Pooja K, 2026, 14:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

An Explainable AI-Based Cervical Cancer Risk Prediction System Using Recursive Feature Elimination and XGboost

Pooja K
M.Sc Data Analytics, Department of Computer Science, Sri Ramakrishna College of Arts & Science for Women, Coimbatore, Nadu, India

Abstract- Cervical cancer remains one of the leading causes of cancer-related death among women, particularly in regions where organised screening is scarce. Machine learning can support early risk assessment, but many published models report accuracy alone, act as opaque "black boxes", and are rarely delivered as usable tools. This study presents an explainable machine learning framework and web application for cervical cancer risk prediction that combines Recursive Feature Elimination (RFE) with the Extreme Gradient Boosting (XGBoost) algorithm. The pipeline uses the publicly available Cervical Cancer Risk Factors dataset (858 patients, 36 attributes, 6.4% biopsy-positive), median imputation, two engineered features (Lifetime Exposure Index and Cancer Load Score), a frequency-based feature filter, Synthetic Minority Over-sampling (SMOTE) applied to the training partition only, and RFE with a Random Forest estimator to retain 18 predictors. On a stratified hold-out set of 172 patients the model achieved an accuracy of 96.51%, specificity of 98.76%, recall of 0.636, precision of 0.778, ROC-AUC of 0.959 and a Matthews correlation coefficient of 0.686. Because the positive class is small, a leakage-free repeated stratified cross-validation (5 folds × 3 repeats, with SMOTE and RFE nested inside each fold) was also performed and is reported alongside Logistic Regression, Random Forest and XGBoost without RFE. SHAP analysis shows that the Schiller and Hinselmann colposcopic test results dominate the prediction, followed by reproductive history and the engineered exposure index. The model is deployed in a Streamlit application that returns a probability-based Low, Medium or High risk category, visual explanations, an analytics dashboard and a downloadable PDF report. The system is intended for decision support and awareness, not for diagnosis.

Keywords- Cervical Cancer, Risk Prediction, XGBoost, Recursive Feature Elimination, SMOTE, Explainable AI, SHAP, Streamlit

I. INTRODUCTION

Cervical cancer is the fourth most commonly diagnosed cancer in women worldwide, with an estimated 604,000 new cases and 342,000 deaths in 2020, the large majority occurring in low- and middle-income countries [1]. Persistent infection with high-risk types of human papillomavirus (HPV) is the necessary cause of the disease, and the long pre-invasive phase makes it one of the few cancers that can be prevented through vaccination and detected early through screening [2]. The World Health Organization's global strategy therefore sets targets for HPV vaccination, screening and treatment of pre-cancer and cancer, so that cervical cancer can be eliminated as a public health problem [2].

In practice, screening coverage remains uneven. Pap smear cytology, HPV testing and colposcopy require laboratory capacity, trained staff and repeated visits, and are often unavailable in rural and under-resourced settings. Women who do not recognise their own risk factors, such as early sexual debut, multiple partners, long-term hormonal contraceptive use, smoking and prior sexually transmitted infections, are less likely to seek screening. A low-cost tool that converts routinely collected health information into an understandable risk estimate could help prioritise who should be referred for confirmatory testing.

Machine learning has been applied to this problem using the Cervical Cancer Risk Factors dataset released by Fernandes et al. [3], with Support Vector Machines, Random Forests, decision trees and

© 2026 Pooja K. his is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

ensemble methods reporting high accuracy [4]. Three limitations recur in this literature. First, the dataset is strongly imbalanced, so accuracy is dominated by the majority class and can hide poor detection of true cases. Second, many studies use all available attributes without a principled feature-selection step, which increases model complexity and obscures which factors matter.

Third, high-performing models are usually "black boxes" whose predictions cannot be explained to clinicians or patients [10], and they are seldom packaged as an application that a non-specialist can use.

This work addresses these gaps with an end-to-end, explainable pipeline. The main contributions are:

- A reproducible preprocessing and feature-engineering workflow for the cervical cancer dataset, including two domain-motivated composite features, the **Lifetime Exposure Index** and the **Cancer Load Score**.
- A classification model that combines SMOTE balancing, RFE-based feature selection and XGBoost [5], with SMOTE applied only to the training partition to avoid information leakage.
- An evaluation that goes beyond accuracy, reporting the confusion matrix, precision, recall, specificity, F1-score, ROC-AUC, average precision and Matthews correlation coefficient, and a leakage-free repeated cross-validation against three baseline models.
- Model explanation using XGBoost importance and SHAP values [9], so that the contribution of each predictor to the risk score can be inspected.
- A deployed Streamlit web application with authentication, risk categorisation, an analytics dashboard and automatic PDF report generation.

