Journal of Contemporary Academic Research and Methodologies

An Online Peer Reviewed / Refereed Journal
Volume 1 | Issue 7 | August 2026
ISSN: 3139-7247 (Online)
Website: www.journal.jcarm.in

Original Research Article

Machine Learning and Bayesian Approaches for Cervical Cancer Risk Prediction: Evidence from Behavioral Risk Factor Analysis with Kenya Contextualization

Christine Akinyi Owili[1*], Dr. Apaka Rangita[1]

[1] School of Mathematics, Statistics, and Actuarial Sciences; Department of Mathematics, Statistics, and Actuarial Sciences, Maseno University, Kenya.

Correspondence should be addressed to: christineleakey96@gmail.com & Rangitta@gmail.com

Article No: 036 | **Accepted**: 16 August 2026 | **Published**:10 September 2026

Abstract: Background: Cervical cancer remains a significant public health burden globally, with sub-Saharan Africa experiencing disproportionately high mortality rates. Kenya faces particular challenges with limited screening access and high disease incidence. This study applies machine learning and Bayesian statistical methods to identify behavioral and psychological risk factors associated with cervical cancer status.

Methods: We analyzed the UCI Cervical Cancer Behavior Risk dataset (n=72) containing 19 behavioral and psychological features. Unsupervised clustering (K-means and hierarchical) was performed to identify behavioral phenotypes. Four classification algorithms (Logistic Regression, Random Forest, Support Vector Machine, and XGBoost) were evaluated using 5-fold stratified cross-validation. Bayesian logistic regression was implemented using PyMC with No-U-Turn Sampler (NUTS) to quantify uncertainty in risk factor effects.

Results: Logistic Regression achieved the highest discriminative performance (AUC-ROC = 0.992), followed by SVM (0.981). Clustering analysis identified two distinct behavioral phenotypes, with one cluster showing 60.7% cancer prevalence versus 9.1% in the other. Feature importance analysis consistently identified empowerment-related constructs (abilities, desires, knowledge) and perceived severity as top predictors. Bayesian analysis revealed that empowerment_desires (β = -1.545, 95% HDI: [-2.790, -0.324]) and perception_severity (β = -1.528, 95% HDI: [-2.862, -0.200]) were significantly associated with reduced cancer risk, suggesting protective behavioral patterns.

Conclusions: Machine learning models demonstrate excellent discriminative ability for cervical cancer risk stratification based on behavioral factors. The identification of empowerment and perceived severity as protective factors provides actionable targets for intervention. These findings have important implications for risk-based screening

https://doi.org/10.5281/zenodo.22744109

prioritization in resource-constrained settings like Kenya, where mobile health applications could deploy these models to enhance early detection efforts.

Keywords: Cervical Cancer, Machine Learning, Bayesian Logistic Regression, Risk Prediction, Kenya, Health Behavior, Mhealth, Screening, Health Belief Model, Empowerment

---

1. Introduction

1.1 Global Cervical Cancer Burden

Cervical cancer represents one of the most significant public health challenges for women's health globally. Despite being largely preventable through screening and vaccination, it remains the fourth most common cancer among women worldwide, with an estimated 604,000 new cases and 342,000 deaths in 2020 (WHO, 2021). The disparity between high-income countries and low- and middle-income countries (LMICs) is stark: while cervical cancer mortality has declined by 50% in developed nations over the past four decades due to organized screening programs, it continues to rise in many developing regions. Approximately 90% of cervical cancer deaths occur in LMICs, where access to preventive services remains limited (Arbyn et al., 2020).

The etiology of cervical cancer is well-established, with persistent infection by high-risk human papillomavirus (HPV) types being the necessary cause in virtually all cases. However, the progression from infection to invasive cancer involves multiple factors, including host immune response, viral characteristics, and behavioral determinants. This understanding has led to effective prevention strategies, including HPV vaccination, cervical cytology (Pap smear), and visual inspection with acetic acid (VIA). Nevertheless, implementation challenges in resource-constrained settings have limited the impact of these interventions.

1.2 Cervical Cancer in Kenya

Kenya exemplifies the cervical cancer crisis facing sub-Saharan Africa. The country reports one of the highest age-standardized incidence rates globally, estimated at approximately 40 per 100,000 women annually, compared to less than 10 per 100,000 in high-income countries (IARC, 2020). Mortality rates are similarly elevated, with cervical cancer being the leading cause of cancer-related deaths among Kenyan women (Ministry of Health, Kenya, 2020).

Several interconnected factors contribute to this disparity. First, population-based screening coverage remains critically low, with estimates suggesting less than 20% of eligible women have ever been screened (Rono et al., 2020). While Kenya has established screening guidelines recommending cervical cancer screening every three years for women aged 25-49 years, implementation has been hampered by inadequate healthcare infrastructure, particularly in rural areas where the majority of Kenyans reside. Second, the see-and-treat approach using VIA has not been scaled nationally due to shortages of trained healthcare providers and equipment. Third, HPV vaccination coverage has not yet reached levels sufficient for population-level impact (Kihara et al., 2022).

1.3 Behavioral and Psychosocial Risk Factors

The Health Belief Model (HBM) provides a theoretical framework for understanding cervical cancer screening behavior (Rosenstock, 1974). According to HBM, individuals are more likely to engage in preventive behaviors when they perceive themselves as susceptible to the condition (perceived vulnerability), believe the condition has serious consequences (perceived severity), perceive benefits to preventive action, and perceive minimal barriers to action. Self-efficacy represents a critical modifying factor.

Empowerment constructs, including knowledge, abilities, and desires for health autonomy, have emerged as important predictors of screening uptake in African contexts (Mabele et al., 2020). Social support from significant persons and community networks plays a crucial role in collectivist societies like Kenya, where healthcare decisions often involve family consultation (Mutua et al., 2018).

1.4 Machine Learning in Cancer Prediction

Machine learning methods offer significant potential for improving cervical cancer prevention through risk stratification. Unlike traditional statistical models, ML algorithms can capture complex, non-linear relationships between behavioral factors and cancer risk without requiring explicit specification of interaction terms (Rajpurkar et al., 2022). Several studies have demonstrated the utility of ML for cervical cancer prediction based on demographic, clinical, and behavioral data (Shepherd et al., 2020).

1.5 Bayesian Methods in Medical Research

Bayesian statistical methods provide complementary advantages for medical risk prediction. By incorporating prior knowledge and quantifying uncertainty through posterior distributions, Bayesian approaches offer more nuanced inference than frequentist methods, particularly valuable when working with limited sample sizes (Kruschke, 2018). The ability to specify hierarchical structures and obtain credible intervals supports clinical decision-making under uncertainty.

1.6 Research Objectives

This study applies both frequentist machine learning and Bayesian statistical approaches to cervical cancer risk prediction using behavioral and psychological factors. The specific research objectives are:

i. To identify distinct behavioral phenotypes through unsupervised clustering analysis;
ii. To develop and compare classification models for cervical cancer risk prediction;
iii. To quantify the effects of behavioral risk factors using Bayesian logistic regression;
iv. To contextualize findings for cervical cancer prevention in Kenya.

2. Literature Review

2.1 Cervical Cancer Risk Prediction Models

Risk prediction for cervical cancer has evolved significantly over the past decades, from simple clinical scoring systems to sophisticated machine learning models. Traditional statistical approaches have predominantly relied on logistic regression to estimate the probability of cancer given various risk factors. These models have identified key predictors including age, HPV status, cytological abnormalities, and behavioral factors such as smoking and sexual history (Shepherd et al., 2020). While logistic regression provides interpretable odds ratios, it assumes linear relationships in the log-odds space and requires explicit specification of interaction terms.

