International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, Delaware, USA
Volume 13– No. 2, April 2026 – www.ijais.org

An Interpretable Predictive Model for Cervical Cancer Risk Prediction using Hybrid Feature Selection and Ensemble Learning

A.S.M. Sabiqul Hassan
Department of Computer Science and Engineering
Northern University Bangladesh
Dhaka, Bangladesh

Md. Mohsin Uddin Azad
Department of Textile Engineering
Northern University Bangladesh
Dhaka, Bangladesh

Goutam Paul
Department of Computer Science and Engineering
Northern University Bangladesh
Dhaka, Bangladesh

Muhammed Samsuddoha Alam
Department of Computer Science and Engineering
Northern University Bangladesh
Dhaka, Bangladesh

Md. Ruhul Amin
Department of Computer Science and Engineering
Southeast University
Dhaka, Bangladesh

ABSTRACT
Cervical cancer has been known as a continuous health threat among women for a long time. There are some screening techniques available to identify this disease, but those are not compatible now due to their high-cost and time-consuming issues. In this study, a ML based interpretable model has been proposed to analyze the risk factors of cervical cancer for trustworthy decision making. Several ML algorithms: KNN, LR, DT, RF, NB, SVM, XGBoost, and Ensemble Learning (Soft Voting) were applied on a publicly available cervical cancer dataset collected from the UCI dataset repository. The result analysis demonstrated that tree or plane based models: DT, RF, and SVM generated best accuracy but ensemble model maintained a balanced result in all evaluation metrics (accuracy: 0.959, precision: 0.643, recall: 0.818, f1-score: 0.720, and roc-auc: 0.882). Additionally, a web version of this model was deployed with explainability based on the top features. This type of decision making tool can be used in the healthcare sector in future with further improvements.

Keywords
Cervical Cancer Prediction, Risk Factors Analysis, Class Imbalance, Explainable AI (XAI), Decision Making

1. INTRODUCTION
Cervical cancer is considered globally as a prevalent and preventable disease affecting women's health. Although there are some advances in screening methodologies, for example, Pap smears and Human Papillomavirus (HPV) testing, late diagnosis still leads a large number of people to death, particularly in low-resource settings [1]. If cervical cancer risk assessment is completed earlier with accuracy, then it can facilitate timely clinical intervention, improve patient prognosis and reduce healthcare costs. Traditional screening methods are clinically valuable but they have limitations, for example, dependency on specialized equipments and trained personnel, inter-observer variability, and barriers to access in underserved populations [2]. These challenges have motivated people to explore the computational approaches for assistance in early risk identification based on routinely collected clinical and behavioral data.

As ML has the ability to model complex nonlinear relationships and interactions among diverse clinical features, a considerable promise has been noticed in disease risk prediction [3]. To classify cervical cancer risk based on demographic, clinical, and behavioral attributes, supervised learning techniques have been applied in several studies recently [4]. However, there are some limitations in the previous works, for example, reliance on a single model without comparative evaluation across multiple algorithms, inadequate handling of class imbalance common in medical datasets, and limited explainability of model predictions that hinders clinical trust and adoption.

This study proposes a comprehensive ML framework for cervical cancer risk prediction motivated by these gaps. The dataset used in this project is publicly available in the UCI dataset repository [5]. The proposed framework included robust data processing, effective handling of class imbalance through SMOTE and training of multiple base models: KNN, LR, DT, RF, NB, SVM, and XGBoost. Each model was optimized via possible hyperparameter tuning. Then, an ensemble model: soft voting classifier was applied to decide the final result, calculating the probabilities among the base models. Rigorous cross-validation was applied to ensure stability of the final predictive model. Performance of this model was analyzed using the standard evaluation metrics for classification tasks.

Additionally, SHAP was applied for model interpretability that provided both global and local explanations of feature contributions [6]. To create a simplified deployment model suitable for real-time risk prediction, the most influential

features identified by SHAP were further used. Then, the utility of the trained model for practical clinical screening support was demonstrated by implementing a flask based web application.

The main contributions of this paper are as follows. For cervical cancer risk classification, a comparative evaluation among multiple base classifiers was accomplished with several performance metrics. SMOTE was integrated to address class imbalance issues and hyperparameters were tuned to optimize the performance of trained models. SHAP was applied for explainability to enhance feature interpretability and facilitate trust in clinical prediction models. A deployment ready model using top influential features was developed for interactive risk assessment of cervical cancer.

The rest of this paper is structured as follows. Section 2 reviews the previous works related to cervical cancer risk prediction. Section 3 describes the dataset and methods employed in this study. Section 4 presents the experimental results of this study with explanation. Finally, Section 5 concludes the study with future guidance for extension.

2. RELATED WORKS

Recently, a trend has been noticed on early detection and clinical decision support for disease risk prediction including cervical cancer using ML techniques. Traditional classification models have been explored in several studies on clinical and demographic datasets. Single supervised models trained on questionnaires based on clinical variables were employed for early research on cervical cancer risk prediction. For example, SVM and LR algorithms were used to classify risk levels based on demographic and behavioral factors. As the sample size was small and there was a lack of robust preprocessing methods, this approach showed promising performance with limited generalization [7].

Similarly, several studies have achieved acceptable classification metrics using DT and NB algorithms. But they often suffered from class imbalance and overfitting due to skewed distributions of positive cases [8]. In some studies, model comparison included tree based ensemble approaches for example, RF and GBM algorithms. As these methods reduce variance and capture feature interactions more effectively, they tend to outperform single models. For example, RF and XGBoost algorithms were applied to clinical data and reported improved roc-auc scores compared to baseline models in a recent study. But there was not thorough addressing of interpretability or deployment considerations [9].

Although there was performance gain, many prior studies emphasized predictive accuracy overlooking model explainability, an essential factor for medical adoption. In other medical prediction tasks, explainable AI approaches for example, LIME and SHAP have been proposed to provide insight into how features influence model decisions [10]. However, their application to cervical cancer risk prediction remained relatively limited. A recent study used SHAP to interpret XGBoost predictions on a small cervical dataset and it highlighted key risk factors. But this study did not investigate feature selection, ensemble learning or web based deployment [11].

There was another gap in existing studies is the practical integration of prediction models into clinical workflows. Most studies didn't develop real-time tools for interactive testing and stopped at offline model evaluation. Web or mobile interfaces for risk assessment have been considered in limited studies. But they used a fixed subset of predictors without rigorous feature stability analysis or deployment optimization [12]. In summary, previous studies demonstrated the feasibility of ML techniques for cervical cancer risk classification but several limitations remain in those studies. Firstly, many studies were dependent on a single model without comprehensive comparison across the base classifiers. Secondly, class imbalance and hyperparameter tuning were often insufficiently addressed. Thirdly, there was a lack of interpretability and deployment ready systems that can be interactively used for clinicians or patients.