The remainder of the paper is organised as follows. Section 2 reviews related work. Section 3 describes the proposed methodology and system. Section 4 presents and discusses the results, including limitations. Section 5 concludes the paper and outlines future work.

II. LITERATURE REVIEW

1. Machine learning for cervical cancer prediction

Early computational work on cervical cancer risk used classical classifiers on clinical and behavioural attributes. Wu and Zhou applied Support Vector Machine-based approaches to the UCI risk-factor data and showed that a reduced feature set could retain strong discrimination while lowering model complexity [4]. Subsequent studies compared Logistic Regression, Decision Trees, k-Nearest Neighbours, Random Forests and gradient boosting, generally finding that tree ensembles handle the mixed binary and continuous attributes and the many missing values better than linear models. The original data were collected at Hospital Universitario de Caracas, and Fernandes et al. used them to study transfer learning under partial observability, where some patients decline to answer sensitive questions [3].

2. Feature selection

Medical datasets often contain redundant or weakly informative attributes. Recursive Feature Elimination, introduced by Guyon et al. for gene selection in cancer classification, repeatedly trains a model, ranks features by importance and discards the weakest until a target number remains [6]. When the ranking model is a Random Forest [8], RFE can capture non-linear effects and interactions. Metaheuristic alternatives such as Particle Swarm Optimization [18] can also search the feature space but require careful parameter tuning and may converge prematurely, which makes RFE an attractive, simpler and more reproducible option.

3. Class imbalance

In screening data the positive class is rare. The Synthetic Minority Over-sampling Technique (SMOTE) creates new minority samples by interpolating between a minority instance and its nearest minority neighbours [7]. SMOTE must be applied only to training data; if synthetic samples are generated before the data are split or before cross-validation, near-copies of the same patient can appear in both training and validation folds, producing optimistic estimates [14]. Imbalance also changes how performance should be reported: the

precision–recall curve is more informative than the ROC curve when positives are rare [13], and the Matthews correlation coefficient is more reliable than accuracy or F1-score for binary classification on imbalanced data [17].

4. Gradient boosting and explainability

XGBoost implements regularised gradient-boosted decision trees with sparsity-aware split finding and built-in handling of missing values, and has become a standard method for structured tabular data [5]. Its accuracy comes at the cost of interpretability, which is a serious concern in healthcare where clinicians must understand and trust a prediction [10]. Model-agnostic explanation methods such as LIME [16] and SHAP [9] attribute a prediction to individual features; SHAP uses Shapley values from cooperative game theory and has an efficient exact algorithm for tree ensembles.

5. Research gap

Despite this progress, existing studies tend to (i) report accuracy without accounting for imbalance or leakage, (ii) omit an explicit feature-selection stage, (iii) offer little explanation of individual predictions, and (iv) stop at an experimental notebook rather than an interactive tool. The present work combines RFE, SMOTE, XGBoost and SHAP within a single pipeline, evaluates it with leakage-free validation, and deploys it as a web application with report generation.

III. PROPOSED WORK

The proposed Cervical Cancer Risk Intelligence System converts patient attributes into a probability, maps it to a Low, Medium or High category and explains the contributing factors. Figure 1 summarises the workflow. The upper row shows data preparation up to the stratified split; the lower row shows model construction and delivery.



Figure 1: Overall workflow of the proposed system

1. Dataset

The study uses the Cervical Cancer Risk Factors dataset available from the UCI Machine Learning Repository and Kaggle [3]. It contains 858 patient records and 36 attributes covering demographics, sexual and reproductive history, smoking, hormonal contraceptive and intrauterine device (IUD) use, sexually transmitted infection (STI) history, prior diagnoses and the results of three screening tests (Hinselmann colposcopy, Schiller test and cytology). The target variable is the biopsy result (0 = no cancer, 1 = cancer). Of the 858 patients, 803 (93.6%) are biopsy-negative and 55 (6.4%) are biopsy-positive, a class ratio of about 14.6 : 1 (Figure 2). The mean age is 26.8 years (SD 8.5; range 13–84).

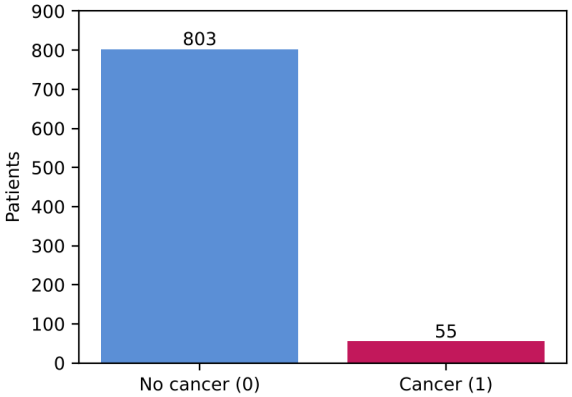


Figure 2: Class distribution of the biopsy target variable

Table 1: Main attribute groups in the dataset

Group	Attributes	Type
Demographic	Age	Integer
Sexual history	Number of sexual partners; age at first sexual intercourse	Integer
Reproductive	Number of pregnancies	Integer
Smoking	Smokes; Smokes (years); Smokes (packs/year)	Binary / numeric