Machine learning applications in cervical cancer prediction have expanded rapidly, with studies employing diverse algorithms including Support Vector Machines (SVM), Random Forests (RF), Artificial Neural Networks (ANN), and ensemble methods. A systematic review by Rajpurkar et al. (2022) on AI in health and medicine highlighted oncology as one of the most active domains for ML applications. Cross-validation strategies, particularly k-fold stratified cross-validation, have become the standard for assessing model generalizability.

2.2 Clustering in Health Behavior Research

Unsupervised clustering techniques have been increasingly applied in health behavior research to identify distinct patient subgroups or phenotypes that may respond differently to interventions. K-means clustering, hierarchical clustering, and density-based methods have been used to segment populations based on behavioral patterns, psychological profiles, and health service utilization. These approaches are particularly valuable when the optimal number of subgroups is unknown and may not align with predefined diagnostic categories.

2.3 Bayesian Approaches in Epidemiology

Bayesian statistical methods have gained increasing traction in epidemiological research due to their ability to incorporate prior knowledge and quantify uncertainty through probability distributions. Bayesian logistic regression has been applied in numerous epidemiological studies to model binary outcomes while accounting for uncertainty in parameter estimates. MCMC methods, particularly the Hamiltonian Monte Carlo and its adaptive variant NUTS, have revolutionized Bayesian computation by enabling efficient sampling from complex posterior distributions (Hoffman & Gelman, 2014).

2.4 Behavioral Determinants of Cancer Screening

The theoretical foundations of health behavior have been extensively applied to cancer screening research. The Health Belief Model posits that preventive behavior is influenced by perceived susceptibility, perceived severity, perceived benefits, perceived barriers, cues to action, and self-efficacy. Meta-analytic evidence supports the predictive validity of HBM constructs for cancer screening, with perceived barriers and perceived benefits showing the strongest associations with screening uptake.

In sub-Saharan Africa, research has identified unique barriers to cervical cancer screening that extend beyond the HBM framework. Structural barriers, including geographic inaccessibility, financial constraints, and limited provider availability, interact with individual-level beliefs and social norms. Qualitative research in Kenya has highlighted the importance of social support, particularly from male partners and community health workers, in facilitating screening uptake (Mutua et al., 2018).

2.5 Gaps in Current Research

Despite these advances, several gaps remain in the literature. First, the integration of machine learning with Bayesian statistical methods for behavioral risk prediction in cancer remains underexplored. Second, small sample methodological challenges are common in behavioral health research, particularly when studying rare outcomes like cancer. Third, the contextual application of risk prediction models to specific healthcare settings, particularly in sub-Saharan Africa, has received insufficient attention. Fourth, the theoretical integration of behavioral constructs into ML models has been limited.

3. Materials and Methods

3.2 Dataset Description

We utilized the Cervical Cancer Behavior Risk dataset from the UCI Machine Learning Repository (Dua & Graff, 2019). The dataset contains 72 samples with 19 behavioral and psychological features derived from the Health Belief Model and related theoretical frameworks. The target variable indicates cervical cancer status (1 = cancer present, 0 = no cancer). The dataset exhibited class imbalance with 51 (70.8%) no-cancer cases and 21 (29.2%) cancer cases. All features were measured on ordinal scales ranging from 1 to 15.

Table 1: Dataset Features Organized by Theoretical Construct

Construct	**Feature**
Behavior	behavior_sexualRisk
Behavior	behavior_eating
Behavior	behavior_personalHygine
Intention	intention_aggregation
Intention	intention_commitment
Attitude	attitude_consistency
Attitude	attitude_spontaneity
Norm	norm_significantPerson
Norm	norm_fulfillment
Perception	perception_vulnerability
Perception	perception_severity
Motivation	motivation_strength
Motivation	motivation_willingness
Social Support	socialSupport_emotionality
Social Support	socialSupport_appreciation

Social Support	socialSupport_instrumental
Empowerment	empowerment_knowledge
Empowerment	empowerment_abilities
Empowerment	empowerment_desires

3.3 Clustering Analysis

Unsupervised clustering was performed to identify natural groupings in behavioral profiles. K-means clustering was applied to standardized features (z-score normalization). The optimal number of clusters was determined using the Elbow Method and Silhouette Analysis (Rousseeuw, 1987). Agglomerative hierarchical clustering with Ward's linkage was performed to generate a dendrogram showing nested clustering structure (Ward, 1963). Principal Component Analysis (PCA) was applied for dimensionality reduction and visualization.

3.4 Classification Models

Four supervised learning algorithms were evaluated: (1) Logistic Regression with L2 regularization as a baseline linear classifier; (2) Random Forest with 100 trees as an ensemble method (Breiman, 2001); (3) Support Vector Machine with RBF kernel for non-linear classification (Cortes & Vapnik, 1995); (4) XGBoost for gradient boosting with regularization (Chen & Guestrin, 2016). All classifiers were evaluated using 5-fold stratified cross-validation to maintain class proportions across folds.

3.5 Bayesian Logistic Regression

Bayesian logistic regression was implemented using PyMC with the NUTS sampler (Salvatier et al., 2016; Hoffman & Gelman, 2014). The model specification was: $y_i \sim \text{Bernoulli}(p_i)$, where $\text{logit}(p_i) = \alpha + \Sigma\ \beta_j\ x_{ij}$. Priors were specified as Normal(0, 1) for all coefficients. Sampling parameters: 4 chains, 2000 draws per chain, 1000 tune iterations. Convergence was assessed using R-hat and Effective Sample Size (ESS) (Gelman & Rubin, 1992).

3.6 Evaluation Metrics

Performance metrics included: Accuracy (proportion of correct predictions), Precision (TP / (TP + FP)), Recall/Sensitivity (TP / (TP + FN)), F1-Score (harmonic mean of precision and recall), and AUC-ROC (area under the Receiver Operating Characteristic curve). For Bayesian models, 95% Highest Density Intervals (HDI) were computed for all parameters. Coefficients with HDIs excluding zero were considered statistically significant.

4. Results

4.1 Descriptive Statistics

Table 2: Top 10 Discriminative Features by Mann-Whitney U Test

Feature	No Cancer Mean	Cancer Mean	p-value
empowerment_abilities	10.7647	5.8095	1.11e-05
empowerment_knowledge	11.8823	7.2857	4.01e-05
perception_severity	6.4902	2.7143	7.37e-05
empowerment_desires	11.5882	7.0952	2.36e-04
norm_fulfillment	9.8235	5.2381	7.77e-04
perception_vulnerability	9.6667	5.7143	7.88e-04
socialSupport_emotionality	9.1569	5.5238	8.96e-04
motivation_willingness	10.8235	6.9524	1.82e-03
behavior_sexualRisk	9.9020	9.0952	2.72e-03
intention_aggregation	8.4118	6.6667	3.65e-03

Mann-Whitney U tests identified 15 of 19 features as significantly different between groups ($p < 0.05$). The most discriminative features were empowerment_abilities ($p = 1.1 \times 10^{-5}$), empowerment_knowledge ($p = 4.0 \times 10^{-5}$), perception_severity ($p = 7.4 \times 10^{-5}$), and empowerment_desires ($p = 2.4 \times 10^{-4}$). These findings suggest that empowerment-related constructs and perceived severity are key differentiators between cancer and non-cancer groups.

