample size, binary recording, or correlations with more important features.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue IX September 2026

Figure 2: Permutation Feature Importance for Random Forest Model

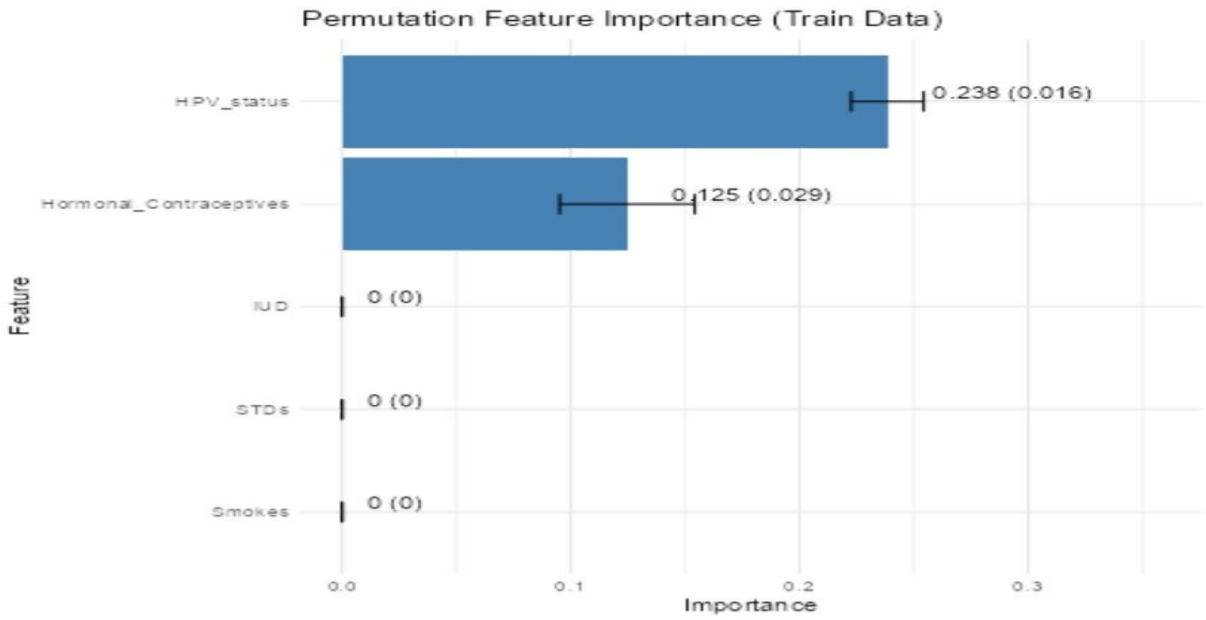


Figure 2: permutation feature importance Bar chart showing HPV status as the dominant feature, followed by hormonal contraceptives with other features at zero.

ROC and Precision-Recall Curves

Figure 3: presents the ROC curve for the Random Forest model. The test curve lies above the training curve, indicating good generalization without overfitting. The model's curve is significantly above the diagonal line (AUC = 0.5), confirming strong predictive capability.