The proposed framework in this study addressed these gaps by using SMOTE to handle imbalance, integrating multiple tuned models, applying explainability via SHAP, and building a deployment ready web application based on the top influential features. It can be used later for decision making purposes in the healthcare sector [13-14].

3. MATERIALS AND METHODS

The development process of the predictive ML model for cervical cancer risk is described below in Figure 1.

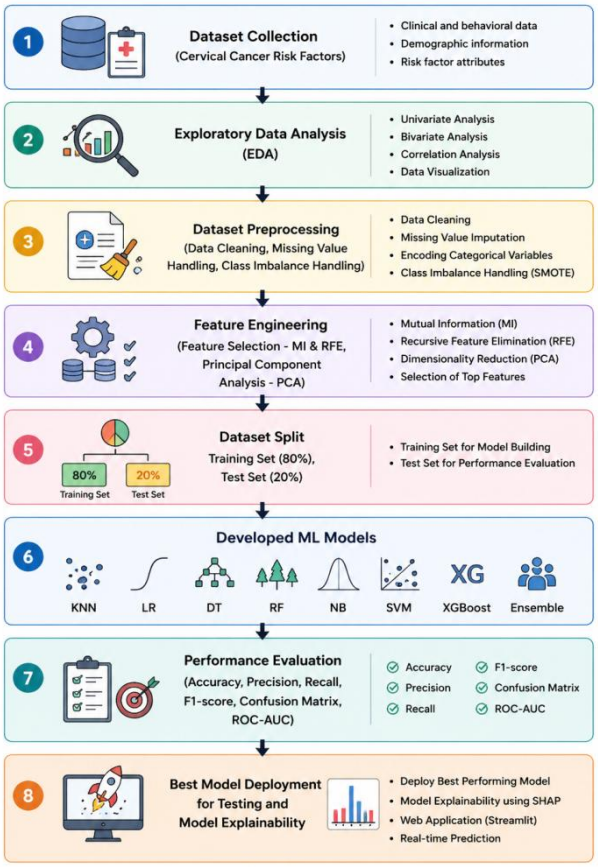

Figure 1. Workflow of Risk Prediction Model for Cervical Cancer

3.1 Dataset Description

This study used a publicly available cervical cancer risk prediction dataset from the UCI dataset repository [5]. The dataset was derived from a hospital based survey associated with cervical cancer. It contains demographic, behavioral and medical risk factors where 858 instances and 36 features are available. But the biopsy is marked as a label representing

binary values (true-1 and false-0) for cervical cancer risk prediction.

This dataset contains a significant number of missing values (?), as there were incomplete medical records and patient non-responses. Additionally, there was imbalance in class distribution which could make the model biased. These issues reflected real-world clinical conditions which should be processed carefully to develop a trustworthy disease diagnosis tool for decision making purposes in the healthcare sector.

Table 1. Description of the Dataset

Feature	Data Example	Data type
Age	13, 25, 84	N
Number of sexual partners	2, 3, 5	N
First sexual intercourse	15, 17, 21	N
Num of pregnancies	1, 2, 4	N
Smokes	0, 1	B
Smokes (years)	0, 1.26, 37	N
Smokes (packs/year)	0, 0.51, 37	N
Hormonal Contraceptives	0, 1	B
Hormonal Contraceptives (years)	0, 3, 15	N
IUD	0, 1	B
IUD (years)	0, 6, 7	N
STDs	0, 1	B
STDs (number)	0, 1, 2	N
Condylomatosis (STD)	0, 1	B
cervical condylomatosis (STD)	0, 1	B
vaginal condylomatosis (STD)	0, 1	B
vulvo-perineal condylomatosis (STD)	0, 1	B
syphilis (STD)	0,1	B
pelvic inflammatory disease (STD)	0, 1	B
genital herpes (STD)	0, 1	B
molluscum contagiosum	0, 1	B
AIDS (STD)	0, 1	B
HIV (STD)	0, 1	B
Hepatitis B (STD)	0, 1	B
HPV (STD)	0, 1	B
Number of diagnosis (STD)	0, 1, 3	N
Time since first diagnosis (STD)	1, 15, 19	N
Time since last diagnosis (STD)	1, 15, 19	N
Cancer (Dx)	0,1	B
CIN (Dx)	0,1	B
HPV (Dx)	0,1	B
Dx	0,1	B
Hinselmann	0,1	B
Schiller	0,1	B
Citology	0,1	B
Biopsy (label)	0,1	B

Legend: Dx = Diagnosis History, STD = Sexually Transmitted Disease, N = Numerical, and B = Boolean

3.2 Exploratory Data Analysis (EDA)

To understand the structure and quality of the dataset, EDA was conducted in this study. It included statistical summaries, visualization of feature distributions using histograms and boxplots, analysis of missing values, and correlation analysis using Pearson correlation coefficients. A correlation heatmap was generated to identify highly correlated features. It provided initial insights into potential redundancy among variables. Additionally, class distribution analysis revealed a strong imbalance between the target classes. It motivated the use of SMOTE, a class imbalance handling technique in subsequent steps of this study [15].

3.3 Data Preprocessing

As median imputation is robust to outliers and commonly used in medical dataset, it was applied to numerical features for addressing missing values in this study [16]. Then, to ensure the consistent feature scaling across different ML models all features were standardized using z-score normalization via the StandardScaler. The dataset was split into training and test subsets using an 80:20 ratio. It was ensured that the test set remained completely unseen during model training and tuning.

3.4 Handling Class Imbalance

As there was skewed class distribution in the dataset, SMOTE technique was applied to the training dataset only [17]. Actually, SMOTE generated synthetic samples of the minority class by interpolating between existing minority samples and improved classifier learning without duplicating data points. This approach helped mitigate bias towards the majority class and improved recall for minority class predictions.

3.5 Feature Selection

Two complementary feature selection techniques: Mutual Information (MI) and Recursive Feature Elimination (RFE) were employed to reduce dimensionality and enhance model generalization. MI technique measured the dependency between input features and the target variable. RFE technique iteratively removed less important features using a tree based estimator. The intersection of features selected by both methods was retained to ensure robustness and stability of selected predictors. This approach helped reduce noise, computational complexity and overfitting risk [18].

3.6 Dimensionality Reduction

Principal Component Analysis (PCA) was applied in this study to the selected features to examine variance distribution and feature redundancy for experiment analysis [19]. As transformed components are not interpretable for real-time user input in clinical applications, PCA was solely used for comparative research purposes and not included in the deployment pipeline.