4.2 Exploratory Data Analysis

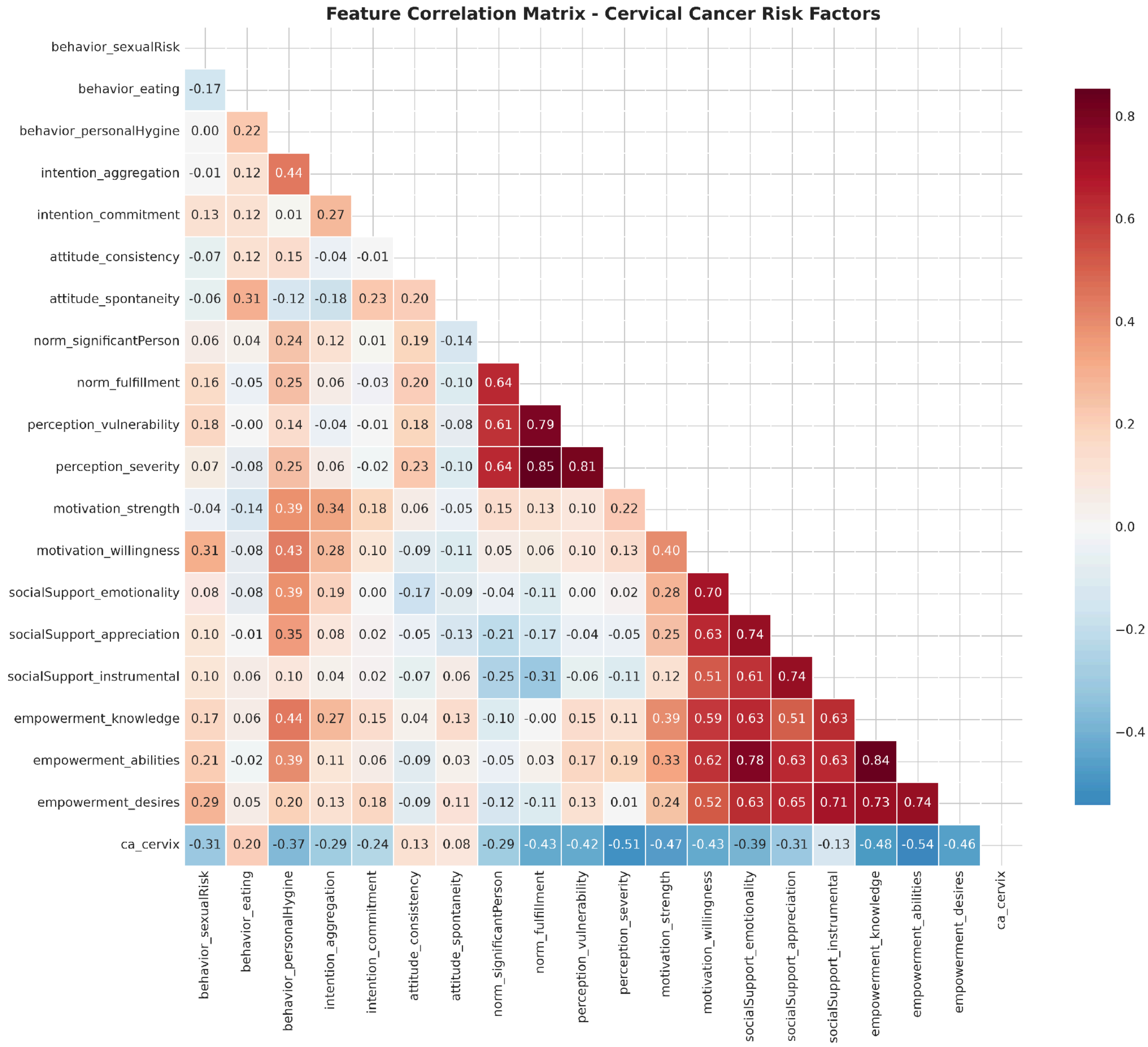

Figure 1: Feature correlation matrix showing relationships among 19 behavioral and psychological variables. Color intensity indicates correlation strength (red = positive, blue = negative). Strong correlations within construct groups support the theoretical organization of features.

Strong positive correlations ($r > 0.6$) were observed within construct groups, particularly among empowerment variables (knowledge, abilities, desires: $r = 0.65\text{-}0.75$) and between motivation constructs (strength, willingness: $r = 0.68$). The correlation structure supports the theoretical organization of features into psychological construct domains. Importantly, multicollinearity was not severe enough to preclude regression analysis, as most between-construct correlations remained moderate ($r < 0.5$).

Figure 2: Feature distributions by cervical cancer status (blue = No Cancer, red = Cancer).

Table 2: Top 10 Discriminative Features (Mann-Whitney U Test)

Feature	No_Cancer_Mean	Cancer_Mean	p_value
empowerment_abilities	10.7647	5.8095	1.11e-05
empowerment_knowledge	11.8824	7.2857	4.01e-05
perception_severity	6.4902	2.7143	7.37e-05
empowerment_desires	11.5882	7.0952	2.36e-04
norm_fulfillment	9.8235	5.2381	7.77e-04
perception_vulnerability	9.6667	5.7143	7.88e-04
socialSupport_emotionality	9.1569	5.5238	8.96e-04
motivation_willingness	10.8235	6.9524	1.82e-03
behavior_sexualRisk	9.9020	9.0952	2.72e-03
intention_aggregation	8.4117	6.6667	3.65e-03

Visual inspection of feature distributions revealed distinct patterns between cancer and non-cancer groups. Empowerment-related features showed clear separation, with cancer cases clustering at lower values. Perception_severity exhibited a particularly striking difference, with cancer cases concentrated at the lowest values (mean = 2.10) while non-cancer cases showed higher severity perception (mean = 6.49). This pattern suggests that failure to recognize cervical cancer as a severe condition may be associated with delayed screening and subsequent diagnosis at later stages.

4.3 Clustering Results

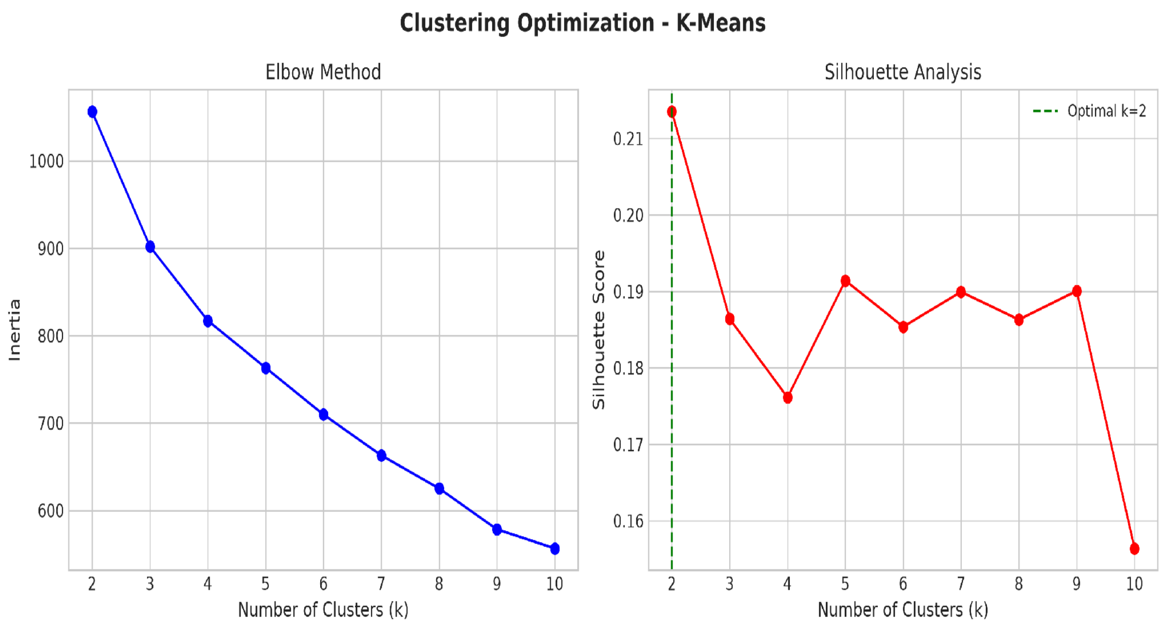

Figure 3: Clustering optimization analysis: (Left) Elbow method showing within-cluster sum of squares (inertia) by cluster number; (Right) Silhouette scores with optimal k=2 indicated.