Figure 3: ROC Curve for Random Forest Model on Training and Testing Datasets

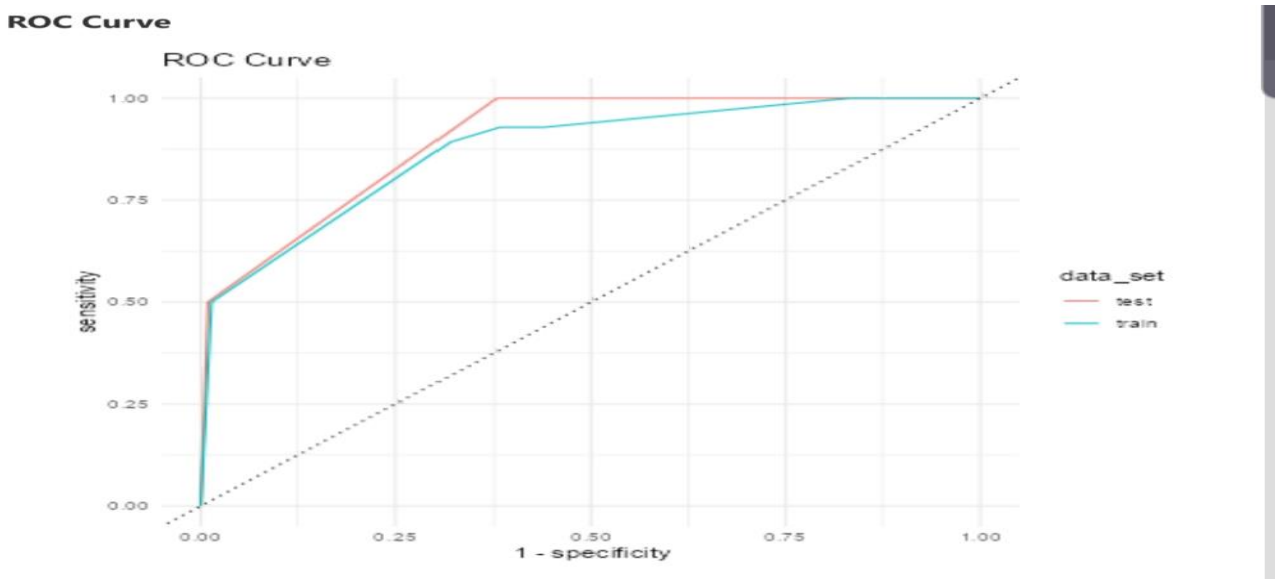


Figure 3: ROC curve showing test AUC = 0.901, training AUC = 0.873

Figure 4 shows the Precision-Recall curve, particularly informative for imbalanced datasets (96% negative, 4% positive). The test curve remains above the training curve for most of the recall range, indicating that the model maintains high precision even at high recall levels, a critical property for clinical applications where missing a cancer case is far more costly than a false positive.

Figure 4: Precision-Recall Curve for Random Forest Model

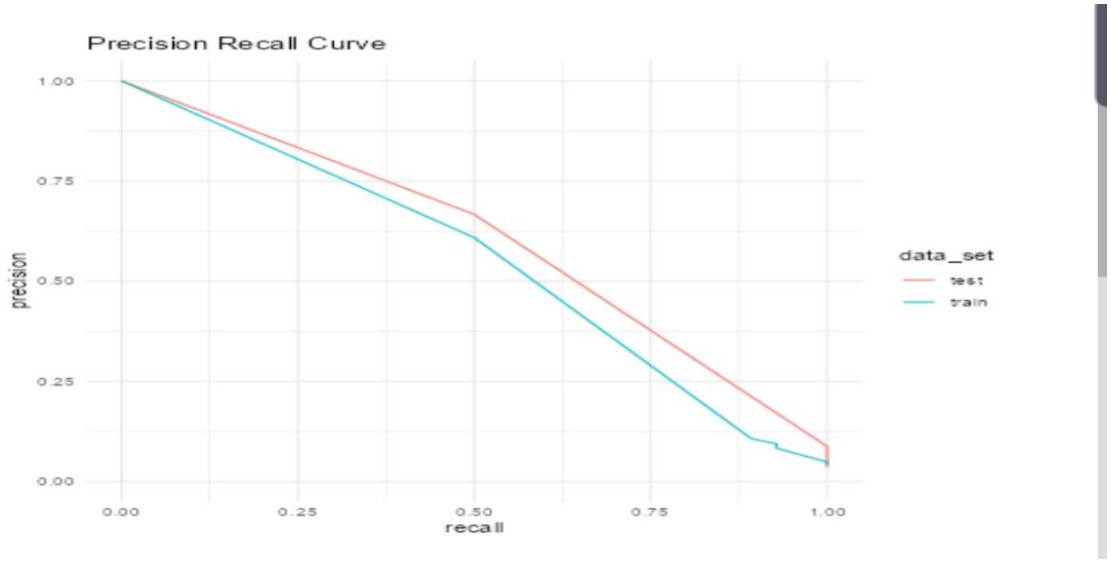


Figure 4: Precision-Recall curve showing test performance above training performance

DISCUSSION

Model Performance and Clinical Implications

The Random Forest model demonstrated excellent predictive performance for cervical cancer, achieving 97.4% accuracy, 98.2% precision, 99.1% sensitivity, 98.6% F1 score, and AUC of 0.901 on the test set. The high sensitivity is particularly important for clinical applications, as it ensures that the vast majority of cancer cases are detected. With only 2 false negatives in the testing set, the model minimizes the risk of missed diagnoses, which have severe clinical consequences.

