Journal of Multidisciplinary Healthcare

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

A Machine Learning and Bayesian Belief Network Approach to Predicting Cervical Cancer Risk: Implications for Risk Management

Khaled Toffaha[1], Mecit Can Emre Simsekler[1], Andrei Sleptchenko[1], Michael A Kortt[1], Laurette L Bukasa[2]

[1]Department of Management Science & Engineering, Khalifa University of Science & Technology, Abu Dhabi, United Arab Emirates; [2]Abu Dhabi Health Data Services, M42, Abu Dhabi, United Arab Emirates

Correspondence: Khaled Toffaha; Mecit Can Emre Simsekler, Department of Management Science & Engineering, Khalifa University of Science & Technology, Abu Dhabi, United Arab Emirates, Email Khaled.mToffaha@ku.ac.ae; Emre.Simsekler@ku.ac.ae

Introduction: Cervical cancer remains a major global health challenge, necessitating enhanced risk stratification and early detection methodologies. This study proposes a comprehensive predictive framework for cervical cancer leveraging advanced machine learning (ML) algorithms and Bayesian Belief Networks (BBNs), illustrating the transformative role of digital technologies in healthcare and education within an increasingly digitized society.

Methods: A cohort of 858 patients was analyzed, addressing data challenges, including missing values, class imbalance, and nonlinear feature interactions, that frequently compromise the reliability of predictive modeling. Methodologically, this study integrated advanced data science approaches, including multiple imputation, feature selection, and imbalance mitigation, advancing medical analytics to ensure robust model generalizability.

Results: High predictive performance was observed across different cervical cancer screening tests. The combined target ML model achieved an accuracy of 95.6%, an area under the receiver-operating characteristic curve (AUROC) of 0.958, and an F1-score of 0.945. The BBN, built upon the Bayesian Additive Regression Trees (BART) model, demonstrated a positive prediction rate (sensitivity) of 91.3% and a negative prediction rate (specificity) of 86.8%.

Discussion: These results validate the technical efficacy of the proposed framework and underscore its potential for integration into clinical decision-support systems. Beyond clinical applications, this research contributes to computational oncology by demonstrating the synergistic potential of combining probabilistic graphical models with ML techniques. The study highlights the critical role of interdisciplinary collaboration between clinical experts and data scientists in creating effective AI healthcare solutions. It also emphasizes the need for upskilling healthcare workers and optimizing healthcare delivery processes to fully realize the benefits of precision medicine.

Keywords: risk management, cervical cancer risk prediction, future of healthcare, cancer risk factors, Bayesian belief network, machine learning, digital health, patient safety

Introduction

Cervical cancer continues to represent a major global public health concern, particularly in low- and middle-income countries (LMICs), where approximately 90% of cervical cancer deaths occur despite the availability of effective screening and vaccination programs in high-income nations.[1] It is estimated that more than 600,000 new cases of cervical cancer are diagnosed each year worldwide, leading to more than 340,000 deaths.[2] The disparity in outcomes is mainly attributable to inequalities in access to preventive healthcare services, timely diagnosis, and treatment availability. In the United States alone, about 13,820 new cases of invasive cervical cancer were expected to be diagnosed in 2024, with an estimated 4360 women projected to die from the disease.[3] In addition to the human suffering, the economic burden is considerable, with the incremental cost of treating local and regional cervical cancers reaching up to $30,917 within the first year post-diagnosis.[4,5]

Received: 21 February 2025
Accepted: 28 May 2025
Published: 25 August 2025

Screening remains the most effective method for reducing cervical cancer incidence and mortality. Common screening modalities include the Papanicolaou (Pap) smear, which examines exfoliated cervical cells for cytological abnormalities; visual inspection with acetic acid (VIA), which highlights suspicious lesions using acetic acid; and visual inspection with Lugol's iodine (VILI), which relies on glycogen uptake by normal epithelial cells (Schiller test).[6,7] These are often followed by colposcopy and biopsy for confirmatory diagnosis. Despite their efficacy, these screening methods face significant limitations in LMICs due to costs, the need for trained personnel, and logistical barriers.[8,9] In such settings, there is a pressing need for scalable, low-cost alternatives to prioritize high-risk individuals for further screening.

ML has emerged as a promising strategy for risk stratification in cervical cancer screening. ML algorithms can learn complex, non-linear relationships from high-dimensional data, including sociodemographic, behavioral, and clinical variables. Such models can assist in identifying women at elevated risk, thereby enabling resource-efficient allocation of clinical screening services.[10,11] However, their interpretability is a critical barrier to the clinical adoption of ML models. Many high-performing models, including deep neural networks and gradient boosting machines, function as "black boxes", offering little insight into the rationale behind their predictions.[12,13] This lack of transparency undermines trust among clinicians and patients, complicating efforts to integrate such models into real-world workflows.

In response to this challenge, there has been a growing emphasis on developing explainable artificial intelligence (XAI) systems in healthcare. Interpretable models, such as decision trees, ensemble tree-based methods, and probabilistic graphical models like BBNs, strike a balance between predictive performance and explainability.[14,15] Decision trees structure decisions hierarchically, facilitating visual inspection and logical tracing of model outputs. Similarly, BBNs represent joint probability distributions over a set of variables using directed acyclic graphs, encoding conditional dependencies transparently and intuitively. These models are especially valuable in public health contexts where explainability is paramount for guiding behavioral change, promoting patient education, and facilitating shared decision-making.

This study uses patient survey data to leverage explainable ML models for cervical cancer risk prediction. Three primary objectives are pursued: (1) evaluation of the models' predictive performance in terms of discrimination and calibration, (2) assessment of their interpretability through feature importance analysis, and (3) identification and ranking of key risk factors contributing to cervical cancer risk. The approach is grounded in the hypothesis that explainable models can provide actionable insights for clinical practice and public health policymaking, particularly in low-resource environments where traditional screening infrastructure is lacking.