3.7 Machine Learning Algorithms

In this study, several supervised ML classifiers were trained and evaluated using KNN, LR, DT, RF, NB, SVM and XGBoost algorithms. RandomizedSearchCV was applied to perform hyperparameter optimization in possible cases. It efficiently explored the hyperparameter space while reducing computational cost [20]. Each model was trained on the balanced training dataset and optimized independently. Soft voting classifier, an ensemble learning technique was constructed by combining the best performing tuned models to improve predictive robustness. It aggregated the predicted class probabilities, enabling the ensemble to capture complementary strengths of individual learners and reduce variance [21].

3.8 Model Evaluation and Deployment

To provide a comprehensive assessment, model performance was evaluated using several evaluation metrics: confusion matrix, accuracy, precision, recall, f1-score, and roc-auc in this study [22-23]. Confusion matrices and roc curves were generated for visual evaluation of the models. To evaluate model stability and generalization across different data partitions, k-fold cross-validation was employed in this study

[24]. A confusion matrix includes several parameters: TP (True Positive), FP (False Positive), TN (True Negative), and FN (False Negative), which summarize the performance of each trained model.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1-score = 2 * \frac{Precision * Recall}{Precision+Recall} \quad (4)$$

Accuracy mentioned in the Eq. (1) measures the percentage of correctly predicted instances for both true positives and true negatives among all predicted instances. It performs well when the classes are evenly distributed in the dataset. Precision mentioned in the Eq. (2) measures the percentage of true positives among all predicted positives. It reflects the model's capability in avoiding false positives. Recall mentioned in the Eq. (3) measures the percentage of actual positives among the correctly predicted by a model. It is important particularly in situations like disease detection where the case of a missing value can be costly.

F1-score mentioned in the Eq. (4) is calculated as the harmonic mean of precision and recall values. It can balance the trade-off between these two metrics and performs better when the class distribution in the dataset is imbalanced. ROC-AUC measures the capability of a classification model to distinguish between classes. The ROC curve plots the true positive rate against the false positive rate. The values of AUC range from 0.5 (random) to 1 (perfect classifier).

SHAP was used to analyze feature contributions to enhance transparency and interpretability [25]. It provided both global feature importance and local explanations for individual predictions and assists clinicians to understand how each feature influences model decisions. Finally, to create a simplified deployment ready predictive model, the most influential features identified through SHAP analysis were further selected. For real-time testing and usability, a lightweight deployment model was developed based on that. This application demonstrated the practical applicability of the proposed framework for immediate risk predictions of cervical cancer.

4. RESULT ANALYSIS

4.1 EDA Results

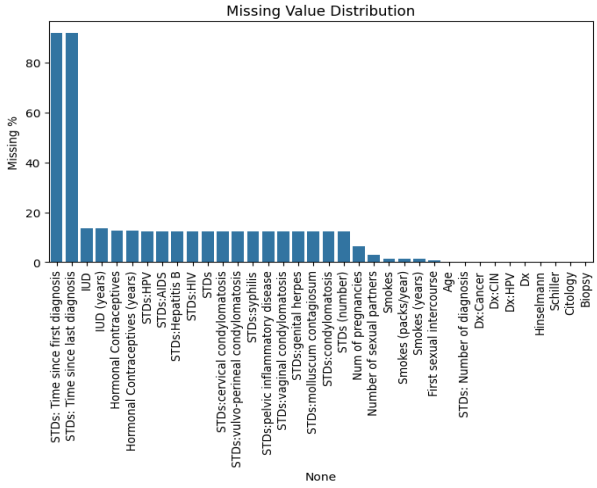

Figure 2: Missing Value Distribution across the Features

EDA applied in the dataset revealed substantial variability across the clinical and behavioral risk factors. Several attributes displayed skewed distributions and the presence of outliers. It justified the use of median based imputation and robust scaling techniques. Figure 2 represents the missing value distribution across the features in the utilized dataset of this study. Correlation analysis showed moderate correlations among a subset of features in Figure 3. However, no excessive multicollinearity was found that would require feature removal depending on correlation thresholds.

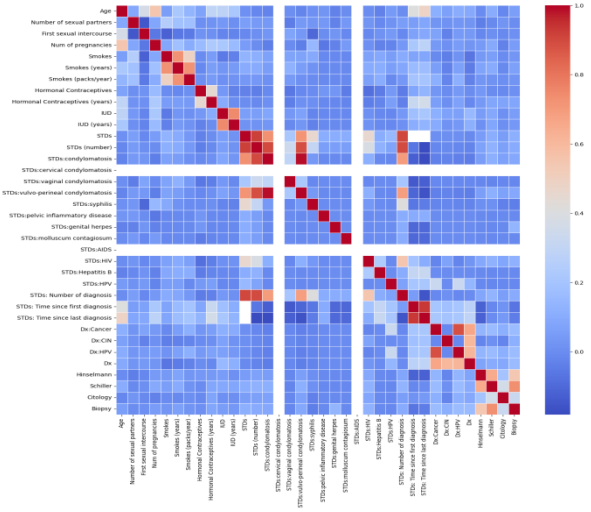

Figure 3: Correlation Analysis across the Features

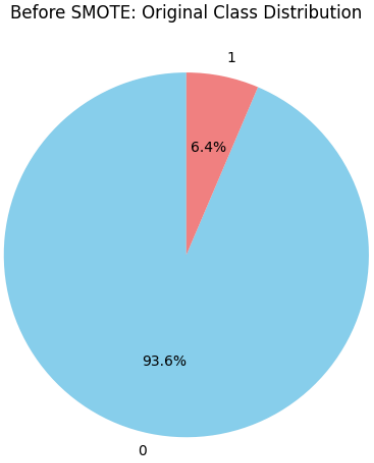

Figure 4: Original Class Distribution Before SMOTE

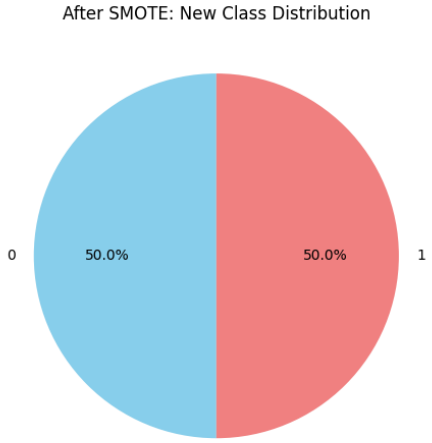

Figure 5: New Class Distribution After SMOTE

The class distribution analysis step confirmed a significant imbalance between cervical cancer positive and negative cases and it reflected real-world screening data. This type of class imbalanced scenario requires the application of SMOTE technique prior to model training to prevent bias towards the majority class. Figure 4 and 5 represent the effect of SMOTE technique application on the dataset used in this study.