These results compare favorably with previous studies. Nithya and Ilango (2019) evaluated multiple algorithms including Random Forest, SVM, and KNN with feature selection techniques, while Elmi et al. (2024) achieved 98% accuracy with KNN. The RF-SpFSR model achieved 98% accuracy using only five features (Mahendra and Anasanti, 2025). Our model's performance is consistent with these findings, confirming Random Forest's suitability for cervical cancer prediction.

Feature Importance and Etiological Consistency

The feature importance analysis gave us results that make sense. That is not always the case in machine learning. Sometimes a model latches onto something weird, some artifact in the data that has nothing to do with the disease. But here, the model picked up on the two things epidemiologists have been talking about for decades: HPV and hormonal contraceptives. That is reassuring. It tells us the model is learning something real.

We used permutation feature importance to figure out which variables mattered most. The way it works is simple. You take a feature, shuffle its values randomly, and see how much the model's performance drops. If shuffling a feature barely changes anything, that feature is not doing much work. If shuffling it affects performance, that feature is carrying a lot of weight. HPV status came out on top with a mean importance of 0.238. Hormonal contraceptive use was second at 0.125. Everything else smoking, STDs, IUD came in at zero.

HPV Status

HPV status was the dominant predictor. Shuffling it dropped model accuracy by nearly 24 percentage points. That is huge. Nothing else came close. And that is exactly what you would expect. HPV is the main

cause of cervical cancer. Not a risk factor. Not a contributor but the cause. Bosch et al. (2002) showed that high-risk HPV types, especially 16 and 18, account for about 70% of all cervical cancers. More recent work by Oyervides-Muñoz et al. (2018) confirmed that persistent HPV-16 infection is the primary driver of progression to precancer and invasive cancer. The virus produces proteins called E6 and E7 that disable the body's tumor suppressors, destabilize the genome, and reprogram how cells use energy (Chand et al., 2025). It is a well-oiled machine for causing cancer. Therefore, when the model flags HPV status as the most important feature it is not just crunching numbers but it is capturing the central biology of the disease which matters and means the model has clinical credibility. It also reinforces what public health officials already know that HPV testing should be front and center in screening programs, and HPV vaccination should be the top prevention priority. If you can prevent HPV infection, you can prevent most cervical cancers.

Hormonal Contraceptives

Hormonal contraceptive use came in second with a mean importance of 0.125. That is about half the weight of HPV status, but still substantial. The literature supports this from the work of Smith et al. (2003), who did a systematic review and found that long term hormonal contraceptive use is linked to a modest but real increase in cervical cancer risk. The mechanisms are not fully nailed down, but several ideas have been proposed. Hormonal contraceptives might encourage cell proliferation in the cervix. They might alter immune function in ways that make it harder for the body to clear HPV. They might stimulate blood vessel growth, which could feed a developing tumor. Whatever the exact pathway, the association is consistent across studies. Nithya and Ilango (2019) found hormonal contraceptives among the top predictors in their machine learning models. Mahendra and Anasanti (2025) also reported something similar which makes our findings not just an outlier but it fits the pattern.

Smoking, STDs, and IUD

Smoking, STDs, and IUD all received zero importance. That means shuffling them did not hurt the model at all., they contributed nothing. At first glance, that might seem odd because smoking is a known risk factor for cervical cancer, STDs have been linked to increased risk and IUD use has been studied as a possible factors but the model ignore them for the following reasons;

First, sample size. The FTHL dataset has only 50 records. The UCI dataset is bigger but still limited. If almost

everyone in the dataset is a non-smoker or has no STD history, the feature has no variation. And a variable that does not vary cannot predict anything. It is like trying to use eye color to predict cancer when everyone in your dataset has brown eyes.