The remainder of the manuscript is structured as follows. Existing literature concerning ML applications in cervical cancer prediction is reviewed in Related Works, focusing on explainable modeling approaches. The dataset, feature selection process, and preprocessing techniques are described in Methods, along with the methodological framework and algorithms employed. Results from the experimental evaluations are presented in Results, while the implications of the findings and potential avenues for future research are discussed in Discussion and Conclusion.

Related Work

Over the past decade, the application of ML to cervical cancer risk prediction has been an active area of research. One of the most frequently utilized resources in this domain is the UCI Cervical Cancer (Risk Factors) dataset,[16] which encompasses a diverse range of variables from sociodemographic indicators to detailed medical histories. This section provides an in-depth review of seminal studies and methodological approaches that have attempted to enhance the early detection of cervical cancer using various ML and probabilistic frameworks, with particular emphasis on strategies to improve model interpretability and address data imbalance challenges.

Nithya et al have contributed to this field by evaluating several ML algorithms, including Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), in combination with advanced feature selection techniques such as Recursive Feature Elimination (RFE), the Boruta algorithm, and Simulated Annealing (SA).[17] Their systematic assessment revealed that the Boruta algorithm, by selecting features based on their contribution to the model, offered a clearer interpretative framework for understanding the underlying risk factors, thereby enhancing the clinical relevance of the predictive models. Notably, however, the study did not sufficiently address the persistent challenge of data imbalance, which is common in medical datasets and can lead to biased predictive performance.

Alsmariy et al approached the cervical cancer prediction problem by integrating an ensemble voting mechanism that combined the predictive power of Decision Trees (DTs), Logistic Regression (LR), and RF classifiers.[18] In addition, their incorporation of the Synthetic Minority Over-sampling Technique (SMOTE) mitigated the adverse effects of imbalanced class distributions, while Principal Component Analysis (PCA) was employed to reduce the dimensionality of the feature space. This methodological synergy resulted in marked improvements in classification accuracy, sensitivity, and the area under the receiver operating characteristic (AUROC) curves, thus highlighting the importance of pre-processing steps in achieving robust ML outcomes.

Similarly, Al Mudawi et al implemented a comprehensive workflow combining SMOTE and PCA to concurrently address issues of class imbalance and high-dimensionality.[19] Evaluating several algorithms ranging from DTs and RF to AdaBoost and SVM- their study reported high classification scores for multiple screening targets. While these results highlight the potential of ensemble learning, they also raise concerns regarding overfitting and the need for careful external validation. Additionally, by integrating a survey-based assessment of cervical cancer awareness, the study underscored the broader socio-educational context within which cervical cancer interventions must be situated.

More recently, Ashtagi et al have demonstrated that a Bagging Classifier, based on DT estimators, achieves a high accuracy (97%) for the early detection of cervical cancer.[20] Despite the high classification performance, this study did not fully explore strategies for addressing data imbalance or implementing feature selection methodologies, which are critical for ensuring the transparency and generalizability of predictive models.

Beyond tree-based models, the application of BBNs represents a paradigm that inherently embeds interpretability into the modeling process. Langberg et al illustrated how BBNs could be employed to capture the causal relationships between patient risk factors and cervical cancer outcomes.[15] This approach facilitates high-quality predictions and offers a robust framework for sensitivity analysis and scenario simulation that is particularly valuable for formulating public health policies in resource-limited settings. Despite the promising utility of BBNs, challenges remain in terms of the robust estimation of network parameters, particularly when available data are sparse or noisy.

Complementing these methodological advancements, Zhao et al have investigated the incorporation of cost-sensitive learning within cervical cancer screening models tailored for low-resource settings, thereby optimizing decision thresholds to balance the cost of false negatives against the operational constraints of healthcare systems.[8] Such cost-sensitive frameworks are increasingly relevant in guiding resource allocation decisions and ensuring that predictive technologies are economically viable and clinically impactful.

In summary, the literature emphasizes the trade-offs between predictive performance and model interpretability in the context of cervical cancer risk prediction. While ensemble methods and deep learning approaches offer high accuracy, their lack of transparency limits clinical acceptance. Contrarily, explainable models like BBNs, though sometimes less competitive in terms of raw predictive metrics, provide critical insights into the risk factors underpinning cervical cancer development. Building on these insights, the present study focuses on advancing explainable ML frameworks and thoroughly evaluating their performance, interpretability, and clinical utility in a real-world setting.

Methods

Ethical Statement

This study uses publicly available anonymized cervical cancer data, obtained from the UCI Machine Learning Repository and collected at Hospital Universitario de Caracas in Caracas, Venezuela, exempting it from ethics review under institutional guidelines. This dataset complies with UCI's terms of use and relevant ethical guidelines for secondary data analysis.

The dataset comprises demographic information, habits, and historical medical records of 858 patients. It contains 36 attributes, including 32 input features and 4 target variables (Hinselmann, Schiller, Cytology, Biopsy), as presented in Table 1. Several patients declined to answer some questions due to privacy concerns, resulting in missing values for specific attributes. This real-world dataset provides a valuable opportunity to develop and evaluate predictive models for cervical cancer while addressing the common challenge of incomplete data in medical research.