4.2 Feature Selection and Dimensionality Reduction Analysis

A stable subset of informative features strongly associated with cervical cancer risk was found as result by the combination of MI and RFE techniques. These features primarily represented reproductive history, sexually transmitted disease indicators and screening-related variables and they aligned with established clinical findings reported in the related studies [26].

PCA was applied for experimental analysis and it demonstrated a reduced number of components could explain the majority of variance in the dataset. But PCA transformed features were excluded from the deployment pipeline to preserve interpretability. This approach also facilitated real-time user input, consistent with recommendations for clinical decision support systems [27]. Figure 6 represents the effect of PCA for dimensionality reduction in the dataset.

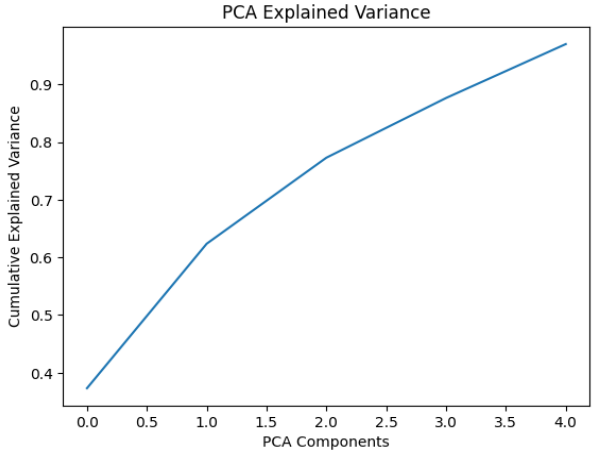

Figure 6: PCA for Dimensionality Reduction

4.3 Model Performance Evaluation

Table 2 represents the results of hyperparameter tuning on the trained ML models except NB and Ensemble models.

Table 2. Results of Hyperparameter Tuning across Models

ML Model	Best Hyperparameter after Tuning
KNN	'weights': 'distance', 'n_neighbors': 7
LR	'penalty': 'l2', 'C': 0.1
DT	'min_samples_split': 10, 'min_samples_leaf': 4, 'max_depth': None
RF	'n_estimators': 100, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_depth': None
NB	N/A
SVM	'kernel': 'rbf', 'gamma': 'auto', 'C': 1
XGBoost	'n_estimators': 300, 'max_depth': 3, 'learning_rate': 0.05
Ensemble	N/A

The NB model did not require hyperparameter tuning due to its nature. Ensemble model also did not need to tune hyperparameters as it aggregates the results of the selected classifiers. Table 3 represents the cross-validation (CV) results of trained ML models using 5-folds and roc-auc values.

Table 3. Statistics of Confusion Matrix across Models

ML Model	CV Result
KNN	0.986 ± 0.005
LR	0.976 ± 0.008
DT	0.989 ± 0.003
RF	0.996 ± 0.003
NB	0.953 ± 0.012
SVM	0.988 ± 0.008
XGBoost	0.996 ± 0.003
Ensemble	0.995 ± 0.003

Table 4 and 5 summarize the predictive performance of all trained ML models: KNN, LR, DT, RF, NB, SVM, XGBoost and Ensemble after evaluation. Performance of these models was analyzed using different evaluation metrics: confusion matrix, accuracy, precision, recall, f1-score and roc-auc.

Table 4. Statistics of Confusion Matrix across Models

ML Model	TN	FP	FN	TP
KNN	156	5	2	9
LR	156	5	2	9
DT	159	2	2	9
RF	159	2	3	8
NB	155	6	2	9
SVM	157	4	2	9
XGBoost	160	1	3	8
Ensemble	156	5	2	9

Table 5. Statistics of Evaluation Metrics across Models

ML Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
KNN	0.959	0.643	0.818	0.720	0.897
LR	0.959	0.643	0.818	0.720	0.888
DT	0.977	0.818	0.818	0.818	0.898
RF	0.971	0.800	0.727	0.762	0.883
NB	0.953	0.600	0.818	0.692	0.896
SVM	0.965	0.692	0.818	0.750	0.842
XGBoost	0.977	0.889	0.727	0.800	0.941
Ensemble	0.959	0.643	0.818	0.720	0.882

Figure 7 to Figure 14 represent the confusion matrix of all trained models: KNN, LR, DT, RF, NB, SVM, XGBoost and Ensemble. Each confusion matrix contains four parameters: TN, FP, FN and TP.