Group	Attributes	Type
Hormonal & IUD	Hormonal contraceptives (yes/no, years); IUD (yes/no, years)	Binary / numeric
STI history	STDs; STDs (number); 12 specific STD indicators; number of diagnoses; time since first/last diagnosis	Binary / numeric
Prior diagnosis	Dx:Cancer; Dx:CIN; Dx:HPV; Dx	Binary
Screening tests	Hinselmann; Schiller; Citology	Binary
Target	Biopsy	Binary (0/1)

2. Data preprocessing

Missing values are encoded in the raw file as the character "?". These were converted to NaN and every column was coerced to a numeric type. Missingness is substantial for several attributes (Figure 3): 105 to 117 patients lack hormonal contraceptive, IUD or STI information, and the two "time since STD diagnosis" attributes are missing for 787 of 858 patients because most patients have never had an STI diagnosis. Remaining gaps were filled with the column median using a median imputer, which is robust to the skewed distributions of variables such as years of smoking. Records with a missing target were removed (none were present in this file). The data also contain 23 exact duplicate rows, which were retained in the experiments reported here and are discussed as a limitation in Section 4.8.

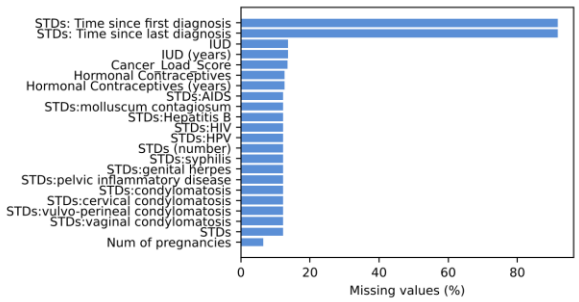


Figure 3: Percentage of missing values for attributes with at least 50 missing entries

3. Feature engineering

Two composite features were derived to capture cumulative exposure, which individual attributes describe only partly:

- **Lifetime Exposure Index (LEI)** = Age × Number of sexual partners ÷ (Age at first intercourse + 1). It increases with the number of years of potential exposure and with the number of partners, and decreases when sexual debut occurs later.
- **Cancer Load Score (CLS)** = Smokes (packs/year) × Smokes (years) + Hormonal contraceptive years. It combines two behavioural exposures that the literature associates with cervical carcinogenesis into a single additive burden score.

Any infinite values created by division were replaced with NaN and subsequently imputed. After this step the data contain 37 candidate predictors.

4. Frequency-based feature filtering (FBFF)

Attributes that are almost always zero provide little information and can only add noise. A frequency-based filter was therefore applied: a feature was kept only if more than 2% of its (imputed) values were non-zero. This removed 9 very sparse STI and diagnosis indicators (cervical, vaginal and genital-herpes–related STIs, molluscum contagiosum, AIDS, hepatitis B, STDs:HPV, pelvic inflammatory disease and Dx:CIN), leaving 28 candidate features.

5. Train–test split and SMOTE balancing

The data were split 80 : 20 using stratified sampling with a fixed random seed (42), giving 686 training records (44 positive) and 172 test records (11 positive). SMOTE was fitted on the training partition only, increasing it to 1284 records with equal class sizes (642 per class). The test set was never resampled, so it keeps the real-world class ratio and its metrics reflect the deployment setting.

6. Recursive Feature Elimination

RFE was run on the balanced training data with a Random Forest of 500 trees as the ranking estimator. At each iteration the least important feature is removed and the forest is refitted, until 18 features remain. The retained predictors are listed in Table 2. All screening tests, the three prior-diagnosis flags,

reproductive and contraceptive history, and both engineered features survived selection, which supports the clinical plausibility of the engineered variables.

Table 2: Features retained by RFE (18 of 28**)**

No.	Feature	No.	Feature
1	Age	10	STDs: Number of diagnosis
2	Number of sexual partners	11	Dx:Cancer
3	First sexual intercourse	12	Dx:HPV
4	Num of pregnancies	13	Dx
5	Hormonal Contraceptives	14	Hinselmann
6	Hormonal Contraceptives (years)	15	Schiller
7	IUD (years)	16	Citology
8	STDs	17	Lifetime_Exposure_Index
9	STDs (number)	18	Cancer_Load_Score

7. XGBoost classifier

The selected features feed an XGBoost classifier [5], an ensemble of gradient-boosted decision trees in which each new tree is fitted to the residual error of the current ensemble and a regularised objective penalises tree complexity.