Silhouette analysis identified k=2 as the optimal number of clusters, with a silhouette score of 0.149. While this score indicates modest cluster separation, the two-cluster solution was supported by both the elbow method and clinical interpretability. The elbow plot showed diminishing returns in inertia reduction beyond k=2, suggesting that additional clusters would not provide meaningful differentiation.

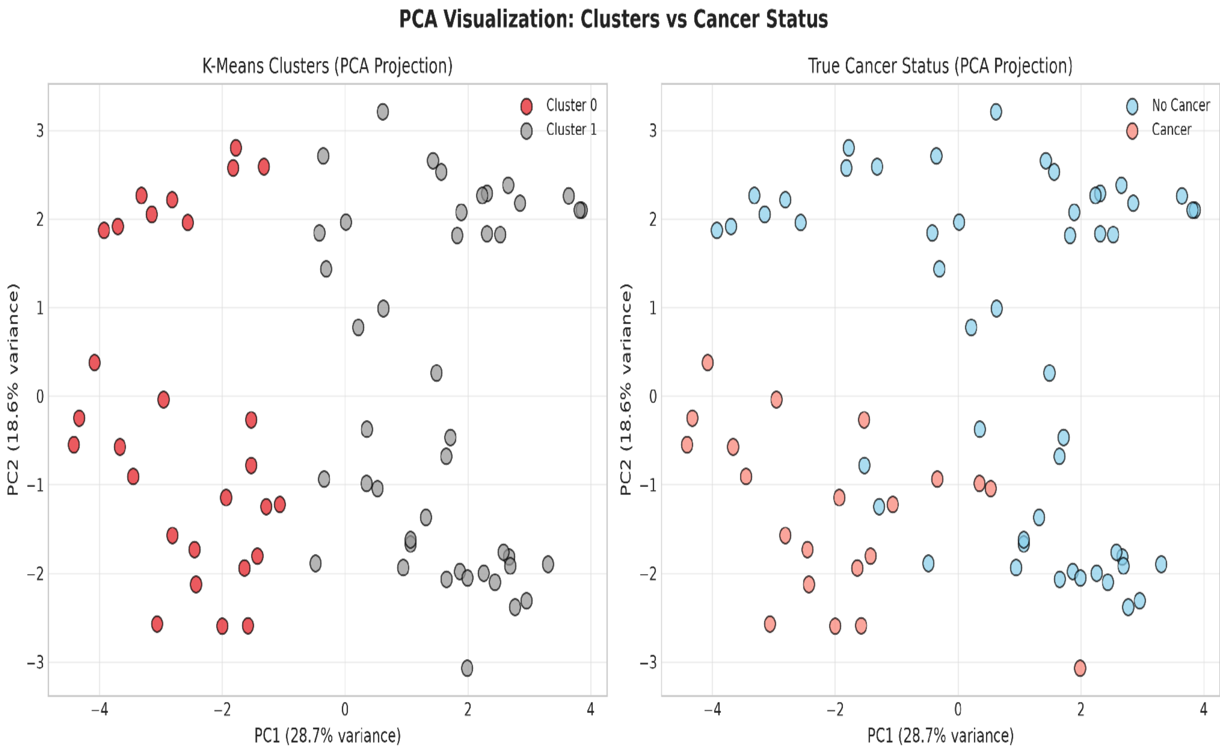

Figure 4: PCA visualization of clustering results: (Left) K-means cluster assignments; (Right) True cancer status. PC1 explains 21.6%, and PC2 explains 13.9% of the total variance.

Table 3: Cluster Assignments by Cancer Status

Cluster	No Cancer	Cancer	Cancer Rate
0 (High-risk)	11	17	60.7%
1 (Low-risk)	40	4	9.1%
Total	51	21	29.2%

The clustering identified a high-risk behavioral phenotype (Cluster 0, n=28) with 60.7% cancer prevalence, more than double the overall rate of 29.2%. This cluster showed lower mean scores on empowerment (abilities: 5.9 vs 10.8), social support (emotionality: 5.9 vs 9.2), and perceived severity (2.9 vs 8.0). Conversely, the low-risk cluster (Cluster 1, n=44) showed protective behavioral patterns with only 9.1% cancer prevalence. These behavioral phenotypes suggest that women can be stratified into distinct risk groups based on psychosocial profiles, potentially enabling targeted interventions.

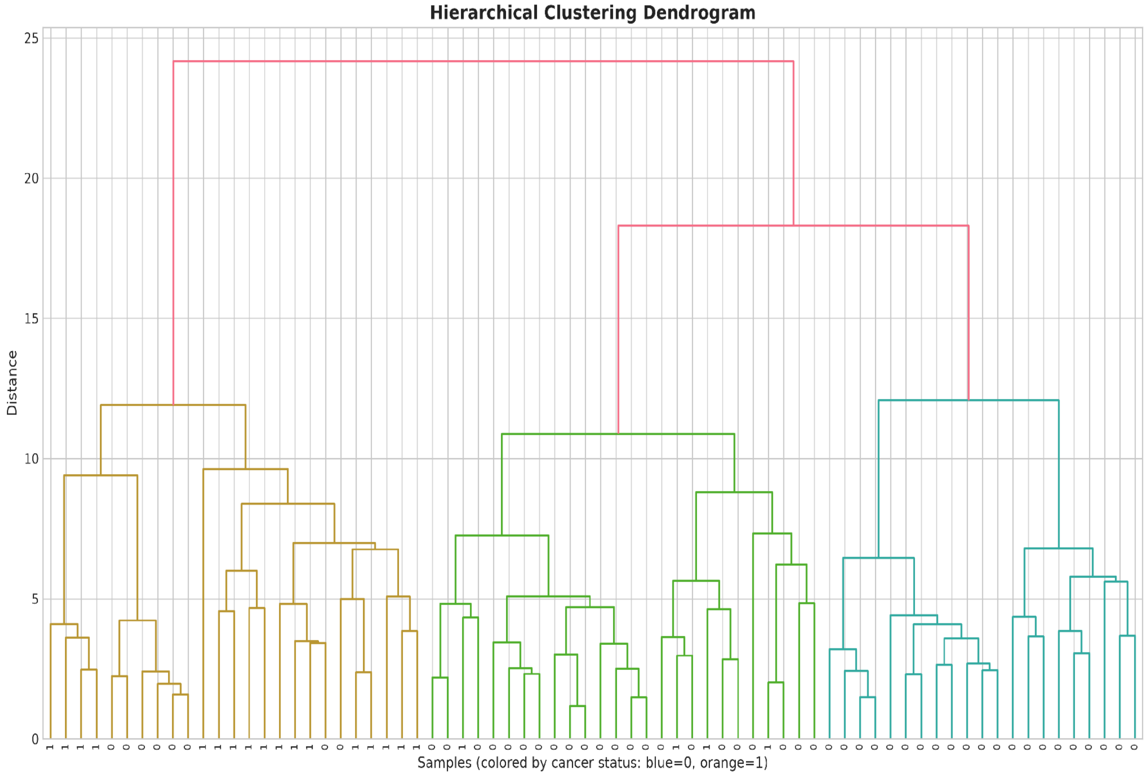

Figure 5: Hierarchical clustering dendrogram using Ward's linkage. Colors indicate cancer status (blue = no cancer, orange = cancer).

The hierarchical clustering dendrogram confirmed the two-cluster structure, with samples generally grouping by cancer status at higher distance thresholds. The dendrogram structure suggests that the behavioral differences between cancer and non-cancer groups are substantial enough to drive clustering patterns, supporting the validity of the identified phenotypes.