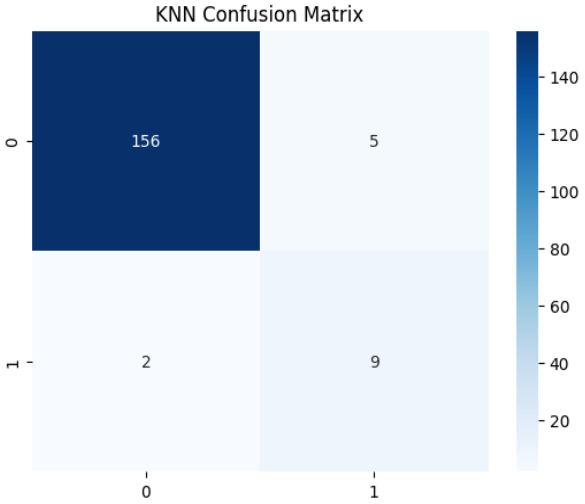

Figure 7: Confusion Matrix of KNN Model

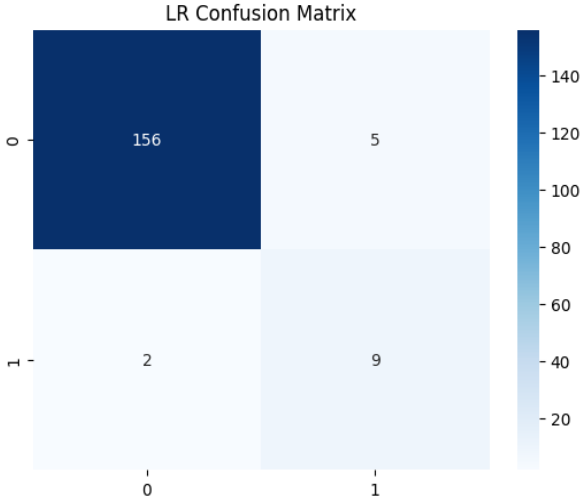

Figure 8: Confusion Matrix of LR Model

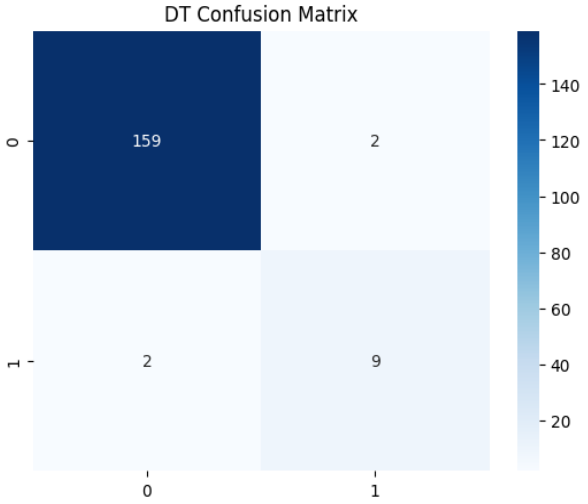

Figure 9: Confusion Matrix of DT Model

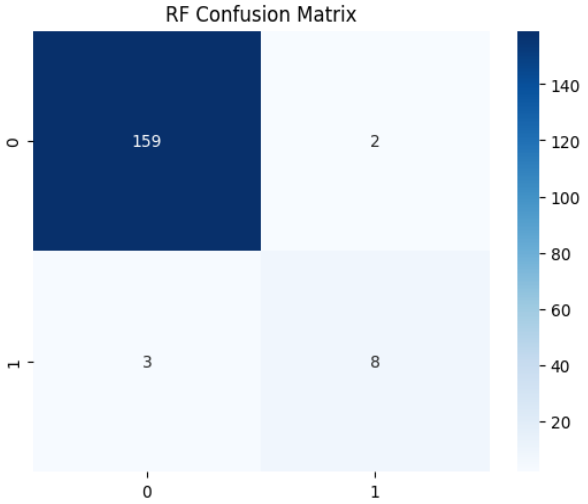

Figure 10: Confusion Matrix of RF Model

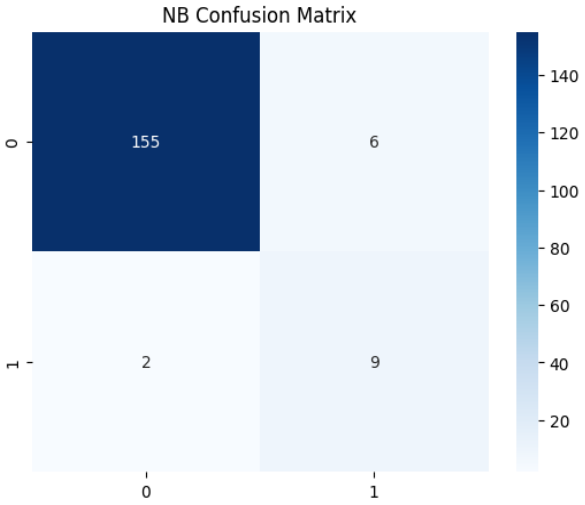

Figure 11: Confusion Matrix of NB Model

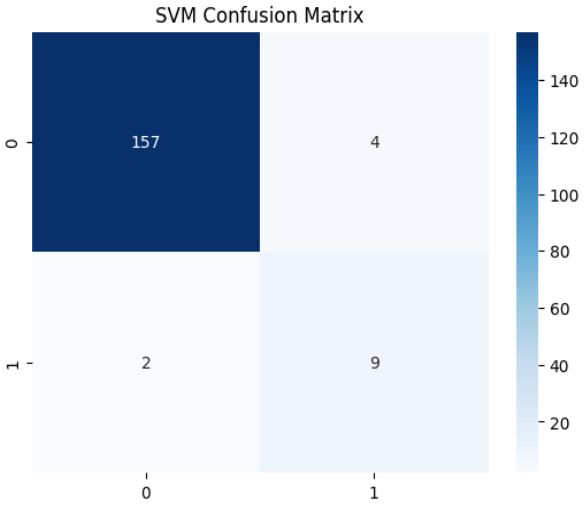

Figure 12: Confusion Matrix of SVM Model

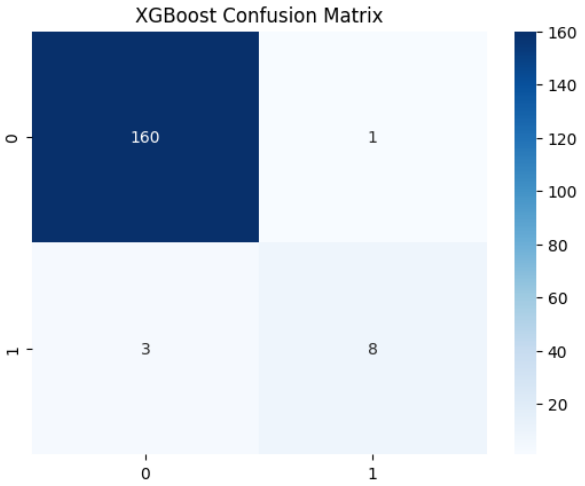

Figure 13: Confusion Matrix of XGBoost Model

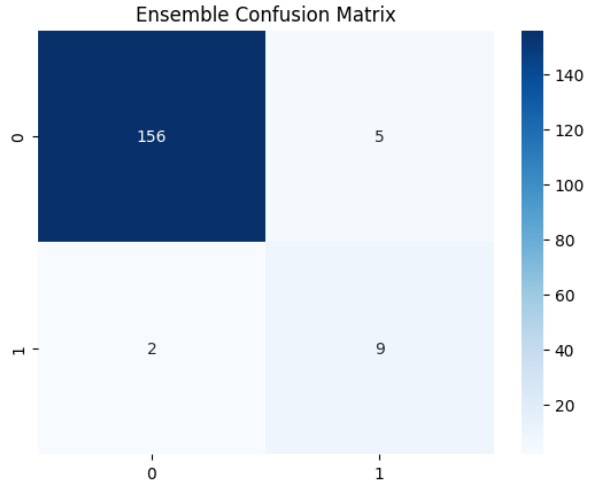

Figure 14: Confusion Matrix of Ensemble Model

Firstly, the accuracy values for KNN, LR, DT, RF, NB, SVM, XGBoost and Ensemble classifiers were 0.959, 0.959, 0.977, 0.971, 0.953, 0.965, 0.977 and 0.959 respectively. XGBoost, DT, RF classifiers achieved the highest accuracy value and other classifiers generated almost close accuracy values. Figure 15 represents accuracy comparison across ML models used in this study.