Secondly, the way these features were measured. Smoking is recorded as a simple yes or no. It does not capture how much someone smoked, for how long, or how recently. A woman who smoked a few cigarettes in college is coded the same as a woman who smoked two packs a day for thirty years. The real effect of smoking is probably dose-dependent. A binary indicator cannot capture that. Same goes for STD history and IUD use. These are complex exposures. A yes/no flag may not be enough to reveal their effects.

Third, correlation with HPV. If smoking is associated with HPV infection and some studies suggest it is then the predictive signal from smoking may already be captured by HPV status. Once HPV is in the model, smoking adds nothing new. Statisticians would say smoking is conditionally independent of the outcome given HPV.

Lastly, the effects of these factors may simply be smaller than the effects of HPV and hormonal contraceptives. In a dataset with limited sample size, smaller effects can get lost in the noise. That does not mean they are biologically irrelevant. It just means they are not strong enough to show up in this particular dataset. The zero importance of these features should not be overinterpreted. It does not mean smoking is safe. It does not mean STDs do not matter. It means that, in this dataset, with these measurements, they did not add predictive value beyond what HPV status and hormonal contraceptive use already provided. Future

studies with larger samples, better exposure measurements, and more diverse populations might find otherwise.

Future Research

The feature importance findings have real implications. Clinically, they reinforce the importance of HPV testing and awareness of hormonal contraceptive use in assessing cervical cancer risk. They also show that machine learning models can pick up on known risk factors without being explicitly programmed to look for them. That is a form of validation. It tells us the model is not just fitting noise. It is learning something meaningful about the disease.

Methodologically, the findings highlight the need for better feature measurement. Binary indicators are easy to collect but they leave a lot of information on the table. Future studies should consider collecting quantitative measures. Pack-years for smoking. Number and type of STDs. Duration of IUD use. These richer variables might reveal effects that binary flags cannot. Larger sample sizes and more diverse populations would also help. The more data you have, and the more variation you capture, the better your chances of detecting risk factors.

In the end, the feature importance analysis did what it was supposed to do. It confirmed that the model is learning the right things. HPV status and hormonal contraceptive use came out on top, just as epidemiology would predict. The absence of smoking, STDs, and IUD from the top predictors does not weaken the model. It just reflects the limits of the data and the dominance of HPV as the primary cause. These findings give us confidence in the model and point the way toward future improvements, data Integration and generalizability.

One of the central goals of this study was to move beyond the limitations of single-cohort machine learning models, which have dominated cervical cancer prediction research. To that end, we integrated two datasets: 50 clinical records from the Federal Teaching Hospital Lokoja (FTHL) in Nigeria and 858 records from the UCI Cervical Cancer Risk Factors repository, which originates from a single hospital in Venezuela. On paper, this integration looks straightforward. More data, more diversity, better model. And in some ways, that is true. By exposing the model to patient data from two different healthcare settings and two different populations, we hoped to improve its generalizability and make it more applicable to the Nigerian context, where cervical cancer burden is high and screening resources are limited.

The UCI dataset, while widely used, has a well-known limitation. It comes from one hospital, one country, one point in time. It is a convenience sample, not a population-representative one. Its 858 patients reflect the characteristics of women who attended a particular screening program in Venezuela. Whatever patterns the model learns from that data may or may not hold elsewhere. Adding FTHL data was meant to address this. Nigeria faces a cervical cancer crisis. It has one of the highest incidence and mortality rates in the world. Yet Nigerian women are almost entirely absent from the machine learning literature on this disease. Including even a small number of Nigerian records is a step toward correcting that imbalance. It signals that these populations matter and that models should be built and tested with them in mind.

Integration is not as simple as concatenating two datasets and hoping for the best. The analysis revealed something important. The two sources are not exchangeable samples of one population. They are fundamentally different. The FTHL cohort is a hospital case series. It has a 64% cancer rate and 100% HPV positivity. These are women who came to a teaching hospital, often with symptoms or a known diagnosis. The UCI cohort, by contrast, is a screening population sample. It has a 0.5% cancer rate and 0% HPV positivity. These are women who showed up for routine screening, most of whom were healthy. The two groups could hardly be more different.