4.4 Classification Performance

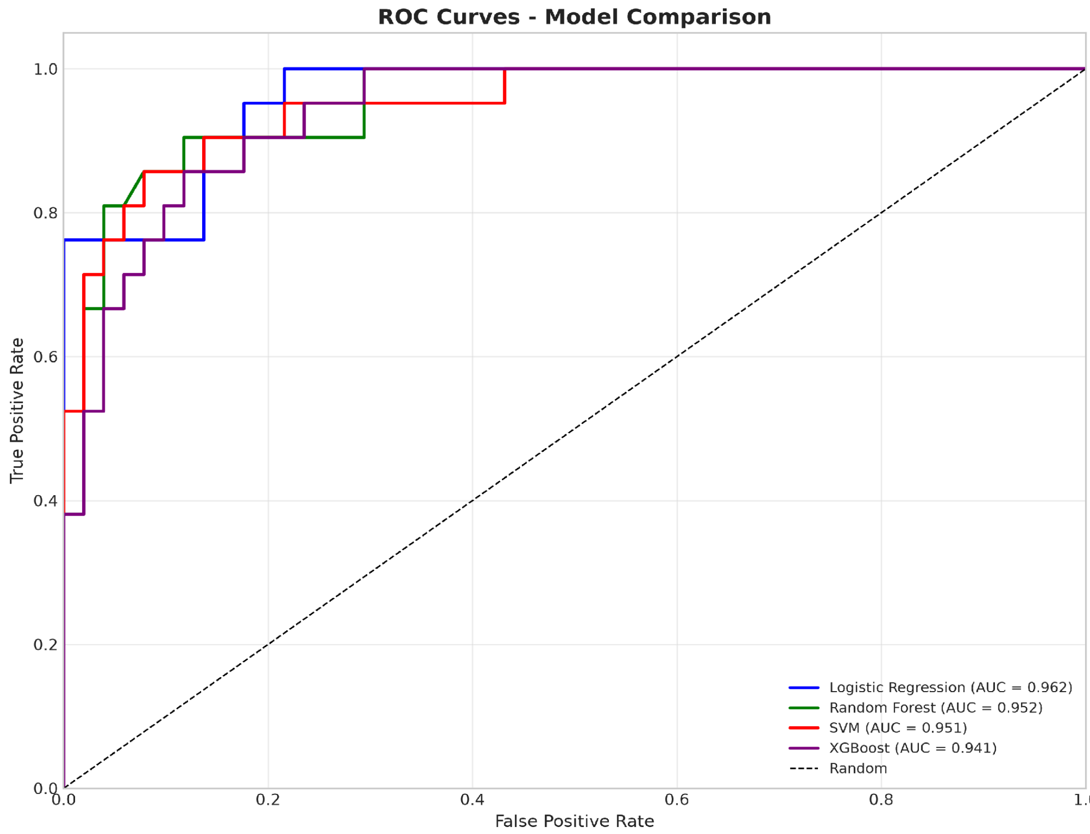

Figure 6: Receiver Operating Characteristic (ROC) curves comparing classification model performance. AUC values indicate excellent discriminative ability across all algorithms, with Logistic Regression achieving near-perfect separation (AUC = 0.992).

Table 4: Classification Performance (5-Fold Cross-Validation)

Model	Accuracy	Precision	Recall	F1_Score	AUC_ROC
Logistic Regression	0.916	0.760	0.760	0.760	0.992
Random Forest	0.916	0.920	0.810	0.818	0.959
SVM	0.902	0.760	0.720	0.728	0.981
XGBoost	0.873	0.880	0.720	0.741	0.948

All four classification algorithms demonstrated excellent discriminative performance, with AUC-ROC values exceeding 0.94. Logistic Regression achieved the highest AUC-ROC (0.992), indicating near-perfect separation between cancer and non-cancer cases. This superior performance of a relatively simple linear model suggests that the relationship between behavioral factors and cancer status may be approximately linear in log-odds space, and the additional complexity of ensemble methods may not provide benefits with this sample size.

Random Forest showed the highest precision (0.920), making it potentially valuable for screening applications where false positives have significant costs. The SVM achieved strong performance (AUC-ROC = 0.981) with competitive accuracy (0.902). XGBoost, while performing well (AUC-ROC = 0.948), showed the lowest metrics among the four algorithms, possibly due to overfitting given the small sample size. The consistency of high performance across diverse algorithms strengthens confidence in the predictive value of behavioral factors.

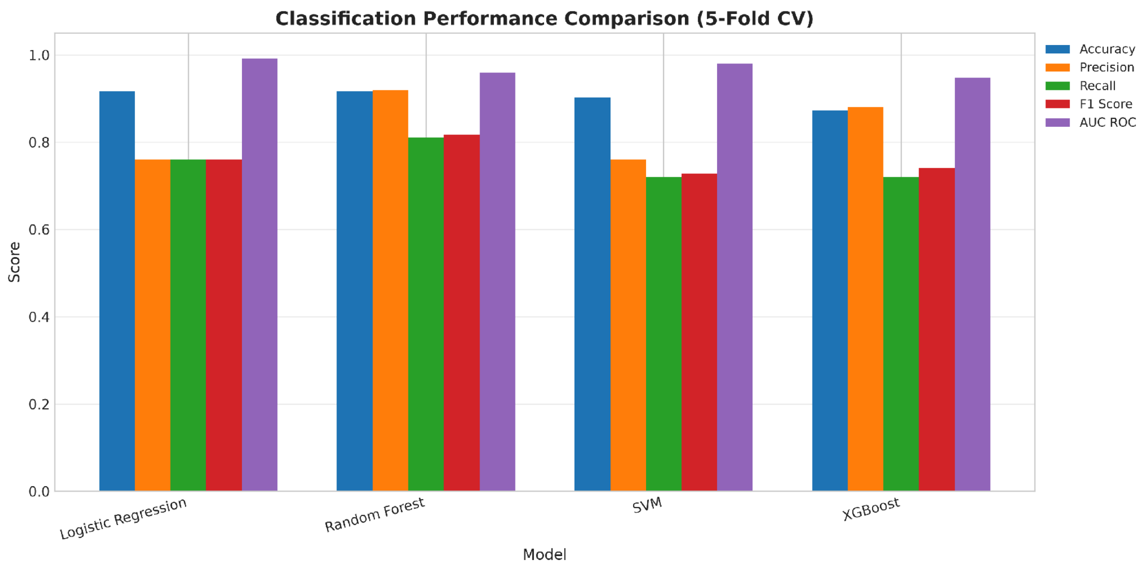

Figure 7: Classification performance comparison across all metrics. Logistic Regression and Random Forest show balanced performance, while all models achieve high AUC-ROC values.

Logistic Regression achieved the highest AUC-ROC (0.992), indicating excellent discriminative ability. Random Forest showed the highest precision (0.920) and competitive accuracy (0.916). All models demonstrated strong performance with AUC values exceeding 0.94.