Table 1 Summary of Features

S. No.	Attribute Name	Type	Description
1	Age	Int	Age of the patient
2	Number of sexual partners	Int	Number of sexual partners
3	First sexual intercourse (age)	Int	Age at first sexual intercourse
4	Num of pregnancies	Int	Number of pregnancies
5	Smokes	Bool	Whether the patient smokes
6	Smokes (years)	Bool	Number of years the patient has smoked
7	Smokes (packs/year)	Bool	Number of packs per year the patient smokes
8	Hormonal contraceptives	Bool	Whether the patient uses hormonal contraceptives
9	Hormonal contraceptives years	Int	Number of years the patient has used hormonal contraceptives
10	IUD	Bool	Whether the patient uses an intrauterine device (IUD)
11	IUD (years)	Int	Number of years the patient has used an IUD
12	STDs	Bool	Whether the patient has a history of Sexually Transmitted Disease
13	STDs (number)	Int	Number of STDs the patient has had
14	STDs: condylomatosis	Bool	Whether the patient has had condylomatosis
15	STDs: cervical condylomatosis	Bool	Whether the patient has had cervical condylomatosis
16	STDs: vaginal condylomatosis	Bool	Whether the patient has had vaginal condylomatosis
17	STDs: vulvo-perineal condylomatosis	Bool	Whether the patient has had vulvo-perineal condylomatosis
18	STDs: syphilis	Bool	Whether the patient has had syphilis
19	STDs: pelvic inflammatory disease	Bool	Whether the patient has had pelvic inflammatory disease
20	STDs: genital herpes	Bool	Whether the patient has had genital herpes
21	STDs: molluscum contagiosum	Bool	Whether the patient has had molluscum contagiosum
22	STDs: AIDS	Bool	Whether the patient has had AIDS
23	STDs: HIV	Bool	Whether the patient has had HIV
24	STDs: Hepatitis B	Bool	Whether the patient has had Hepatitis B
25	STDs: HPV	Bool	Whether the patient has had HPV
26	STDs: Number of diagnosis	Int	Number of STD diagnoses
27	STDs: time since first diagnosis	Int	Time since first STD diagnosis
28	STDs: time since last diagnosis	Int	Time since last STD diagnosis
29	Dx: cancer	Bool	Diagnosis of any cancer
30	Dx: CIN	Bool	Diagnosis of cervical intraepithelial neoplasia (CIN)
31	Dx: HPV	Bool	Diagnosis of HPV
32	Dx	Bool	Diagnosis of any condition
33	Hinselmann: target variable	Bool	Hinselmann test result (target variable)

(*Continued*)

Table 1 (Continued).

S. No.	Attribute Name	Type	Description
34	Schiller: target variable	Bool	Schiller test result (target variable)
35	Cytolagy: target variable	Bool	Cytology test result (target variable)
36	Biopsy: target variable	Bool	Biopsy test result (target variable)

Missing Data Imputation

To address the issue of missing data and maximize the utility of the available information, four different imputation methods were employed: missing flag, forward filling using mean, MissForest, and MICE (Multivariate Imputation by Chained Equations). The missing flag approach created binary indicators to denote the presence or absence of missing values for each feature, allowing models to learn patterns related to missingness potentially. Forward filling using the mean provided a simple baseline imputation strategy by replacing missing values with the mean of the respective feature. MissForest, a non-parametric method, uses random forests to impute missing values based on the observed values of other variables, capturing complex relationships between variables and handling mixed data types. Finally, MICE created multiple imputations for multivariate missing data using chained equations, which is particularly useful for handling complex missing data patterns and preserving relationships between variables.

By employing these diverse imputation techniques, the impact of different missing data handling approaches on the predictive performance of the models was assessed. This comprehensive approach enables a robust evaluation of the dataset's potential for cervical cancer risk prediction.

Feature Selection

To identify the most relevant predictors for cervical cancer and optimize model performance, four feature selection methods were implemented: full original data, Genetic Algorithm (GA), Recursive Feature Elimination (RFE), and Sequential Forward Selection (SFS). The full original data approach retained all features from the original dataset as a baseline for comparison. The Genetic Algorithm, an evolutionary approach, was used to select an optimal subset of features, allowing for the exploration of complex feature interactions. Recursive Feature Elimination recursively removes features based on their importance in the model, helping identify a compact set of highly predictive features. Sequential Forward Selection built the feature set incrementally by adding features based on their contribution to performance improvement. The application of these diverse feature selection methods enables us to compare their effectiveness in identifying the most informative predictors for cervical cancer risk. This multi-faceted approach to feature selection provides a comprehensive evaluation of the dataset's predictive potential.

Machine Learning Models

A comprehensive set of ML algorithms was trained on the datasets resulting from each feature selection method. The algorithms included decision trees, ensemble methods (eg, Random Forests, Extra Trees), boosting algorithms (eg, XGBoost, LightGBM), and BART. This diverse set of algorithms allows us to explore different modeling approaches and identify the most effective predictors for cervical cancer risk.

Decision trees provide interpretable models with clear decision rules, while ensemble methods leverage the power of multiple models to improve predictive performance and robustness. Boosting algorithms utilize sequential learning to create strong predictive models, and BART combines decision trees with Bayesian inference for flexible and interpretable modeling. By employing this range of algorithms, the predictive capabilities of the dataset and feature selection methods can be thoroughly evaluated.

Bayesian Belief Network

Building upon the insights gained from the ML models, a BBN was constructed to create a probabilistic model for cervical cancer prediction. The BBN provides a probabilistic framework for reasoning about cervical cancer risk, allowing for interpretable predictions and the incorporation of expert knowledge. By integrating ML algorithms with a Bayesian network approach, the methodology aims to leverage the strengths of both paradigms - the predictive power of modern ML techniques and the interpretability and reasoning capabilities of probabilistic graphical models.

This comprehensive approach allows us to develop a robust and insightful predictive model for cervical cancer risk assessment, combining the strengths of various data imputation, feature selection, and modeling techniques. The resulting model has the potential to provide valuable insights for clinical decision-making and patient care in the context of cervical cancer prevention and early detection.

Results

The dataset contained varying levels of missingness across different variables, ranging from 0.8% to 13.6% of total values. The variables with the highest percentage of missing values (13.6%) were IUD and IUD years, followed closely by Hormonal Contraceptives and Hormonal Contraceptives years at 12.6%. Several STD-related variables had 12.2% missing values. Lower levels of missingness were observed for variables such as the number of pregnancies (6.5%), number of sexual partners (3%), and smoking-related variables (1.5%), as shown in Figure 1. To capture complex relationships and handle mixed data types, missing flags and forward filling were utilized using mean, MissForest, and MICE. The performance of all four methods was compared using cross-validation, focusing on their impact on downstream analyses and predictive models. This comprehensive approach allowed us to develop a robust strategy for handling missing data in the proposed cervical cancer prediction model, potentially combining different methods based on the nature and extent of missingness in each variable.