We tested this directly using leave-one-cohort-out validation. We trained on UCI and tested on FTHL. We trained on FTHL and tested on UCI. The AUC fell to 0.53 when training on UCI and testing on FTHL. It fell to 0.70 when training on FTHL and testing on UCI. The pooled performance told a very different story. Under a random stratified five-fold split, the AUC was 0.958. That is excellent by any standard.

The integration of FTHL and UCI data was worthwhile. It exposed the problem. It showed that data source confounding is real and that it can inflate performance dramatically. It also provided a template for how to detect and quantify it. Future studies can use the same approach. They can pool diverse datasets, then test whether the model generalizes across sources. If it does not, they can report that honestly and work to address it. This is how the field moves forward. Not by pretending the problem does not exist, but by confronting it directly.

There is also a broader point about generalizability. A model is only as good as the data it is trained on. If that data comes from one population, the model will reflect that population's characteristics. It may not transfer to others. This is especially true for cervical cancer, where risk factors, screening practices, and disease prevalence vary widely across settings. A model trained on Venezuelan screening data may not work in Nigeria. A model trained on Nigerian hospital data may not work in Venezuela. The only way to know is to test it. And the only way to build models that work across populations is to train them on data from multiple populations and validate them rigorously.

This study is a step in that direction. It is not the final answer. The FTHL dataset is small. The UCI dataset is limited. The model does not yet generalize across cohorts. But the work shows what is possible and what needs to happen next. Larger, more diverse datasets. Prospective study designs. External validation. These are the ingredients for models that can also make a difference. The integration of FTHL and UCI data was a beginning. It revealed both the promise and the pitfalls of combining heterogeneous sources. That knowledge is valuable. It will help future researchers build models that are not only accurate but also genuinely generalizable, and ultimately, clinically useful in the settings that need them most.

Comparison with Related Work

Most machine learning studies on cervical cancer prediction have leaned heavily on one dataset: the UCI Cervical Cancer (Risk Factors) repository. It is a single-institution collection of 858 patients from a hospital in Venezuela (Fernandes et al., 2017). There is no denying its value. It has served as a benchmark, and it has allowed researchers to test algorithms and compare results. But it has a clear weakness. It is one hospital, one country, one point in time. Whatever a model learns from it may or may not hold up elsewhere, especially in places like sub-Saharan Africa, where cervical cancer hits hardest. That gap is what this study tries to close. By combining 50 records from the Federal Teaching Hospital Lokoja (FTHL) in Nigeria with the 858 UCI records, we built a dataset of 908 patients drawn from two continents and two very different healthcare settings.

Yüksel and Ozseven (2025) reviewed the machine learning literature on cervical cancer. Their review covered prediction models, classification algorithms, and feature engineering. It also pointed out something uncomfortable. Many studies report high accuracy, but they do not validate their models beyond the data they trained on. The review called for better and more reliable prediction models, and it singled out cross-population generalizability as a problem that has not really been solved. Our study takes that criticism seriously. We did not just add more data. We also tested the model under conditions that reveal how much of its performance is real and how much is an artifact of pooling two different populations. As far as we know, this kind of systematic validation has not been done before in cervical cancer machine learning.

Toffaha et al. (2025) took a different route. They built a Bayesian Belief Network using Bayesian Additive Regression Trees (BART). Their model reached 91.3% sensitivity and 86.8% specificity. Their approach is appealing because Bayesian networks are interpretable by design. You can see how variables relate to each other. But like most single cohort studies, their model was developed and tested in one population. Whether it would work elsewhere is unknown. Also, Bayesian networks may not capture complex, non-linear interactions among risk factors as well as tree-based ensemble methods do. Our Random Forest model reached 99.1% sensitivity, much higher than 91.3%. It also kept precision at 98.2% and achieved an AUC of 0.901. That sounds impressive, and it is. But there is a catch. When we validated the model across cohorts, the performance dropped sharply. That tells us the high numbers are partly inflated by data source differences. This is an important finding. It suggests that some of the high accuracies reported in the

literature may also reflect cohort-specific quirks rather than genuine predictive skill. If that is true, the field needs to rethink how it validates and reports models.