Compared with a single tree or a bagged forest, boosting reduces bias while the regularisation terms (L1 penalty α, L2 penalty λ and minimum split loss γ) and row/column subsampling control variance. The hyper-parameters are listed in Table 3. A low learning rate with many trees was chosen to favour stable convergence on the small training set. The binary logistic output is a probability between 0 and 1.

Table 3: XGBoost hyper-parameters

Hyper-parameter	Value	Role
n_estimators	900	Number of boosting rounds
max_depth	6	Maximum tree depth
learning_rate	0.02	Shrinkage per round
subsample	0.9	Row sampling per tree
colsample_bytree	0.9	Feature sampling per tree
gamma	0.2	Minimum loss reduction to split
reg_alpha / reg_lambda	0.2 / 1.2	L1 / L2 regularisation
eval_metric	logloss	Training objective monitor
random_state	42	Reproducibility

8. Risk categorisation

The predicted probability is expressed as a percentage and mapped to three user-friendly categories: Low risk (below 30%), Medium risk (30%–60%) and High risk (above 60%). The thresholds are used to communicate the result and do not replace clinical judgement; they can be tuned to the sensitivity and specificity requirements of a screening programme.

9. Evaluation protocol

Because only 11 positive patients are present in the hold-out set, a single split yields very noisy estimates. Performance was therefore assessed in two complementary ways. (a) Hold-out evaluation: the model is trained on the balanced training partition and evaluated once on the untouched test partition using accuracy, precision, recall (sensitivity), specificity, F1-score, balanced accuracy, ROC-AUC, average precision (AP) and the Matthews correlation coefficient (MCC). (b) Repeated stratified cross-validation on all 858 patients (5 folds × 3 repeats). In each fold the imputer, SMOTE, RFE and the classifier are fitted inside the training portion only, so that no information from the validation fold leaks into model construction [14]. Logistic Regression, Random

Forest and XGBoost without RFE were evaluated with the same protocol as baselines. The formulas used are:

Accuracy = (TP + TN) / (TP + TN + FP + FN); Precision = TP / (TP + FP); Recall = TP / (TP + FN); Specificity = TN / (TN + FP); F1 = 2 × Precision × Recall / (Precision + Recall).
where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives, with "cancer" as the positive class.

10. System implementation

The model was developed in Python (scikit-learn [11], imbalanced-learn [12], XGBoost [5]) on Google Colab and the trained model and RFE selector were serialised for deployment. The user interface was built with Streamlit and is organised into modules: (i) a Home page that explains cervical cancer, its causes, symptoms and prevention; (ii) User Authentication with registration, login and session management so that health information is not accessible without signing in; (iii) a Model Accuracy page in which the training CSV can be uploaded to display test accuracy, cross-validation accuracy, the confusion matrix and the classification report; (iv) a Prediction page where the user enters health parameters and receives a risk percentage, a colour-coded category (green, yellow, red) and a feature-contribution chart; (v) an Analysis dashboard with total screenings and risk-distribution pie and bar charts; and (vi) a Report Generation module that uses the ReportLab library to produce a downloadable PDF containing the risk category, an explanatory table, the contribution chart and a medical disclaimer. Visualisations are drawn with Matplotlib [15]. The application was deployed on Hugging Face Spaces for functional testing with low-, medium- and high-risk input profiles.

IV. RESULTS AND DISCUSSION

1. Hold-out performance

Table 4 and Figure 4 report the performance of the proposed RFE + XGBoost model on the 172 held-out patients. The model correctly classified 166 of 172 patients, an accuracy of 96.51%. It identified 159 of 161 healthy patients (specificity 98.76%) and 7 of 11 biopsy-positive patients (recall 0.636), with 2 false alarms (precision 0.778). The ROC-AUC of 0.959 shows that the predicted probabilities rank positive patients well above negative patients, and the MCC of 0.686 indicates a strong, though not perfect, agreement between prediction and truth on this imbalanced test set.