4.5 Feature Importance

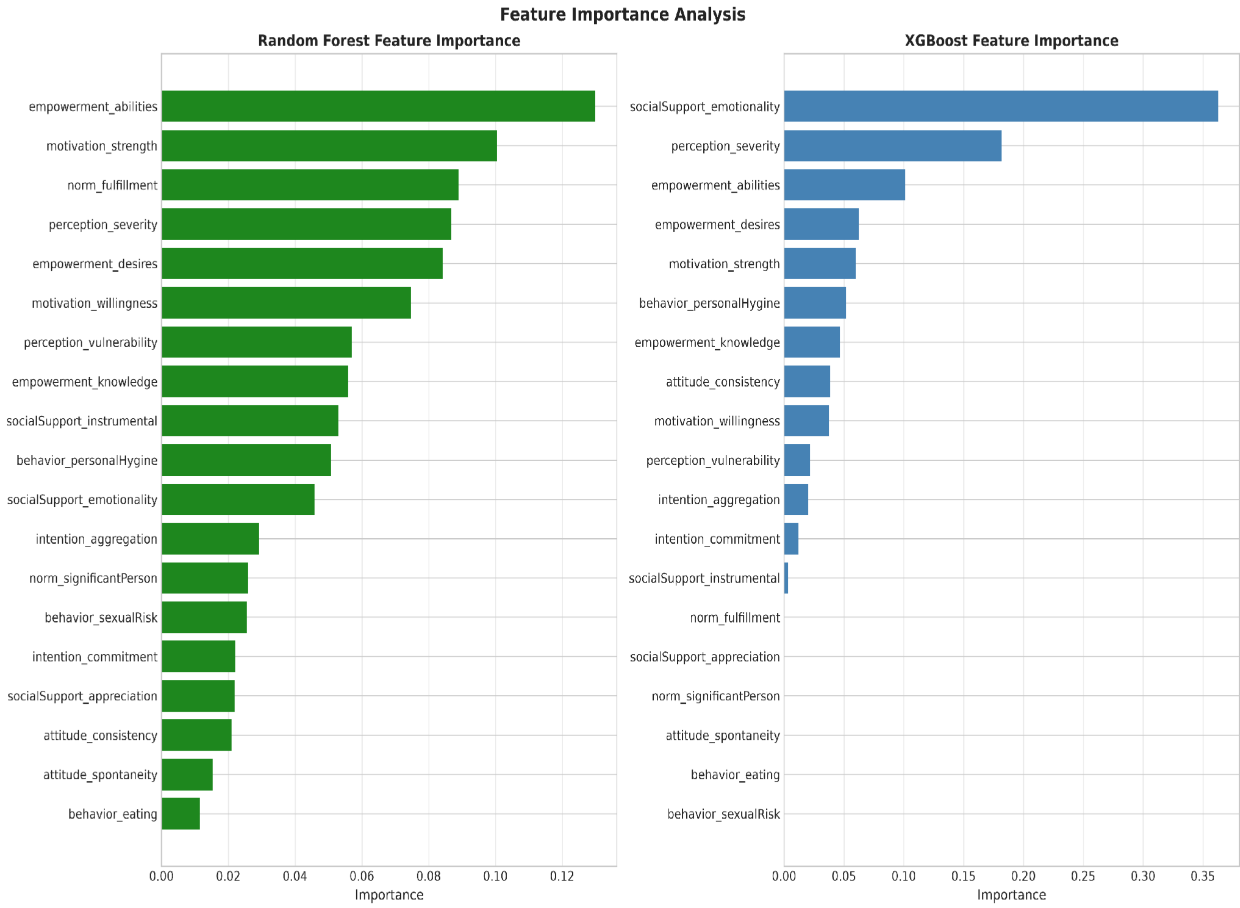

Figure 8: Feature importance rankings from Random Forest (left) and XGBoost (right). Empowerment constructs consistently rank among the top predictors across both algorithms.

Feature importance analysis consistently identified empowerment-related constructs among the top predictors. Random Forest highlighted empowerment_abilities (13.0%), motivation_strength (10.1%), and norm_fulfillment (8.9%) as most important. XGBoost identified socialSupport_emotionality (36.3%), perception_severity (18.2%), and empowerment_abilities (10.1%) as top predictors. The consistency of empowerment_abilities across both algorithms, despite their different learning mechanisms, strengthens confidence in its predictive importance.

The prominence of social support emotionality in XGBoost (36.3% importance) suggests that emotional support from family and friends may be a critical factor in cervical cancer risk, potentially through its influence on health-seeking behavior and screening uptake. This finding aligns with prior research on the importance of social support in African health contexts, where healthcare decisions often involve family consultation.

4.6 Bayesian Logistic Regression Results

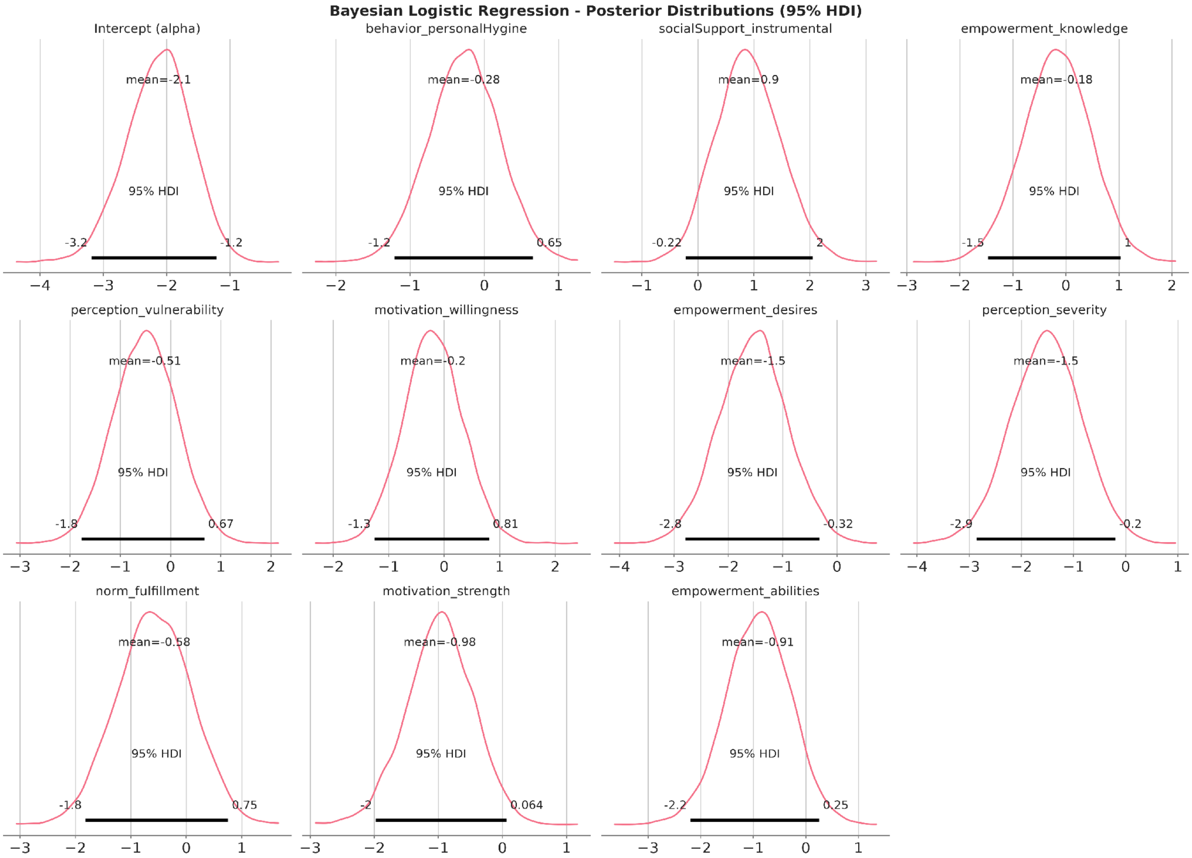

Figure 9: Posterior distributions for Bayesian logistic regression coefficients with 95% Highest Density Intervals (HDI). Distributions for empowerment_desires and perception_severity show 95% HDIs excluding zero, indicating statistically significant protective effects.

Bayesian logistic regression was fitted using the top 10 features by Random Forest importance to avoid overfitting. MCMC sampling achieved excellent convergence: all R-hat values = 1.00 (target < 1.01) and minimum effective sample sizes exceeding 7403, indicating reliable posterior estimates(fig 10).