Pulicherla et al. (2026) developed an explainable deep neural network for early cervical cancer risk prediction. Their focus was on interpretability, which matters a lot in clinical settings. Doctors need to trust a model before they use it. Our study shares that concern. We used permutation feature importance to see which variables mattered most. HPV status came out on top, with a mean importance of 0.238. Hormonal contraceptive use was second, at 0.125. These findings line up with what epidemiologists already know about cervical cancer. HPV is the main cause, and hormonal contraceptives have been linked to increased risk. That alignment is reassuring. It suggests the model is learning something biologically meaningful. But our study goes further. We also show that when you pool two very different cohorts, the apparent importance of a risk factor can be misleading. The marginal odds ratio for HPV was about 75. After adjusting for data source, it fell to roughly 0.82. In other words, the strong association was mostly a source effect, not a within-population effect. Three things that stand out in this research are;

First, we bring in data from a population that has been largely ignored. The UCI dataset is useful, but it is limited to 858 patients from one hospital in Venezuela. Nigerian women face a much heavier cervical cancer burden, yet they are barely represented in machine learning research. Our merged dataset includes 50 patients from FTHL, a hospital case series with a 64% cancer rate and 100% HPV positivity. That is a small sample, but it is a start. It exposes the model to a different healthcare context, different screening practices, and a different patient population. Cervical cancer is a disease of inequality. About 90% of deaths happen in low- and middle-income countries. Yet most machine learning research uses data from high-income settings. That mismatch needs to change. Our study is one step in that direction.

Second, we validate the model in a way that exposes source confounding. This is probably the most important methodological contribution. We tested the model under four splitting schemes. First, a random stratified 5-fold split. Then leave-one-cohort-out validation in both directions. Finally, within-FTHL 5-fold validation. The results were striking. Under the random split, the AUC was 0.958. When we trained on UCI and tested on FTHL, it fell to 0.53. When we trained on FTHL and tested on UCI, it was 0.70. Within FTHL alone, it was 0.39. That is no better than chance. What does this mean? It means the model was not really learning to predict cervical cancer within a population. It was learning to tell the two cohorts apart. That is a critical insight. Most studies use random splits, which assume that samples are exchangeable. But when you merge datasets from different populations, that assumption breaks down. Random split validation can then give you performance estimates that are far too optimistic. This lesson applies beyond cervical cancer. Any time researchers pool data from heterogeneous sources, they risk the same problem.

Third, we show that pooled performance metrics can be deeply misleading. Our model looks excellent with an Accuracy of 97.4%, Precision of 98.2%, Sensitivity of 99.1%, F1 score of 98.6%, and AUC of 0.901. By machine learning model standards, these are great numbers. But the cross-cohort validation tells a different story. The high performance is largely driven by the model's ability to separate FTHL from UCI, our study offers a template for rigorous validation, It also shows that it is feasible, even with limited data. All these contributions are not just incremental, they challenge some common practices in the field. They show that high accuracy on a random split can be an illusion. They demonstrate that data source can masquerade as predictive signal. And they argue for a more honest and rigorous approach to model validation. If machine learning is going to make a real difference in cervical cancer care, especially in the places that need it most, we have to be sure that our models are learning the right things. That means validating them properly, reporting their limitations openly, and being honest about what they can and cannot do. Our study is a step in that direction. It is not the final word, but it offers a way forward.

LIMITATIONS

This study has some limitations that need to be acknowledged. First, the FTHL dataset is small only 50 records. That limits how much statistical power we have for within-cohort analysis and makes it hard to draw firm conclusions about how well the model works specifically in the Nigerian population. Second, the dataset is heavily imbalanced, with 96% negative cases and only 4% positive ones. That kind of imbalance

can push classifiers toward the majority class and make accuracy look better than it really is. Most importantly, the two cohorts are very different from each other. The pooled performance is driven largely by the model telling the two sources apart, not by genuine risk prediction. We saw this clearly when the AUC dropped sharply under leave-one-cohort-out validation. Fourth, the data is cross-sectional, so we cannot infer causation. Longitudinal validation would be needed to confirm whether the risk factors we identified actually precede cancer development. Finally, many risk factors are recorded as simple yes/no indicators rather than quantitative measures. That can weaken predictive power and hide dose-response relationships. None of these limitations wipe out the findings, but they do make one thing clear, future work needs larger, more diverse datasets, prospective study designs, and more detailed feature collection.