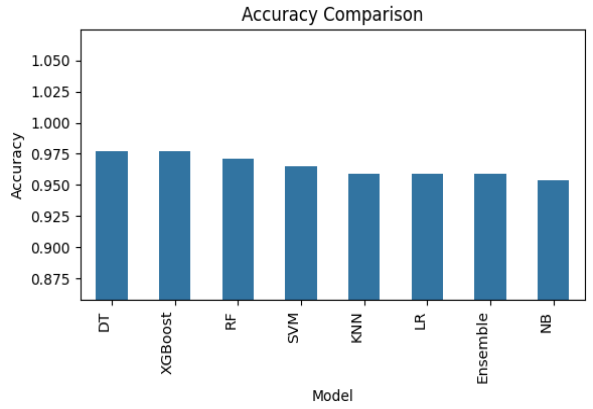

Figure 15: Accuracy Comparison across ML Models

Secondly, the precision values for KNN, LR, DT, RF, NB, SVM, XGBoost and Ensemble classifiers were 0.643, 0.643, 0.818, 0.800, 0.600, 0.692, 0.889 and 0.643 respectively. XGBoost classifiers achieved the highest precision value. DT and RF classifiers achieved the second highest precision value. Other classifiers generated low precision values compared to the mentioned three classifiers. Figure 16 represents precision comparison across ML models used in this study.

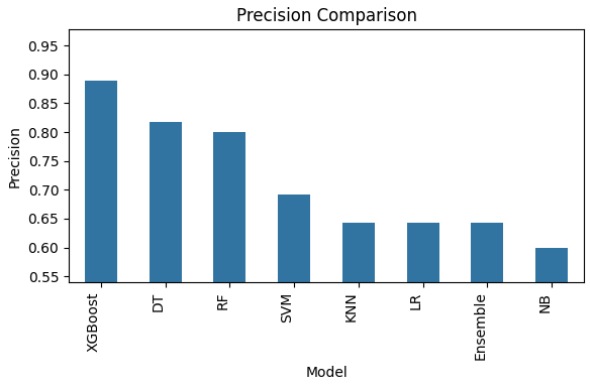

Figure 16: Precision Comparison across ML Models

Thirdly, the recall values for KNN, LR, DT, RF, NB, SVM, XGBoost and Ensemble classifiers were 0.818, 0.818, 0.818, 0.727, 0.818, 0.818, 0.727, and 0.818 respectively. Almost all classifiers achieved the highest recall value 0.818 but RF and XGBoost classifiers achieved a lower recall value 0.727. Figure 17 represents recall comparison across ML models used in this study.

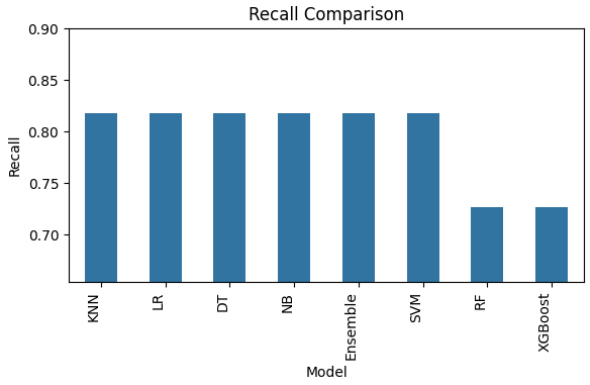

Figure 17: Recall Comparison across ML Models

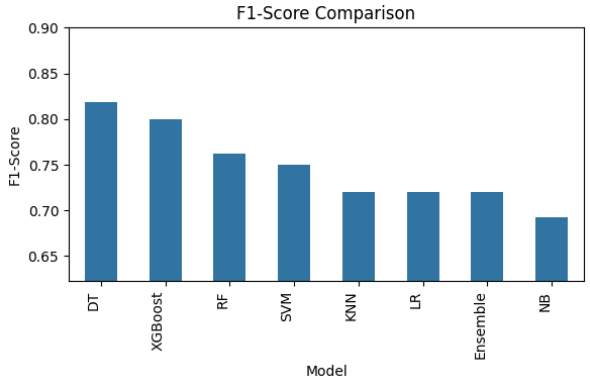

Figure 18: F1-Score Comparison across ML Models

Additionally, the f1-score values for KNN, LR, DT, RF, NB, SVM, XGBoost and Ensemble classifiers were 0.720, 0.720, 0.818, 0.762, 0.692, 0.750, 0.800 and 0.720 respectively. DT and XGBoost classifiers achieved the highest f1-score value. RF and SVM classifiers achieved the second highest f1-score value. Other classifiers achieved low f1-scores compared to the mentioned four classifiers. Figure 18 represents f1-score comparison across ML models used in this study

The final observation of this study is that tree based models, particularly RF and XGBoost continuously achieved better performance over linear and distance based classifiers across most metrics. The soft voting classifier, an ensemble learning technique further improved robustness by aggregating probabilistic outputs from several tuned ML models. It achieved the most balanced performance across all evaluation metrics.

The stability of the ensemble model was confirmed via cross-validation results. There was a trend of low variance across folds and it suggested good generalization capability. A notable reduction in false negatives compared to baseline models was noticed in confusion matrix analysis. It is considered in critical medical risk screening scenarios.

4.4 ROC Analysis

The superiority of ensemble and tree based models were validated via ROC curve analysis. The ensemble classifier model achieved the highest roc-auc score and it indicated strong discriminative capability between risk classes. Simpler models like NB and KNN demonstrated comparatively lower performance in ROC curves, as they had limited capacity to capture complex nonlinear feature interactions.

These findings are consistent with previous related studies and they highlighted the effectiveness of ensemble learning in medical classification tasks involving heterogeneous feature sets. Moreover, the roc-auc values for KNN, LR, DT, RF, NB, SVM, XGBoost and Ensemble classifiers were 0.897, 0.888, 0.898, 0.883, 0.896, 0.842, 0.941 and 0.882 respectively. LR, RF, SVM and Ensemble classifiers achieved the highest roc-auc value around 0.9 and other classifiers achieved comparatively lower roc-auc value around 0.88. Figure 19 to Figure 26 represent ROC curves for KNN, LR, DT, RF, NB, SVM, XGBoost and Ensemble models trained in this study.