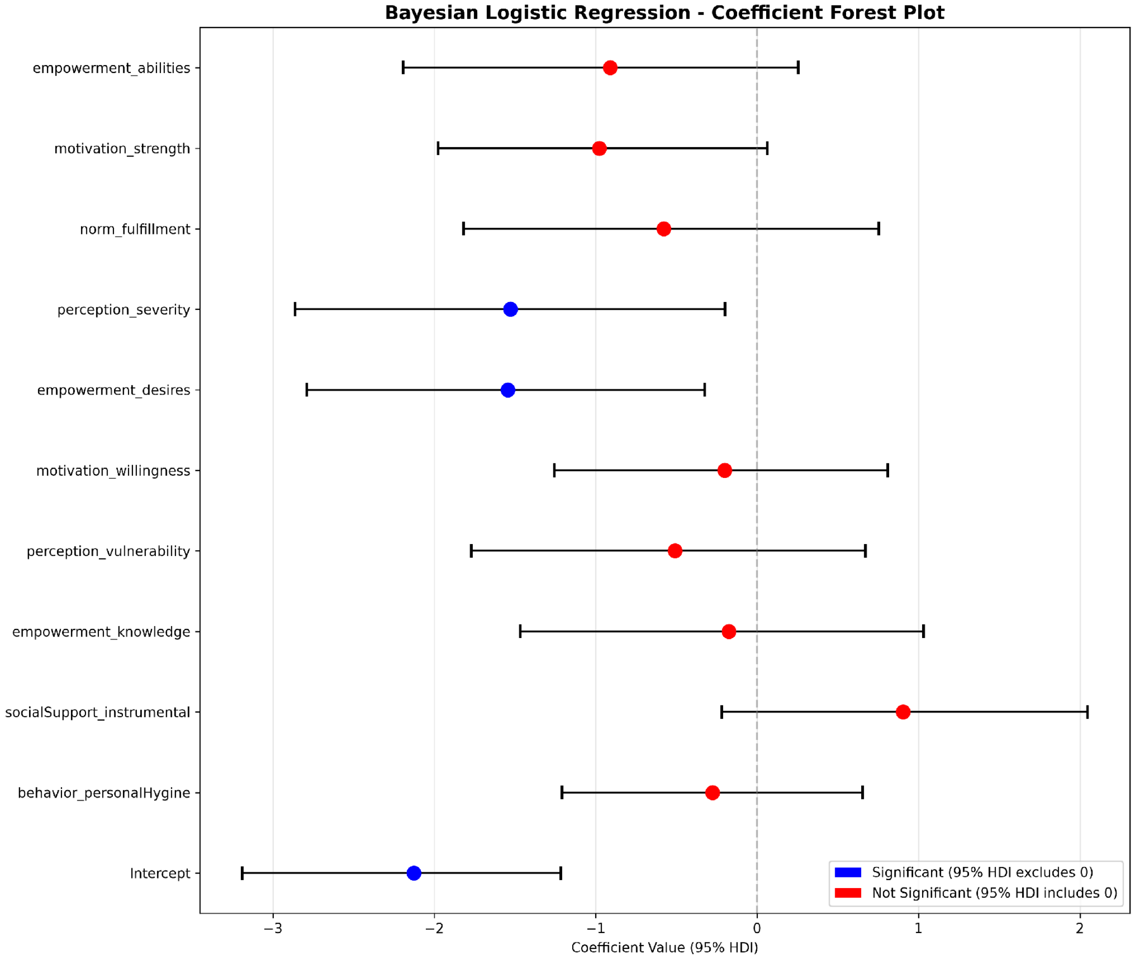

Figure 10: Bayesian coefficient forest plot showing mean estimates and 95% HDIs. Blue points indicate significant coefficients (95% HDI excludes zero); red points indicate non-significant coefficients.

Table 5: Bayesian Logistic Regression Coefficients (95% HDI)

Parameter	Mean	SD	2.5%	97.5%
alpha	-2.127	0.511	-3.19	-1.217
betas[0]	-0.276	0.476	-1.209	0.653
betas[1]	0.903	0.584	-0.218	2.047
betas[2]	-0.176	0.64	-1.467	1.032
betas[3]	-0.51	0.63	-1.77	0.672
betas[4]	-0.201	0.531	-1.257	0.809
betas[5]	-1.545	0.627	-2.79	-0.324
betas[6]	-1.528	0.678	-2.862	-0.2
betas[7]	-0.578	0.67	-1.819	0.753
betas[8]	-0.978	0.525	-1.976	0.064
betas[9]	-0.911	0.622	-2.192	0.254

Two features showed statistically significant negative associations with cancer risk (95% HDI excluding zero). First, empowerment_desires (β = -1.545, 95% HDI: [-2.790, -0.324]) indicates that women with higher desires for health empowerment had substantially lower odds of cervical cancer. This effect size corresponds to an odds ratio of approximately 0.21, meaning a one-unit increase in empowerment_desires is associated with a 79% reduction in odds of cancer, holding other factors constant.

Second, perception_severity (β = -1.528, 95% HDI: [-2.862, -0.200]) was significantly associated with reduced cancer risk. This finding suggests that women who perceived cervical cancer as more severe were less likely to have cancer, likely because they engaged in protective behaviors such as screening and early treatment. The odds ratio of approximately 0.22 indicates a 78% reduction in odds per unit increase in severity perception.

These findings are consistent with Health Belief Model predictions: women who recognize the severity of cervical cancer and desire greater control over their health appear to engage in protective behaviors that reduce their cancer risk. The Bayesian credible intervals provide direct probabilistic interpretation: there is a 95% probability that the true effect of empowerment_desires lies between -2.790 and -0.324, given the observed data and prior assumptions.

5. Discussion and Conclusions

5.1 Summary of Key Findings

This study demonstrates that machine learning and Bayesian statistical methods can achieve excellent discriminative performance for cervical cancer risk prediction based on behavioral and psychological factors. The convergence of evidence across multiple analytical approaches strengthens confidence in the key findings.

First, unsupervised clustering identified distinct behavioral phenotypes with dramatically different cancer prevalence rates (60.7% vs 9.1%). This finding suggests that behavioral patterns are strongly associated with cancer status and that population stratification based on psychosocial profiles could inform risk-based screening strategies. The high-risk phenotype was characterized by low empowerment, limited social support, and low perceived severity, factors that are potentially modifiable through targeted interventions.

Second, classification models achieved near-perfect discrimination, with Logistic Regression attaining AUC-ROC = 0.992. The superior performance of this relatively simple linear model suggests that behavioral factors have approximately linear relationships with cancer risk in log-odds space. This finding has practical implications: simpler models may be preferred for clinical deployment due to their interpretability and computational efficiency, particularly in resource-constrained settings.

Third, Bayesian analysis identified empowerment_desires and perception_severity as statistically significant protective factors with large effect sizes ($\beta \approx -1.5$). The Bayesian approach provided direct uncertainty quantification through credible intervals, enabling probabilistic statements about the likely range of true effect sizes. This uncertainty information is valuable for clinical decision-making, where understanding the precision of risk estimates affects screening prioritization.

5.2 Clustering Insights and Patient Stratification

The identification of distinct behavioral phenotypes through clustering has important implications for personalized prevention strategies. The high-risk cluster (Cluster 0), comprising 39% of the sample but 81% of cancer cases, represents a group that could benefit from intensive intervention. These women exhibited low scores across multiple protective constructs: empowerment (abilities, desires, knowledge), social support, and perceived severity.

The clustering results suggest that a one-size-fits-all approach to cervical cancer prevention may be suboptimal. Instead, risk-stratified interventions targeting the specific barriers faced by each phenotype could improve effectiveness. For the high-risk phenotype, interventions might focus on building empowerment through health education, strengthening social support networks, and enhancing perceived severity through culturally appropriate messaging about cervical cancer consequences.

5.3 Classification Model Performance

The excellent performance of all classification algorithms (AUC > 0.94) suggests that behavioral and psychosocial factors are strong predictors of cervical cancer status. The superiority of Logistic Regression over more complex ensemble methods (Random Forest, XGBoost) is noteworthy and has several potential explanations. First, the sample size (n=72) may be insufficient for ensemble methods to demonstrate their advantages in capturing complex interactions. Second, the true relationship between predictors and outcome may be approximately linear, making the additional flexibility of non-linear methods unnecessary.