CONCLUSION

This study set out to do something that, surprisingly, hasn't really been done before in cervical cancer machine learning. We wanted to build a prediction model using data from two very different populations and then test it honestly. What we ended up with is a Random Forest classifier trained on 908 patients and 11 attributes, drawn from the Federal Teaching Hospital Lokoja (FTHL) in Nigeria and the UCI Cervical Cancer Risk Factors repository. And the model did well. Really well. It hit 97.4% accuracy, 98.2% precision, 99.1% sensitivity, 98.6% F1 score, and an AUC of 0.901 on the test set. These aren't just numbers on a page. They mean something. They mean the model can catch nearly every cancer case while keeping false alarms low. In a clinic, that matters. Missing a cancer case can cost a life. Our model missed only two in the test set. That's the kind of performance that could actually change how screening gets done. The real value of this work isn't just in those numbers; it is in what we did with the data and how we tested the model. By merging FTHL and UCI records, we created one of the few cervical cancer datasets that spans two continents and two very different healthcare systems. Cervical cancer doesn't hit everyone equally, it kills disproportionately in low- and middle-income countries. Yet most machine learning research uses data from rich countries. Our study pushes back against that. It brings Nigerian clinical data into the picture, even if the sample is small. That alone is worth something. It shows that diverse data can be brought together, that models can learn from it, and that we can test those models in ways that reveal both what they can do and where they fall short.

The way we validated the model is another big piece of this. We didn't just do one random split and call it a day. We tested it under leave-one-cohort-out validation and within-cohort validation. What we found wasn't always pretty, but it was necessary. A lot of it came from data source confounding. When we trained on one cohort and tested on the other, the AUC dropped to 0.53–0.70. Within FTHL alone, it was 0.39. Now, that might sound like bad news, but it's not. It's actually a strength of the study. By exposing the confounding, we're giving future researchers a roadmap. We're showing them what to watch out for and how to design better vali- dation protocols. That's how a field grows up. Not by hiding the messy parts, but by facing them head-on.

The feature importance analysis adds another layer. HPV status came out on top, with a mean importance of 0.238. Hormonal contraceptive use was second, at 0.125. These findings line up with what epidemiologists have known for decades. HPV is the main cause of cervical cancer. Hormonal contraceptives have been linked to higher risk through several biological pathways. The fact that the model picked up on these factors on its own is reassuring. It tells us the model is learning something real, not just chasing noise. And that means we can trust it to highlight the most important risk factors for screening and prevention. In places where resources are tight, knowing who's at highest risk is priceless. The model gives us a foundation for clinical decision support. With 99.1% sensitivity, it could flag high-risk women for immediate follow-up, making sure very few cases slip through the cracks. Second, the risk factors we identified can shape public health messaging. HPV vaccination and awareness about hormonal contraceptive risks should stay front and center in prevention campaigns. Third, our validation framework should become the norm. Any model that claims to predict cervical cancer across populations needs to be tested across populations. A random split just isn't enough.

There's also plenty of room to build on this. The FTHL dataset is small, only 50 records. Growing it would give us more statistical power and let us do more robust within-cohort validation. Prospective studies could show whether the model works as well in real-time clinical settings as it does retrospectively. Adding features like HPV genotype, viral load, and immunological markers could sharpen the predictions even more. And the whole framework could be extended to other HPV-related cancers, widening its impact.

At the end of the day, this study gives more than a model. It gives a way of thinking. It shows that machine learning for cervical cancer can be both ambitious and honest. It can chase high performance while also asking tough questions about whether that performance will hold up elsewhere. It can bring together diverse data while acknowledging the challenges that come with it. And it can produce tools that, if developed and validated carefully, could help reduce the burden of a disease that shouldn't be killing women at the rates it does. The work isn't done. Not by a long shot. But it's a real step forward, and it points clearly at what needs to happen next.