To identify the most relevant predictors for cervical cancer and optimize model performance, four feature selection methods were implemented: full original data, GA, RFE, and SFS. Recognizing the limitations of these methods for multi-class feature identification, a comprehensive approach was adopted by testing two paths. In the first path, the four target variables (Hinselmann, Schiller, Cytology, and Biopsy) were merged into a single target. Each feature selection

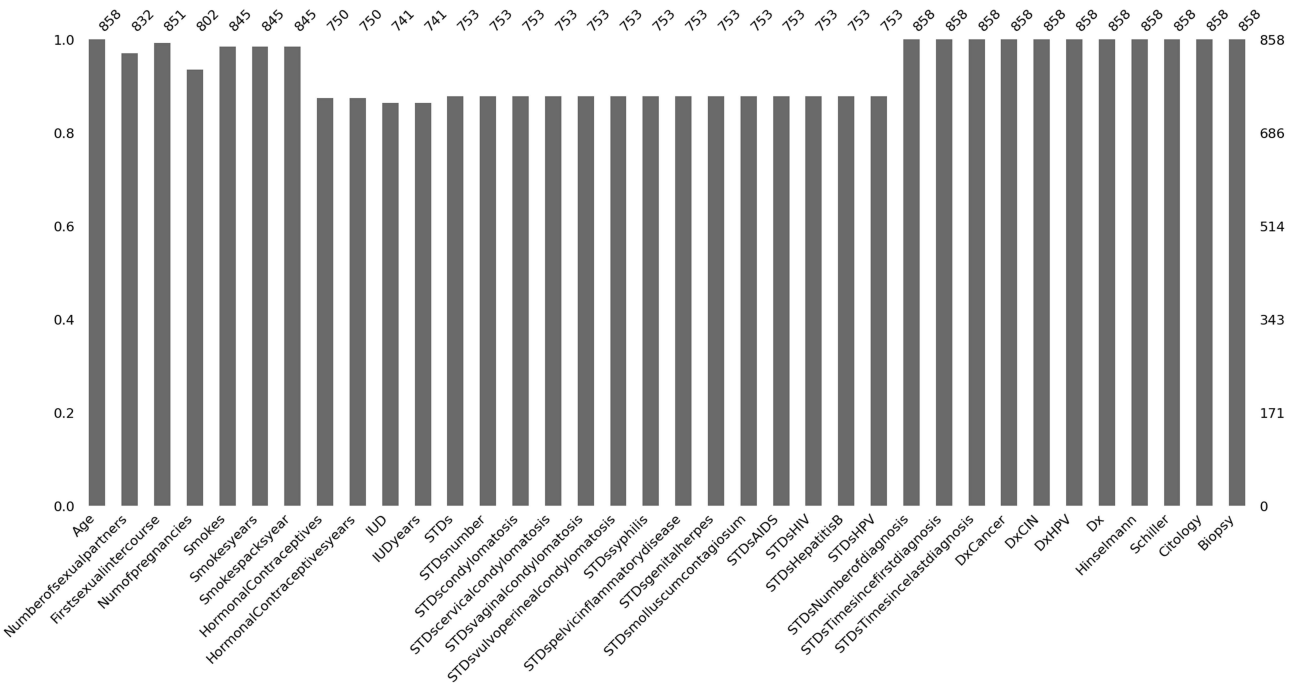

Figure 1 Missing Values Bar Plot.

method was applied to identify essential features for this combined target. In the second path, each feature selection method was applied to each target variable independently before combining the results. By comparing and combining the results from both paths, a robust list of relevant predictors was developed, capturing both overall and specific feature importance as shown in Table 2. This dual-path approach enhances the reliability and interpretability of the feature selection process, providing a solid foundation for subsequent model development and analysis.

After identifying the most important features, several data imbalance treatment methods were employed to address the class imbalance in the dataset. For oversampling, ADASYN (Adaptive Synthetic), Randomized Over-Sampling, and SMOTE (Synthetic Minority Over-sampling Technique) were utilized. These methods aim to increase the number of

Table 2 Summary of the Most Important Features

Combined Targets	Schiller Target	Hinselmann Target	Cytology Target	Biopsy Target
Smokes				Smokes
Smokes years		Smokes years		
Smokes packs/year				
	Num of pregnancies			Num of pregnancies
	Hormonal Contraceptives years	Hormonal Contraceptives years		
	IUD years			
			IUD	
STDs vaginal condylomatosis				
STDs vulvo-perineal condylomatosis	STDs vulvo-perineal condylomatosis		STDs vulvo-perineal condylomatosis	
	STDs pelvic inflammatory disease			
STDs syphilis				
STDs genital herpes	STDs genital herpes			
STDs Time since last diagnosis			STDs Time since last diagnosis	
	STDs molluscum contagiosum			
	STDs HIV			
Dx CIN			Dx CIN	Dx CIN
Dx HPV				Dx HPV
	Dx		Dx	Dx
			First sexual intercourse	First sexual intercourse
			STDs	
				STDs cervical condylomatosis
				STDs AIDS

(*Continued*)

https://doi.org/10.2147/JMDH.S524132

Table 2 (Continued).

Combined Targets	Schiller Target	Hinselmann Target	Cytology Target	Biopsy Target
				STDs Number of diagnosis
				Age
		Number of sexual partners		

minority class samples by creating synthetic examples. For under sampling, Edited Nearest Neighbors, Near Miss, One-Sided Selection, and Randomized UnderSampling techniques were applied to reduce the number of majority class samples and balance the dataset. Additionally, combined approaches including SMOTETomek and SMOTEEN were explored, integrating both oversampling and undersampling strategies. These methods were systematically applied and evaluated to determine their impact on the model's performance in predicting cervical cancer risk.