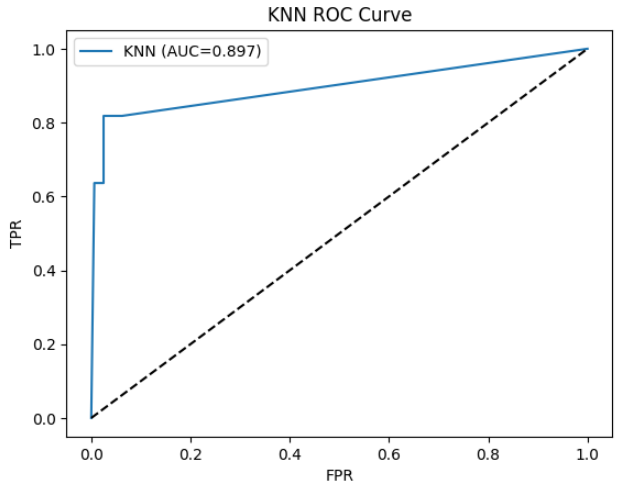

Figure 19: ROC Curve for KNN Model

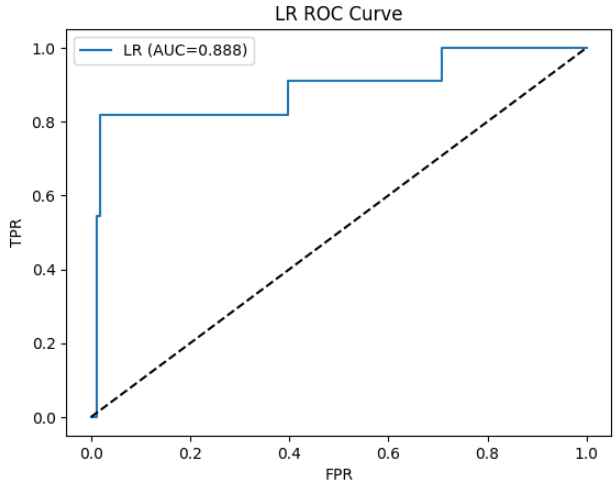

Figure 20: ROC Curve for LR Model

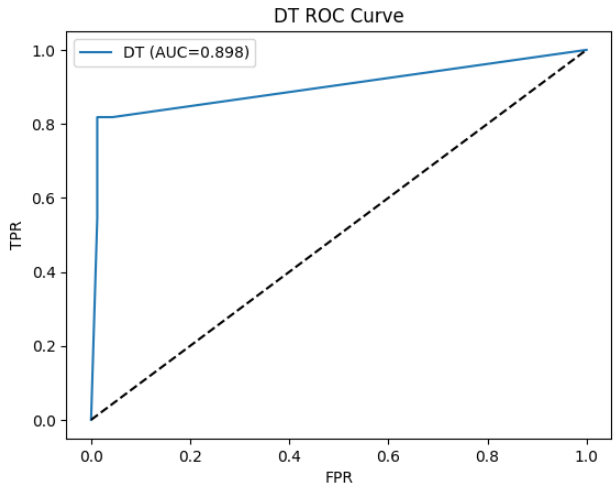

Figure 21: ROC Curve for DT Model

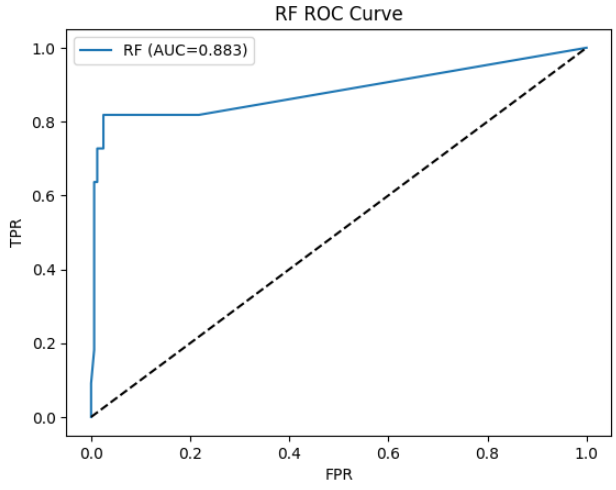

Figure 22: ROC Curve for RF Model

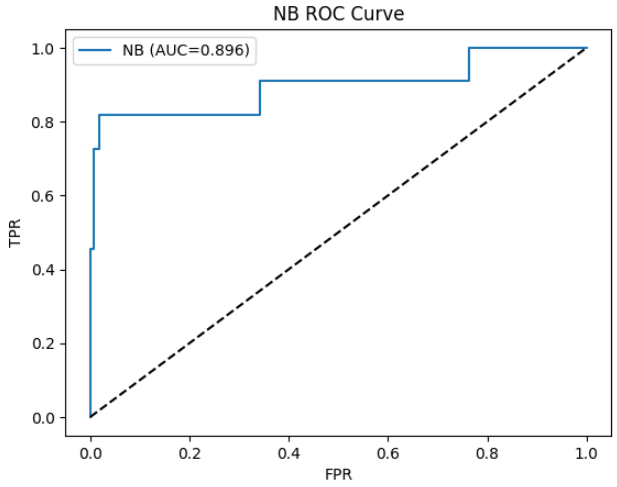

Figure 23: ROC Curve for NB Model

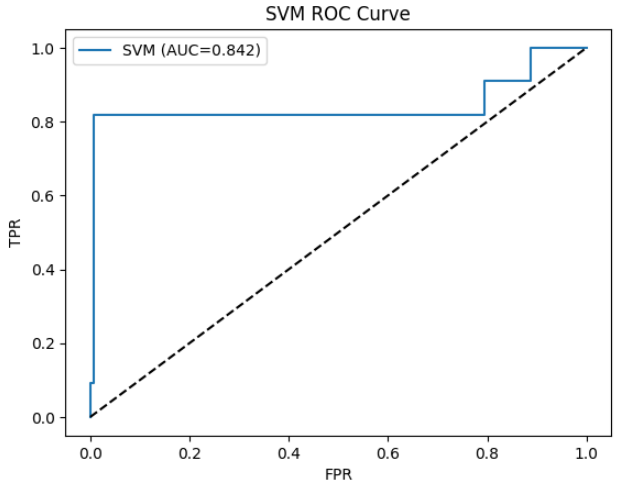

Figure 24: ROC Curve for SVM Model

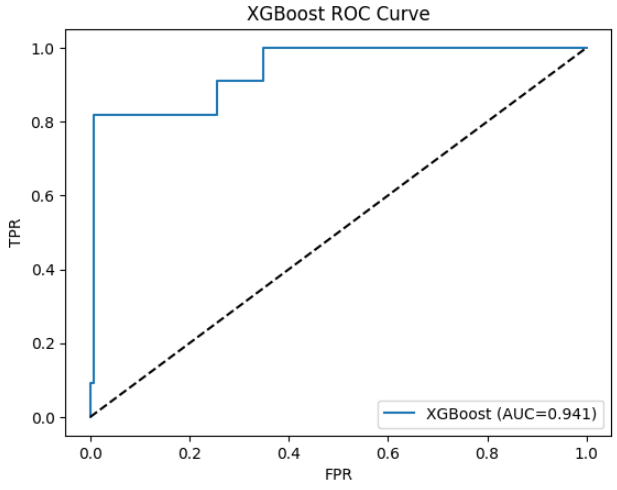

Figure 25: ROC Curve for XGBoost Model

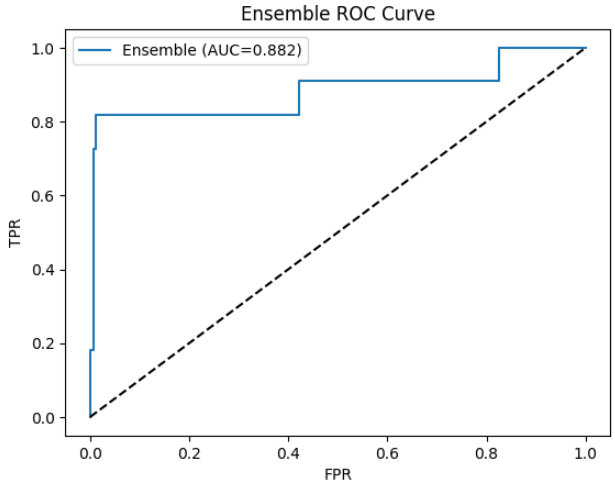

Figure 26: ROC Curve for Ensemble Model

4.5 Model Explainability using SHAP

SHAP was employed to address the interpretability challenge inherent in complex ML models. A small subset of features as dominant contributors to model predictions was identified via SHAP global importance analysis. These features were consistent with known cervical cancer risk factors and reinforced the clinical plausibility of the model. Local SHAP explanations also illustrated the procedure of individual feature values influencing specific predictions. It also provided transparent justification for both high-risk and low-risk classifications. The integration of SHAP enhances trustworthiness and supports potential adoption in clinical decision-support settings. It has been emphasized in recent explainable AI research [28].

Table 6. Statistics of SHAP Importance for Top Features

Feature	SHAP Importance
Schiller	0.366585
Num of pregnancies	0.060345
Dx	0.039135
First sexual intercourse	0.033659
Age	0.024579
Hormonal Contraceptives	0.020251

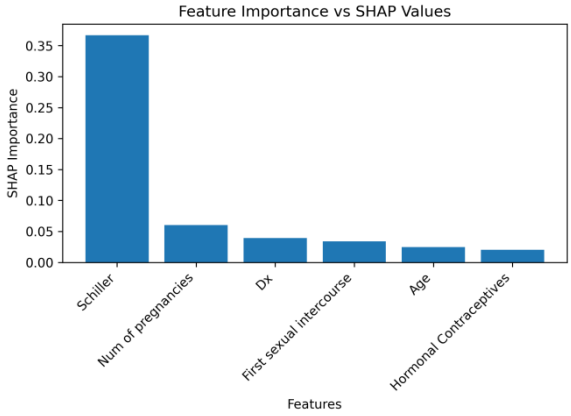

Figure 27: Feature Importance vs SHAP Values Plot