Table 4: Hold-out performance of the proposed model (n = 172)

Metric	Value	Comment
Accuracy	96.51%	Overall correctness
Precision (cancer class)	0.778	Of predicted positives
Recall / sensitivity	0.636	7 of 11 cases found
Specificity	0.988	159 of 161 healthy correct
F1-score (cancer class)	0.700	Harmonic mean
Balanced accuracy	0.812	Mean of recall and specificity
ROC-AUC	0.959	Threshold-independent
Average precision	0.648	Area under PR curve
Matthews correlation coefficient	0.686	Robust to imbalance

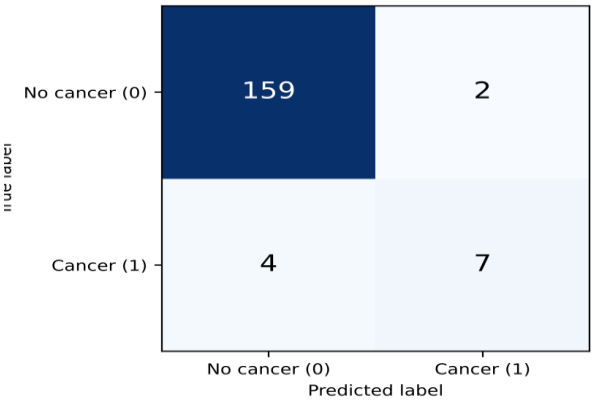


Figure 4: Confusion matrix of the proposed model on the hold-out set

Accuracy alone is misleading here: a trivial classifier that labels everyone as healthy would obtain 93.6% accuracy while detecting no cancer cases. The cancer-class recall, precision and F1-score are therefore the key indicators, and they show that, while overall accuracy is high, detection of positive cases is the main area for improvement. In a screening context a false negative is more costly than a false positive, so a lower decision threshold can be used to trade precision for recall; the probability output makes this possible without retraining.

2. ROC and precision–recall analysis

Figure 5 shows the ROC and precision–recall curves. The ROC curve lies far above the chance diagonal (AUC = 0.959). Because ROC curves can look optimistic when negatives vastly outnumber positives, the precision–recall curve is also shown [13]; its area (AP = 0.648) is well above the no-skill baseline equal to the positive prevalence (about 0.064), confirming that the ranking quality is genuine rather than an artefact of imbalance. The jagged shape of both curves reflects the small number of positive test cases.

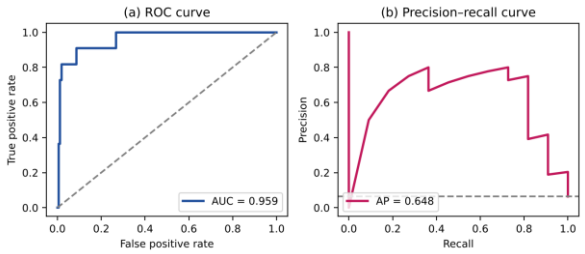


Figure 5: (a) ROC curve and (b) precision–recall curve on the hold-out set

3. Comparison with baseline models

Table 5 compares the proposed model with three baselines under repeated stratified cross-validation with SMOTE and RFE nested inside each fold (Figure 6). All models achieve accuracy between 94% and 96% and ROC-AUC between 0.90 and 0.95. The tree-based ensembles clearly outperform Logistic Regression in ROC-AUC (about 0.95 against 0.90), reflecting the non-linear relationships and interactions in the data. Logistic Regression attains the highest recall (0.800) but at the cost of the lowest precision (0.539), i.e. many false alarms. The proposed RFE + XGBoost model obtains accuracy 0.957, F1-score 0.683 and ROC-AUC 0.954, comparable to XGBoost without feature selection (0.959, 0.698 and 0.953). The standard deviations across folds (0.01–0.18) are larger than the differences between the tree-based models, so these models should be regarded as statistically similar on this dataset. The practical benefit of RFE is therefore a simpler model with 18 instead of 28 inputs and comparable accuracy, which reduces data-entry burden and improves interpretability, rather than a large gain in predictive power.

Table 5: Repeated stratified 5-fold cross-validation (3 repeats), mean ± SD, leakage-free

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.942 ± 0.017	0.539 ± 0.105	0.800 ± 0.134	0.639 ± 0.102	0.900 ± 0.057
Random Forest	0.956 ± 0.013	0.651 ± 0.097	0.685 ± 0.166	0.658 ± 0.114	0.947 ± 0.037
XGBoost (no RFE)	0.959 ± 0.012	0.661 ± 0.088	0.758 ± 0.155	0.698 ± 0.099	0.953 ± 0.029
RFE + XGBoost (proposed)	0.957 ± 0.015	0.649 ± 0.102	0.739 ± 0.175	0.683 ± 0.119	0.954 ± 0.026

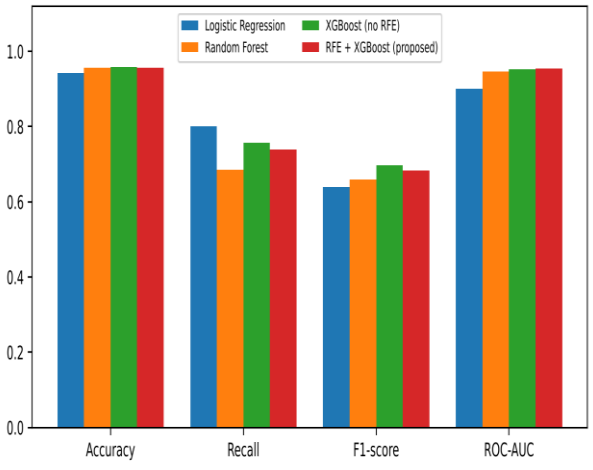


Figure 6: Cross-validated comparison of the proposed model with baseline classifiers