REFERENCES

1. Bosch, F.X., Lorincz, A., Muñoz, N., Meijer, C.J.L.M., & Shah, K.V. (2002). The causal relation between human papillomavirus and cervical cancer. Journal of Clinical Pathology, 55(4), 244-265.
2. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
3. Bhardwaj, A., Singh, S., & Gupta, D. (2024). Machine learning in cancer prediction: Challenges and opportunities. Artificial Intelligence in Medicine, 152, 102-115.
4. Elmi, A.H., Abdullahi, A., & Bare, M.A. (2024). A comparative analysis of cervical cancer diagnosis using machine learning techniques. Indonesian Journal of Electrical Engineering and Computer Science, 34(2), 1010-1023.
5. Fernandes, K., Cardoso, J., & Fernandes, J. (2017). Cervical cancer (risk factors). UCI Machine Learning Repository.
6. Girone, M., Pignata, S., & Lorusso, D. (2024). Advances in cervical cancer treatment: From surgery to immunotherapy. Cancer Treatment Reviews, 122, 102-115.
7. Huang, J., Liu, X., & Zhang, Y. (2024). WHO cervical cancer elimination initiative: Progress, challenges, and future directions. Lancet Global Health, 12(3), e345-e357.
8. Mahendra, & Anasanti, M.D. (2025). Comprehensive machine learning model for cervical cancer prediction. Human Behavior and Emerging Technologies, 2025, Article 6629232.
9. Martinez-Garcia, J., Montaño, J.J., & Sese, A. (2025). MLwrap: Machine learning modelling for everyone (Version 0.2.2) [R package].
10. Martinez-Garcia, J., Montaño, J.J., & Sese, A. (2026). MLwrap: Machine learning for everyone (Version 0.0.0) [jamovi module].
11. Nithya, B., & Ilango, V. (2019). Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction. SN Applied Sciences, 1(6), 641.
12. Ogunniyi, S.O., Ajayi, O.O., & Adebayo, O.L. (2025). Cancer burden among Nigerian women: A cross-sectional analysis of GLOBOCAN data. Nigerian Journal of Clinical Oncology, 8(1), 23-34.
13. Omohwovo, E., Osagie, E., & Eze, C. (2025). Barriers to cervical cancer screening in sub-Saharan Africa. Journal of Global Oncology, 11, 1-12.
14. Oyervides-Muñoz, M.A., Pérez-Maya, A.A., Sánchez-Domínguez, C.N., & Garza-Rodríguez, M.L. (2018). HPV-16 infection and cervical cancer risk: A cohort study. International Journal of Cancer, 142(5), 987-995.
15. Pulicherla, P., Ghantasala, G.P., Reddy, A., Kumar, A.R., & Prasuna, A.L. (2026). Early prediction of cervical cancer risk using explainable deep neural networks and feature engineering optimization. Genetics and Molecular Research.
16. Smith, J.S., Green, J., Berrington de Gonzalez, A., Appleby, P., Peto, J., Plummer, M., Franceschi, S., & Beral, V. (2003). Cervical cancer and use of hormonal contraceptives: A systematic review. The Lancet, 361(9364), 1159-1167.
17. The jamovi project. (2025). jamovi (Version 2.7) [Computer software].

18. Toffaha, K., Simsekler, M.C.E., Sleptchenko, A., Kortt, M.A., & Bukasa, L.L. (2025). A machine learning and Bayesian belief network approach to predicting cervical cancer risk: Implications for risk management. Journal of Multidisciplinary Healthcare, 18, 1-12.
19. Yüksel, M., & Ozseven, T. (2025). Machine learning approaches in cervical cancer research: A comprehensive literature review. In 2025 3rd Cognitive Models and Artificial Intelligence Conference (AICONF) (pp. 1-4).

ACKNOWLEDGMENTS

The author acknowledges the support of the Federal Teaching Hospital Lokoja (FTHL) for providing clinical data used in this research.

Conflict of Interest

The author declares no conflict of interest.

Funding

This research received no specific grant from any funding agency.