Following feature selection and class imbalance treatment, a diverse set of ML algorithms was implemented and evaluated to construct predictive models for cervical cancer risk. The algorithms employed in this study included Random Forest Classifier, Gradient Boosting Classifier, AdaBoost Classifier, Decision Tree Classifier, XGBoost Classifier, and BART. To rigorously assess these models' performance, a comprehensive set of evaluation metrics was utilized: Accuracy, Area Under the Receiver Operating Characteristic curve (AUROC), Geometric Mean (G-Means), Precision, Recall, and F1-score. These metrics were chosen to provide a holistic view of model performance, capturing overall correctness, discriminative ability, and balance between sensitivity and specificity. A stratified k-fold cross-validation strategy (k=5) was employed to ensure robust performance estimation and mitigate overfitting. For each target variable (Combined, Schiller, Hinselmann, Cytology, and Biopsy), all combinations of imputation methods, feature selection techniques, and ML algorithms were systematically evaluated. This comprehensive approach enables the identification of the optimal model configurations for each target variable.

The results revealed that different combinations of methods yielded the best performance for each target. For the Combined Target and Schiller Test, Forward Filling Imputation and Genetic Algorithm feature selection, coupled with BART and SMOTEEN, proved the most effective. The Hinselmann Test benefited from MissForest imputation and Recursive Feature Elimination, while the Cytology Test performed best with MissForest imputation and Genetic Algorithm. For the Biopsy target, Iterative imputation and Sequential Forward Selection were optimal. Notably, BART and SMOTEEN were consistently selected across all optimal models, suggesting their robustness in handling the complexities of this dataset and effectively addressing class imbalance issues. These optimal configurations demonstrated superior performance across all evaluation metrics, with accuracies ranging from 0.956 to 0.992 and AUROC values between 0.833 and 0.958. The Biopsy model achieved the highest overall performance with an accuracy of 0.992 and an AUROC of 0.899. The variation in optimal imputation and feature selection methods across target variables underscores the importance of tailored approaches for each specific diagnostic test.

These results demonstrate the efficacy of the proposed comprehensive modeling approach in accurately predicting cervical cancer risk across various diagnostic modalities. By leveraging advanced ML techniques and a thorough evaluation process, models that show great promise in enhancing clinical decision-making processes for cervical cancer screening and diagnosis were developed. The high performance across multiple metrics suggests that these models could support healthcare professionals in identifying high-risk patients and optimizing screening strategies.

Based on the best-performing ML model, a BBN was constructed to further enhance the understanding of the relationships between variables and improve predictive accuracy for cervical cancer risk. Multiple discretization functions (k-means, MDLP, and CAIM) and learning methods (Chow-Liu, MIIC, TAN, GHC, Tabu, and NaiveBayes) were compared to identify the optimal approach. The Tree-Augmented Naive Bayes (TAN) learning method combined with

the CAIM discretization function yielded the best performance, resulting in a BBN with strong predictive capabilities, achieving a 91.3% positive prediction rate and an 86.8% negative prediction rate.

The structure of the resulting BBN provides valuable insights into the complex interplay of factors contributing to cervical cancer risk, as shown in Figure 2. At the center of the network is the "Diagnosis" node, directly connected to all other variables, underscoring its role as both an influencer and a result of the model's various risk factors and health indicators. This central position highlights the multifaceted nature of cervical cancer diagnosis and the importance of considering a wide range of factors in risk assessment. Surrounding the central diagnosis node, several key clusters of interconnected variables were observed. The smoking-related variables form one such cluster, with "Smokes" directly influencing "Smokes packs/year", reflecting the logical relationship between smoking status and quantity. Another significant cluster involves STD-related variables, revealing potential comorbidities or shared risk factors among conditions like vaginal condylomatosis, genital herpes, and syphilis.

A particularly interesting node in the network is "STDs Time since last diagnosis", which has multiple outgoing connections influencing smoking behavior, HPV diagnosis, and CIN diagnosis. This suggests that the recency of STD diagnoses may play a crucial role in predicting other health outcomes and behaviors, emphasizing the importance of regular screening and follow-up care. The diagnostic nodes for CIN and HPV are both influenced by the time since the last STD diagnosis, pointing to a temporal relationship between STD history and these specific cervical conditions. This connection underscores the potential long-term impacts of STDs on cervical health and the need for ongoing monitoring of patients with a history of STDs.

To further analyze the model's behavior and understand the impact of different variables on the diagnosis, a sensitivity analysis was conducted using the inference graph. The probability of having cervical cancer was set to 100% in the "Diagnosis" node, allowing observation of how this certainty propagated through the network and affected the relative probabilities of other variables, as shown in Figure 3.

By capturing these complex interactions and relationships, the BBN provides a powerful tool for understanding and predicting cervical cancer risk. The integration of the TAN learning method and CAIM discretization function has allowed us to create a model that performs well in predictive accuracy and offers interpretable insights into the factors contributing to cervical cancer risk. This approach enhances the ability to support clinical decision-making and risk assessment, potentially leading to more targeted and effective cervical cancer screening and prevention strategies.