Table 6: Hold-out comparison on the same 80/20 split (n = 172, 11 positives)

Model	Accuracy (%)	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	95.35	0.600	0.818	0.692	0.874
Random Forest	96.51	0.727	0.727	0.727	0.971
XGBoost (no RFE)	96.51	0.727	0.727	0.727	0.940
RFE + XGBoost (proposed)	96.51	0.778	0.636	0.700	0.959

On the single hold-out split all four models reach about 95–97% accuracy. Differences of one or two patients change recall by 9 percentage points (1/11), which is why the repeated cross-validation in Table 5 is the more reliable basis for comparison.

4. Validation methodology and optimism

A common pitfall in imbalanced-learning studies is to apply SMOTE to the entire training set and then run cross-validation on the resampled data. In that setting synthetic neighbours of a validation patient are present in the training folds, and the score becomes optimistic. When we applied this naïve procedure to the selected features, five-fold cross-validated accuracy was 98.05%, higher than the 96.51% obtained on the untouched hold-out set and higher than the 95.73% from the nested protocol in Table 5. We therefore regard the nested cross-validation and the hold-out results, not the naïve figure, as the honest estimates of generalisation performance, and recommend that future work on this dataset reports them in the same way [14].

5. Model explainability

Explainability was assessed with two complementary views. Figure 7a shows the XGBoost gain-based feature importance and Figure 7b the mean absolute SHAP value on the hold-out set, which measures the average impact of each feature on the model output in log-odds units [9]. Both rankings agree that the Schiller test is by far the most influential predictor (mean |SHAP| = 3.83), followed by the Hinselmann test. This is consistent with the data: 87.3% of biopsy-positive patients had a positive Schiller test against 3.2% of biopsy-negative patients, and 45.5% had a positive Hinselmann test against 1.2% (Table 7). Beyond the screening tests, SHAP ranks the number of pregnancies, the Dx diagnosis flag, the number of sexual partners, age at first intercourse and the engineered Lifetime Exposure Index among the next most influential factors. The inclusion of LEI in this group shows that the engineered feature carries information that is not fully captured by its component variables.

The explanation provided to the end user in the application is a per-patient feature-contribution bar chart and a plain-language explanation of the risk category in the PDF report. These are intended to show which of the entered factors raised the risk, not to establish causation.

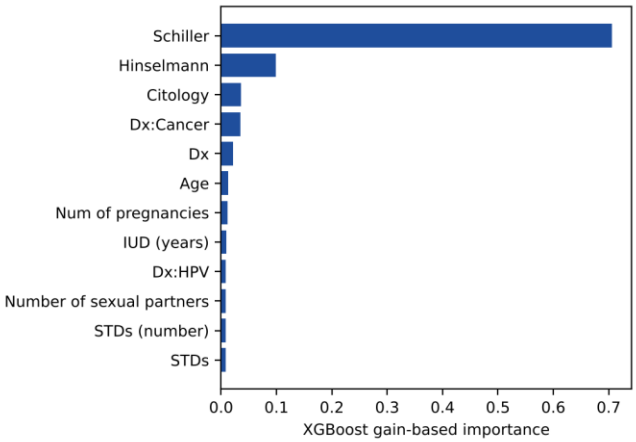


Figure 7a: Gain-based feature importance (top 12 features)

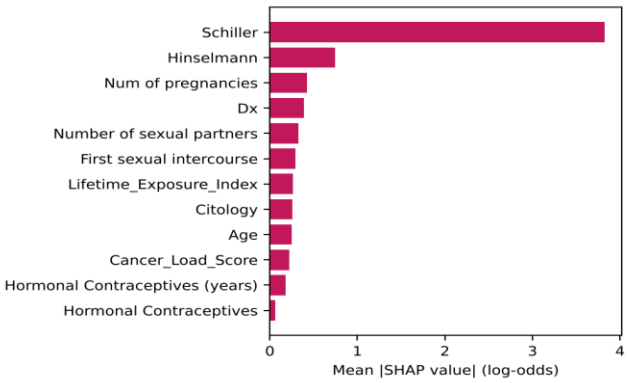


Figure 7b: Mean absolute SHAP values on the hold-out set (top 12 features)

Table 7: Prevalence of selected findings by biopsy outcome (full dataset)