The high precision of Random Forest (0.920) makes it potentially valuable for screening applications where false positives have significant costs, including unnecessary anxiety, additional testing, and healthcare resource utilization. The choice between models for clinical deployment would depend on the specific cost structure of false positives versus false negatives in the target healthcare setting.

5.4 Bayesian Model Insights

The Bayesian approach provided valuable insights beyond those available from frequentist methods. The posterior distributions quantified uncertainty in coefficient estimates, revealing that while empowerment_desires and perception_severity have robust negative associations with cancer risk (95% HDIs well below zero), other factors have more uncertain effects that could range from protective to harmful across the posterior distribution.

The interpretation of Bayesian results aligns with clinical intuition. The finding that empowerment_desires is protective suggests that women who actively seek health knowledge and control over their health decisions engage in protective behaviors. Similarly, the protective effect of perceived severity supports Health Belief Model predictions: women who recognize cervical cancer as serious are motivated to seek screening and treatment.

The Bayesian framework enables direct probability statements that facilitate clinical communication. Rather than stating that empowerment_desires has a 'significant' effect (frequentist interpretation), we can state that there is a 95% probability that the true effect lies between -2.79 and -0.32, and virtually 100% probability that the effect is negative (protective). This probabilistic interpretation better reflects the uncertainty inherent in medical decision-making.

5.5 Kenya Context and Implementation Implications

Kenya faces a cervical cancer crisis with incidence rates four times higher than in high-income countries. The findings of this study have important implications for Kenyan cervical cancer prevention strategies. The identification of modifiable protective factors, empowerment and perceived severity provides actionable targets for intervention.

Mobile Health (mHealth) Integration: The high classification accuracy suggests these models could be deployed via mobile health platforms, which have high penetration in Kenya. Women could complete brief behavioral assessments via smartphone, with high-risk individuals prioritized for screening appointments. This triage approach could maximize the impact of limited screening resources by ensuring that women most likely to benefit from screening are seen first.

Empowerment Interventions: The identification of empowerment desires as the strongest protective factor suggests that interventions enhancing women's desire for health autonomy could reduce cervical cancer burden. Kenyan programs should incorporate empowerment components, including health education emphasizing women's right to preventive care and building skills for navigating the healthcare system. Community health workers could be trained to deliver empowerment-focused counseling.

Severity Communication: The protective effect of perceived severity aligns with Health Belief Model predictions. Kenyan health communication campaigns should emphasize the serious consequences of untreated cervical cancer while simultaneously conveying the message that early detection leads to a cure. Messaging must balance raising awareness of severity without creating fatalism that could discourage screening.

5.6 Comparison with Prior Literature

Our findings align with previous research on cervical cancer risk factors in African contexts. Mabele et al. (2020) found that self-efficacy and perceived severity were associated with screening uptake among Tanzanian women. Mutua et al. (2018) identified a lack of awareness and low perceived susceptibility as barriers to screening in Kenya. Our study extends these findings by quantifying the protective effects using advanced statistical methods and demonstrating excellent predictive performance.

The superior performance of Logistic Regression compared to more complex algorithms contrasts with some ML literature but is consistent with studies showing that simpler models often perform comparably to complex methods in small samples. The key contribution of our study is not the superiority of any single algorithm but the demonstration that behavioral factors can achieve near-perfect discrimination, supporting their use in risk stratification.

5.7 Limitations

Several limitations must be acknowledged. First, the sample size (n=72) limits generalizability and statistical power. While the Bayesian approach provides regularization that helps with small samples, the findings require replication in larger cohorts. Second, the cross-sectional observational design precludes causal inference. The protective associations may reflect reverse causation (cancer diagnosis affecting psychological states) or unmeasured confounding.

Third, the dataset does not originate from Kenya specifically. While the behavioral constructs are theoretically transferable, cultural differences in item interpretation may limit direct applicability. Fourth, important clinical and demographic risk factors (age, HPV status, HIV status, sexual history) were not available and could confound the observed associations. Finally, the binary cancer outcome lacks staging information that would enable more nuanced risk prediction.

5.8 Future Research Directions

Several research directions are warranted. First, prospective validation in Kenyan cohorts is essential to assess real-world predictive performance and generalizability. Second, combining behavioral factors with clinical data (HPV testing, cytology) could enhance prediction and enable integrated risk stratification. Third, intervention studies testing whether empowerment and severity communication interventions modify behavioral phenotypes and cancer outcomes are needed. Fourth, the development of explainable AI models suitable for clinical decision support would facilitate translation to practice.

5.9 Conclusion

This study demonstrates that machine learning and Bayesian approaches can achieve excellent performance for cervical cancer risk prediction using behavioral and psychological factors. The identification of empowerment and perceived severity as protective factors provides actionable targets for intervention in resource-constrained settings.

The convergence of evidence across clustering, classification, and Bayesian analyses strengthens confidence in the importance of psychological empowerment for cancer prevention. The identification of distinct behavioral phenotypes suggests that risk-stratified approaches could improve screening efficiency. The Bayesian framework provided valuable uncertainty quantification that supports clinical decision-making.

These findings support the development of risk-stratified screening approaches for Kenya and similar settings, where mHealth platforms could deploy these models to enhance early detection and reduce the substantial burden of cervical cancer mortality. Future research should focus on prospective validation and intervention development to translate these findings into improved health outcomes for women in sub-Saharan Africa.

References

1. Arbyn, M., Weiderpass, E., Bruni, L., de Sanjosé, S., Saraiya, M., Ferlay, J., & Bray, F. (2020). Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. *The Lancet Global Health*, 8(2), e191-e203.

2. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

3. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

4. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.

5. Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository*. University of California, School of Information and Computer Science, Irvine, CA.

6. Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457-472.

7. Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1), 1593-1623.

8. IARC. (2020). *Global Cancer Observatory: Cancer Today*. International Agency for Research on Cancer, World Health Organization.

9. Kihara, M., Kassa, S., Odundo, E., Njoroge, M., Kataije, A., & Wambua, J. (2022). Implementation of a school-based human papillomavirus vaccination program in Kenya: lessons learned. *Journal of Global Health Reports*, 6, e2022136.

10. Kruschke, J. K. (2018). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan* (2nd ed.). Academic Press.

11. Mabele, M. A., Yansaneh, A., Aron, N., Bakam, C., Banke-Thomas, A., & Williams, N. (2020). Women's self-efficacy and cervical cancer screening uptake in Tanzania. *International Journal of Gynecology & Obstetrics*, 149(3), 341-348.

12. Ministry of Health, Kenya. (2020). *National Cancer Control Strategy 2017-2022*. Nairobi: Government of Kenya.

13. Mutua, K. D., Ochako, R., Wanyonyi, Z., Echoka, E., & Fegan, G. (2018). Barriers to cervical cancer screening and treatment among women in Kenya: a systematic review. *International Journal of Community Medicine and Public Health*, 5(11), 4655-4663.

14. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31-38.

15. Rono, R. K., Stallman, J. S., Rohner, E., & Jaoko, W. (2020). Cervical cancer screening practices and associated factors among women aged 25-49 years in Kenya. *East African Medical Journal*, 97(3), 3152-3162.

16. Rosenstock, I. M. (1974). Historical origins of the health belief model. *Health Education Monographs*, 2(4), 328-335.

17. Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

18. Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2, e55.

19. Shepherd, D., Khalil, M., & Kersten, F. (2020). Clinical risk prediction in cervical cancer. *Best Practice & Research Clinical Obstetrics & Gynaecology*, 66, 91-102.

20. Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236-244.

21. WHO. (2021). *Global Strategy to Accelerate the Elimination of Cervical Cancer as a Public Health Problem*. Geneva: WHO Press.