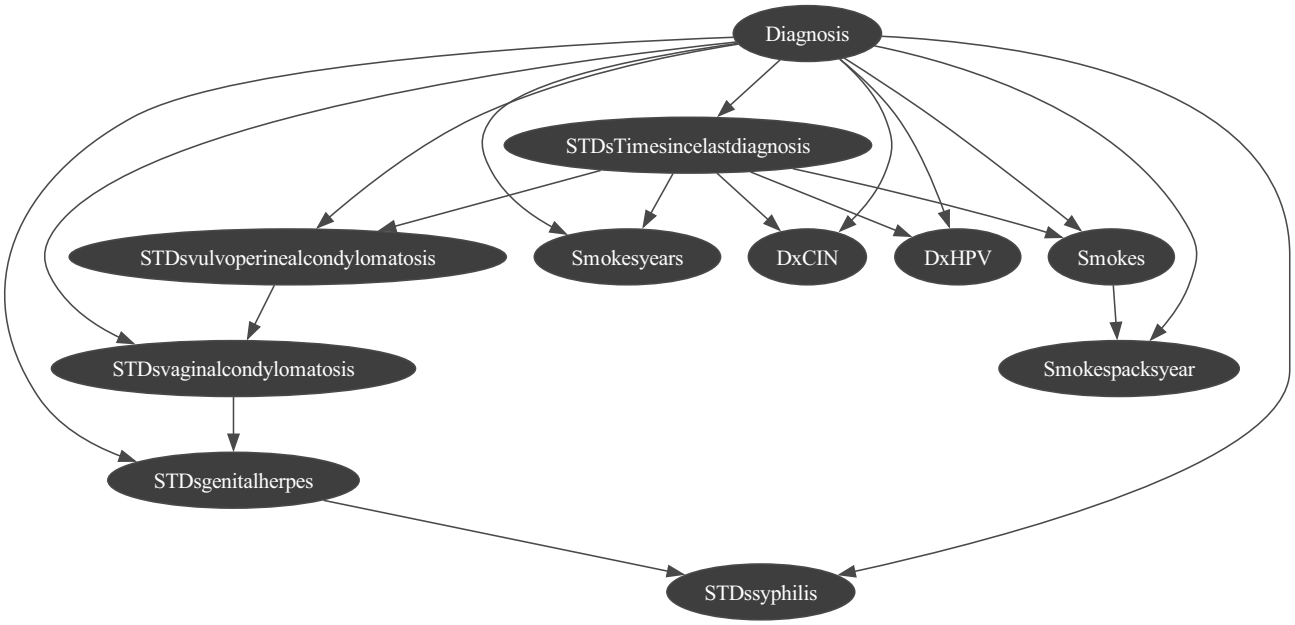

Figure 2 BBN Model.

https://doi.org/10.2147/JMDH.S524132

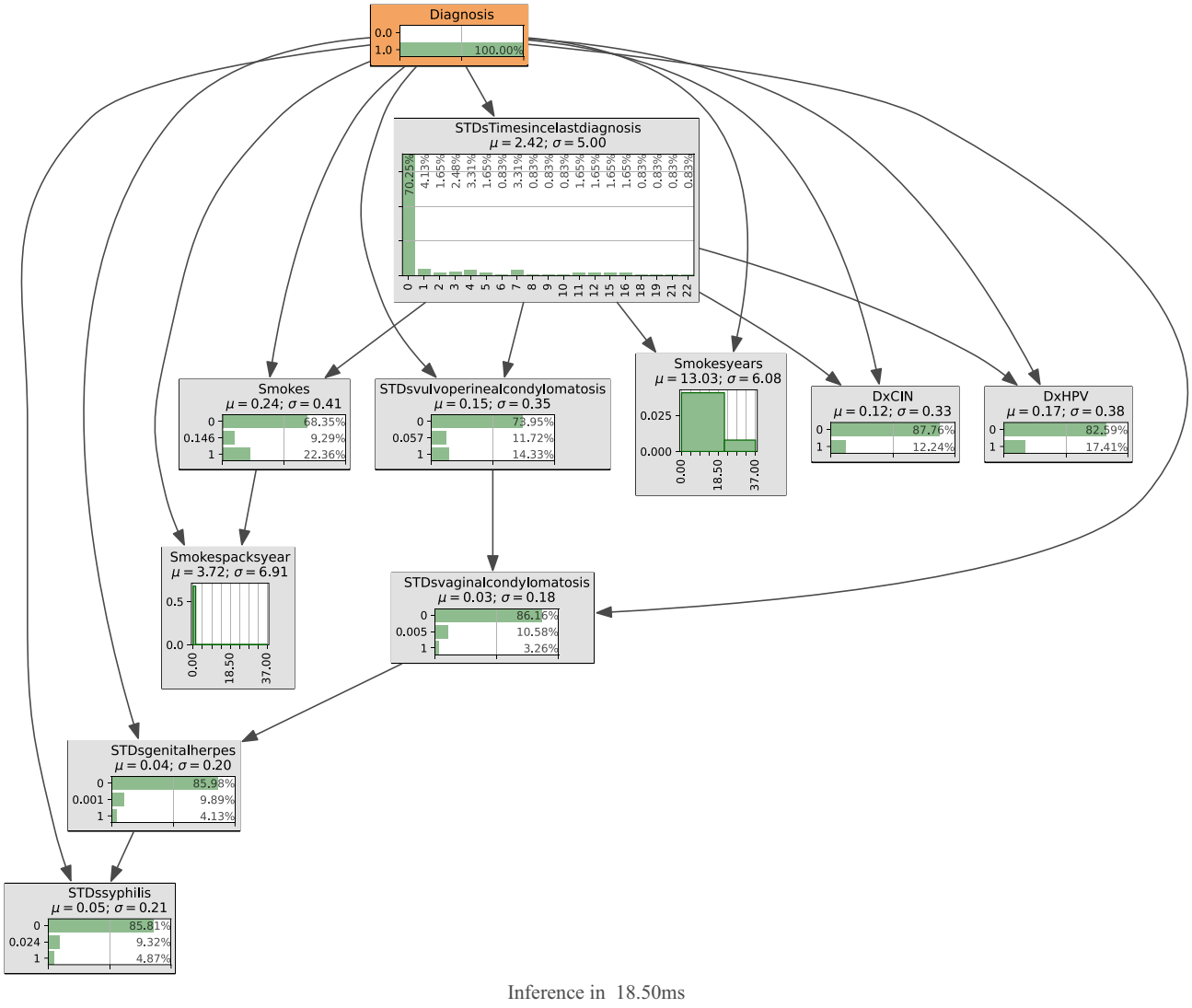

Figure 3 BBN Inference Model.