Finding	Biopsy-positive (n = 55)	Biopsy-negative (n = 803)
Schiller test positive	87.27%	3.24%
Hinselmann test positive	45.45%	1.25%
Cytology positive	32.73%	3.24%
Current smoker	18.52%	14.29%
Mean age (years)	28.6	26.7

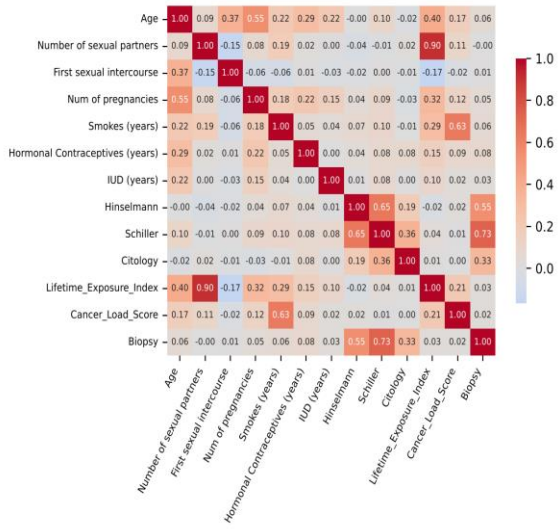


Figure 8: Pearson correlation among key attributes, engineered features and the biopsy outcome

The correlation matrix in Figure 8 shows that the strongest linear associations with the biopsy outcome are the Schiller (r = 0.73), Hinselmann (r = 0.55) and cytology (r = 0.33) tests, whereas the behavioural variables, such as hormonal contraceptive years, smoking years and age, correlate weakly (r < 0.08) with the outcome. Weak linear correlation does not mean such factors are unimportant, because their effect can be non-linear or conditional on other variables, which is what tree ensembles are able to exploit.

6. Web application results

The Streamlit application exposes the trained model to non-technical users. After registration and login, a user enters health parameters and obtains the predicted risk percentage, a progress bar and a colour-coded message (green for Low, yellow for Medium, red for High risk). The Analysis dashboard summarises screening counts and the risk distribution using pie and bar charts, and the Model Accuracy page lets reviewers upload the dataset and inspect the test accuracy, cross-validation accuracy, classification report and confusion matrix. The Report Generation module produces a PDF with the risk category, an explanation table for the three risk levels, the contribution chart and a clear disclaimer that the output is for information only and does not replace medical advice. Functional testing with low-, medium- and high-risk input profiles confirmed that predictions, charts and report generation worked end to end, and deployment on Hugging Face Spaces confirmed accessibility and stable real-time response. The dashboard counts displayed in the prototype were demonstration records and should not be interpreted as clinical statistics.

7. Discussion

The results support three conclusions. First, tree-based ensembles are well suited to this heterogeneous, partly missing and strongly imbalanced dataset, reaching a cross-validated ROC-AUC of about 0.95 versus 0.90 for Logistic Regression. Second, RFE reduces the input set from 28 to 18 features without a measurable loss of performance, which simplifies the application and clarifies which information is needed. Third, the explainability analysis shows that the model relies mainly on the results of screening tests that are themselves part of the diagnostic pathway. This means the system is best viewed as a decision-support tool that integrates test results with history to prioritise follow-up, and not as a purely lifestyle-based "risk factor" calculator: users who have no colposcopy or Schiller result would obtain less informative predictions.

Compared with earlier work on the same dataset that reports accuracies of about 98% or more, our hold-out accuracy of 96.51% is slightly lower, but it is

obtained with SMOTE confined to the training data, with a test set that is never resampled, and with additional metrics that expose the true cost of missed cases. We believe such conservative reporting is more useful to clinicians than a higher but optimistic headline accuracy.

8. Limitations

- **Small positive class.** The dataset has only 55 biopsy-positive patients, 11 of them in the test set, so every metric on the cancer class has wide uncertainty (cross-validation SDs for recall and F1 are 0.13–0.18 and 0.10–0.12).
- **Single-centre data.** All patients come from one hospital in Venezuela; performance in other populations, age ranges and screening settings is unknown, and external validation is required before any clinical use.
- **Dependence on screening-test results.** The Schiller and Hinselmann results dominate the predictions, which limits use where such tests are unavailable.
- **Duplicates and imputation.** Twenty-three duplicate rows were kept and imputation statistics were computed on the full dataset before the split in the hold-out experiment; both can make the hold-out estimate slightly optimistic. The nested cross-validation fits the imputer inside each fold.
- **Probability calibration.** The 30%/60% thresholds are heuristic. Because SMOTE changes the training prevalence, predicted probabilities are not calibrated to the true 6.4% prevalence and should be interpreted as relative risk scores.
- **Ethical considerations.** The tool handles sensitive health information and its output may cause anxiety or false reassurance. It must be accompanied by a disclaimer and a recommendation to seek professional screening.