Discussion

The comprehensive study on cervical cancer prediction highlights significant patient safety risks, with critical findings identified that could profoundly influence clinical practice and guide future research. Through the employment of advanced ML techniques and BBNs, a robust framework for predicting outcomes in cervical cancer patients was developed. This innovative approach not only enhances prediction accuracy but also addresses the urgent need for improved screening methods, aiming to elevate patient care and safety in oncology settings.

Employing multiple methods across different targets, the feature selection process revealed a complex set of predictors for cervical cancer risk. As shown in Table 2, smoking-related variables, STD history, and previous diagnoses of CIN and HPV appeared as consistently important features across multiple targets. This highlights the multifaceted nature of cervical cancer risk and the importance of considering a wide range of factors in risk assessment. The feature selection process aligns well with the existing literature on cervical cancer risk factors[21] while highlighting new insights. As expected, key features relating to STD history, HPV positivity, and CIN cytology surfaced as important predictors across multiple targets, consistent with established knowledge about cervical cancer risk factors. However, the strong importance of smoking-related features was a somewhat surprising finding that refined the understanding of the risk factors. While smoking has been associated with cervical cancer in

Table 3 Performance Metrics for Cervical Cancer Prediction Models

Model	Accuracy	AUROC	G-Means	Precision	Recall	F1-score
Combined Target	0.956	0.958	0.958	0.935	0.958	0.945
Schiller Test	0.973	0.833	0.825	0.837	0.833	0.819
Hinselmann Test	0.990	0.870	0.863	0.880	0.870	0.863
Cytology Test	0.983	0.872	0.866	0.852	0.872	0.845
Biopsy	0.992	0.899	0.898	0.890	0.899	0.891

previous studies, its prominence as a key predictor across multiple targets in the performed analysis highlights its potential significance. It suggests it may play a more significant role in cervical cancer risk than previously recognized.

The high performance of the developed models across various evaluation metrics (Table 3) demonstrates the effectiveness of the implemented approach in predicting cervical cancer risk, with accuracies ranging from 0.956 to 0.992 and AUROC values between 0.833 and 0.958, indicating the potential of advanced analytics for improving cervical cancer prediction. The consistent superiority of BART and SMOTEEN across different targets highlights their robustness when handling complex, imbalanced datasets such as the one analyzed. These findings could inform future studies on medical data analysis, particularly in scenarios characterized by class imbalance.

The BBN in Figure 2, constructed using the Tree-Augmented Naive Bayes (TAN) method and CAIM discretization, achieved not only high predictive accuracy (91.3% positive prediction rate, 86.8% negative prediction rate) but also provided valuable insights into the complex interplay of factors contributing to cervical cancer risk. The central position of the "Diagnosis" node and its connections to all other variables emphasize the multifaceted nature of cervical cancer prediction. The identified clusters of interconnected variables, such as smoking-related factors and STD history, offer a nuanced understanding of risk factor relationships. The sensitivity analysis conducted using the inference graph 3 further elucidated the impact of different variables on cervical cancer diagnosis. The observed increases in probabilities for smoking and STD-related nodes when cervical cancer probability was set to 100% reinforce the strong associations between these factors and cervical cancer risk.

When considering the implications of this study for the future of education and work in the digitized society, several key points emerge. First, the complexity of the developed models and the diverse skill set required for their development and interpretation highlight the need for interdisciplinary education combining medical knowledge with data science and AI proficiency. Understanding which factors in the model are routinely available is essential to identifying high-risk women in real-world settings. Future healthcare professionals must demonstrate proficiency in working with and interpreting AI-driven tools, suggesting a necessary shift in medical education curricula. Moreover, the integration of advanced analytics in healthcare, as demonstrated by this study, may lead to changes in workforce needs. While increased demand for data scientists and AI specialists in healthcare is anticipated, existing healthcare workers may also require retraining to use these new tools effectively. This emphasizes the importance of lifelong learning and adaptable, technology-focused workforce development strategies in healthcare.

Limitations of this study include reliance on a single retrospective dataset that may not reflect local epidemiology or current screening modalities such as primary HPV testing, as well as the introduction of assumptions inherent in multiple imputation and synthetic sampling techniques, and the absence of evaluations related to cost-effectiveness, clinician acceptability, or patient perceptions; to address these gaps, future work should involve external validation on diverse cohorts, pilot integration into clinical workflows to assess operational feasibility, health economic analyses to quantify potential benefits and unintended consequences, and the development of targeted training modules to embed AI literacy within clinical education, thereby fostering a cautious yet systematic pathway toward the safe and equitable implementation of precision screening strategies in oncology practice.

Conclusion

This comprehensive investigation into cervical cancer risk prediction, employing a synergistic framework that integrates advanced ML algorithms with probabilistic graphical modeling via BBNs, has yielded a series of clinically and methodologically significant findings that may inform future research directions and translational applications in preventive oncology. By combining high-capacity learners such as BART with class imbalance correction techniques like SMOTEEN, the proposed models achieved consistently high predictive performance across multiple cervical cancer screening endpoints, with the overall model demonstrating a particularly strong classification profile (accuracy of 0.956, AUROC of 0.958, and F1-score of 0.945), thereby affirming the utility of ensemble-based approaches in medical risk prediction tasks characterized by complex and imbalanced datasets.

The methodological pipeline, which incorporated multiple imputation strategies to address missingness ranging from 0.8% to 13.6%, and deployed targeted feature selection techniques to isolate clinically relevant predictors, effectively mitigated common limitations inherent to real-world clinical datasets. Notably, the TAN-based BBN constructed using CAIM discretization offered not only high predictive fidelity (91.3% positive prediction rate and 86.8% negative prediction rate), but also a transparent, interpretable structure that revealed meaningful conditional dependencies between variables—particularly the roles of smoking, STD history, and prior cervical pathology—in the progression toward a positive cervical cancer diagnosis.

The integrative framework presented in this study successfully addressed multiple methodological challenges, including high-dimensional feature interactions, data sparsity, and class imbalance, illustrating its scalability and adaptability for other medical classification tasks exhibiting similar structural complexities. Despite the retrospective nature of the data and the acknowledged need for external validation across diverse populations and healthcare settings, the findings provide a robust foundation for developing AI-augmented clinical tools that can support more personalized, efficient, and equitable cervical cancer screening strategies.

Ultimately, the proposed models offer a promising avenue toward enhancing clinical decision-making and risk stratification in cervical cancer prevention by facilitating early detection among high-risk individuals through interpretable, data-driven methods. Future research should prioritize prospective validation studies, rigorous impact assessments within real-world screening workflows, and stakeholder-informed implementation frameworks to evaluate model accuracy, clinical utility, cost-effectiveness, and integration feasibility within existing healthcare infrastructures. Through such continued interdisciplinary effort, the transformative potential of ML and BBNs in precision public health may be more fully realized, thereby contributing to improved patient outcomes and population-level disease control.

Disclosure

The authors report no conflicts of interest in this work.

References

1. World Health Organization. Global strategy to accelerate the elimination of cervical cancer as a public health problem. 2020.
2. Arbyn M, Bansi-Matharu L, Cambiano V. Global burden of cervical cancer in 2021: a comprehensive analysis. *Lancet Glob Health*. 2021;9(5): e620–e630. doi:10.1016/S2214-109X(21)00025-5
3. A. C. Society. Cancer facts & figures 2024. Atlanta: American Cancer Society; 2024.
4. Ma L, Wang Y, Gao X, et al. Economic evaluation of cervical cancer screening strategies in urban China. *Chin J Cancer Res*. 2019;31(6):974–983. doi:10.21147/j.issn.1000-9604.2019.06.13
5. Subramanian S, Trogdon J, Ekwueme DU, Gardner JG, Whitmire JT, Rao C. Cost of cervical cancer treatment: implications for providing coverage to low-income women under the medicaid expansion for cancer care. *Women's Health Issu*. 2010;20(6):400–405. doi:10.1016/j.whi.2010.07.002
6. Tehranian A, Ghahghaei-Nezamabadi A, Motiei Langeroudi M, Aghajani R. Comparison of visual inspection methods using either acetic acid solution or Lugol's iodine solution with colposcopy in screening of cervical cancer: a cross sectional study. *J Obstetrics Gynecol Cancer Res*. 2023;8(1):53–56. doi:10.30699/jogcr.8.1.53
7. Ghosh P, Gandhi G, Kochhar PK, Zutshi V, Batra S. Visual inspection of cervix with Lugol's iodine for early detection of premalignant & malignant lesions of cervix. *Indian J Med Res*. 2012;136(2):265–271.
8. Zhao XL, Zhao S, Xia CF, et al. Cost-effectiveness of the screen-and-treat strategies using hpv test linked to thermal ablation for cervical cancer prevention in China: a modeling study. *BMC Med*. 2023;21(1). doi:10.1186/s12916-023-02840-8
9. Tapera O. Innovations in cervical cancer screening in low-resource settings: a systematic review. *Preventive Med Reports*. 2023;29:101–110.
10. Siyam A. Machine learning approaches for cervical cancer risk stratification: advances and challenges. *Int J Med Inform*. 2022;156:104–113.
11. de Oliveira M. Predictive analytics in oncology: lessons from machine learning applications in cervical cancer. *Artif Intell Med*. 2021;112:102–111.

12. Devi S, Gaikwad SR, R H. Prediction and detection of cervical malignancy using machine learning models. *Asian Pac J Cancer Prev*. 2023;24 (4):1419–1433. doi:10.31557/APJCP.2023.24.4.1419
13. Rahimi M, Akbari A, Asadi F, et al. Cervical cancer survival prediction by machine learning algorithms: a systematic review. *BMC Cancer*. 2023;23(1). doi:10.1186/s12885-023-10805-1
14. Ashtagi R, Rajput V, Antad S, et al. Cervical cancer prediction using machine learning. *J Electr Syst*. 2023. doi:10.52783/jes.851
15. Langberg GSRE, Nygård JF, Gogineni VC, Nygård M, Grasmair M, Naumova V. Towards a data-driven system for personalized cervical cancer risk stratification. *Sci Rep*. 2022;12(1):12083. doi:10.1038/s41598-022-16361-6
16. Fernandes K, Cardoso J, Fernandes J. Cervical cancer (risk factors). 2017. doi:10.24432/C5Z310
17. Nithya B, Ilango V. Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction. *SN Appl Sci*. 2019;1(6):641. doi:10.1007/s42452-019-0645-7
18. Alsmariy R, Healy G, Abdelhafez H. Predicting cervical cancer using machine learning methods. *Comput Struct Biotechnol J*. 2020;13:8–17.
19. Al Mudawi N, Alazeb A. A model for predicting cervical cancer using machine learning algorithms. *Sensors*. 2022;22(11):4132. doi:10.3390/s22114132
20. Ashtagi R, Rajput V, Antad S, et al. Cervical cancer prediction using machine learning. *J Electrical Systems*. 2024;20–1s(1s):944–955. doi:10.52783/jes.851
21. Moscicki A-B, Ma Y, Jonte J, et al. The role of sexual behavior and human papillomavirus persistence in predicting repeated infections with new human papillomavirus types, Cancer Epidemiology. *Biomarkers Prevention*. 2010;19(8):2055–2065. doi:10.1158/1055-9965.EPI-10-0394