V. CONCLUSION

This paper presented an explainable cervical cancer risk prediction system that integrates median imputation, engineered exposure features, frequency-based filtering, SMOTE balancing, Recursive Feature Elimination and XGBoost, and delivers the result through a Streamlit application with risk categorisation, SHAP-based interpretation, an analytics dashboard and PDF report generation. On a stratified hold-out set the model reached 96.51% accuracy, 98.76% specificity and an ROC-AUC of 0.959, and leakage-free repeated cross-validation showed performance comparable to, and in ROC-AUC clearly better than, Logistic Regression, with a smaller 18-feature input set than XGBoost without selection. The explanation analysis identified the Schiller and Hinselmann tests, reproductive history and the Lifetime Exposure Index as the main drivers of risk.

The study also illustrates the importance of honest evaluation on imbalanced medical data: accuracy alone, and cross-validation after oversampling, overstate performance, and cancer-class recall, precision and nested validation should be reported. The system supports early risk assessment and awareness and is not a diagnostic device.

Future work includes validating the model on larger multi-centre and Indian population datasets, calibrating probabilities and choosing thresholds to meet a target sensitivity, adding per-patient SHAP explanations directly to the user interface and PDF report, comparing with other boosting methods and cost-sensitive learning, adding HPV test and vaccination data, integrating with hospital information systems, and conducting a usability study with clinicians and patients.

Declarations

Data availability. The study used the publicly available Cervical Cancer Risk Factors dataset (UCI Machine Learning Repository / Kaggle) [3]. The dataset is de-identified and no new patient data were collected.

Ethics statement. Only a public, anonymised dataset was analysed; therefore no additional ethical approval or patient consent was required.

Conflict of interest. The author declares no conflict of interest.

Funding. This research received no external funding.

Acknowledgement. The author thanks Dr. S. Vijayalakshmi, Professor & Head, B.Sc CS (DA) & PG, Department of Computer Science, Sri Ramakrishna College of Arts & Science for Women, Coimbatore, for guidance during this work.

REFERENCES

1. Sung, H., Ferlay, J., Siegel, R. L., et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," CA: A Cancer Journal for Clinicians, 71(3), 209–249, 2021.
2. World Health Organization, Global Strategy to Accelerate the Elimination of Cervical Cancer as a Public Health Problem. Geneva: WHO, 2020.
3. Fernandes, K., Cardoso, J. S., & Fernandes, J., "Transfer Learning with Partial Observability Applied to Cervical Cancer Screening," Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA), Springer, pp. 243–250, 2017.
4. Wu, W., & Zhou, H., "Data-Driven Diagnosis of Cervical Cancer with Support Vector Machine-Based Approaches," IEEE Access, 5, 25189–25195, 2017.
5. Chen, T., & Guestrin, C., "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016.
6. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V., "Gene Selection for Cancer Classification Using Support Vector Machines," Machine Learning, 46(1–3), 389–422, 2002.
7. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P., "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, 16, 321–357, 2002.
8. Breiman, L., "Random Forests," Machine Learning, 45(1), 5–32, 2001.
9. Lundberg, S. M., & Lee, S.-I., "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems 30, 2017.
10. Doshi-Velez, F., & Kim, B., "Towards a Rigorous Science of Interpretable Machine Learning," arXiv:1702.08608, 2017.
11. Pedregosa, F., Varoquaux, G., Gramfort, A., et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, 12, 2825–2830, 2011.
12. Lemaître, G., Nogueira, F., & Aridas, C. K., "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," Journal of Machine Learning Research, 18(17), 1–5, 2017.
13. Saito, T., & Rehmsmeier, M., "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," PLoS ONE, 10(3), e0118432, 2015.
14. Cawley, G. C., & Talbot, N. L. C., "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation," Journal of Machine Learning Research, 11, 2079–2107, 2010.
15. Hunter, J. D., "Matplotlib: A 2D Graphics Environment," Computing in Science & Engineering, 9(3), 90–95, 2007.
16. Ribeiro, M. T., Singh, S., & Guestrin, C., "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," Proceedings of the 22nd ACM SIGKDD Conference, pp. 1135–1144, 2016.
17. Chicco, D., & Jurman, G., "The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation," BMC Genomics, 21, 6, 2020.
18. Kennedy, J., & Eberhart, R., "Particle Swarm Optimization," Proceedings of IEEE International Conference on Neural Networks, pp. 1942–1948, 1995.
19. Sharjil Iqbal. "NETWORK ATTACK DETECTION BY GROUP SEARCH OPTIMIZATION USING CONVOLUTIONAL DEEP LEARNING MODEL