Table 6 represents the top features of the dataset identified using SHAP importance values. Moreover, figure 27 represents a feature importance vs SHAP values plot based on the data available in Table 6. These features were consistently selected across methods and they indicated their strong predictive importance in this study.

4.6 Deployment Oriented Evaluation

Top influential features were selected to train a simplified RF model for deployment based on SHAP importance. The deployment model maintained competitive performance but the feature set was reduced than the original one. It demonstrated that accurate risk prediction can be possible with limited input variables. Usability and data collection was significantly improved via this reduction in real-world applications.

The deployment ready model was integrated into a flask based web application. It enabled real-time risk prediction for cervical cancer and visual explanation through SHAP plots. This practical implementation bridged the gap between offline model development and real-world usability to overcome a limitation frequently noticed in previous cervical cancer prediction studies.

Figure 28 and Figure 29 represent the Testing Interface and Result Interpretation Plot respectively for a sample positive case. The model predicts high risk for cervical cancer. It is primarily influenced by a positive Schiller test, along with factors such as age and early sexual activity. These factors increase the predicted cancer risk. These combined effects lead the model toward a high-risk classification.

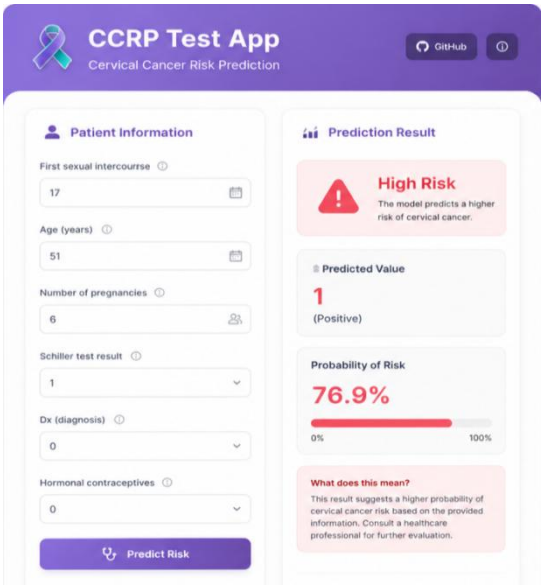

Figure 28: Testing Interface for Positive Case

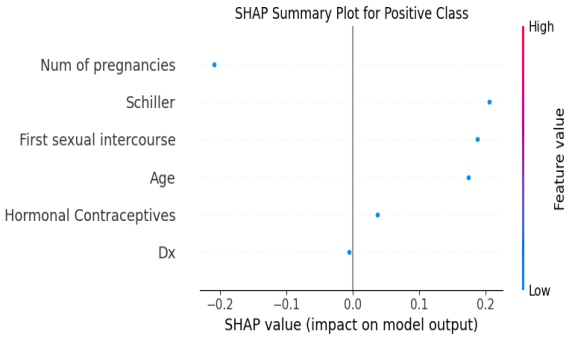

Figure 29: Result Interpretation Plot for Positive Case

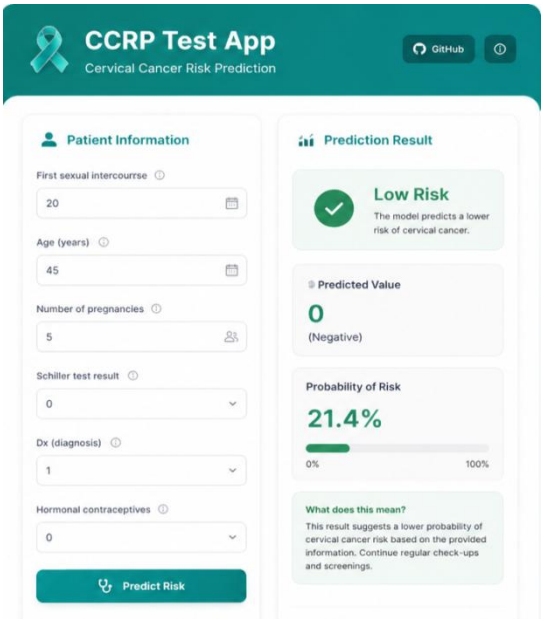

Figure 30: Testing Interface for Negative Case

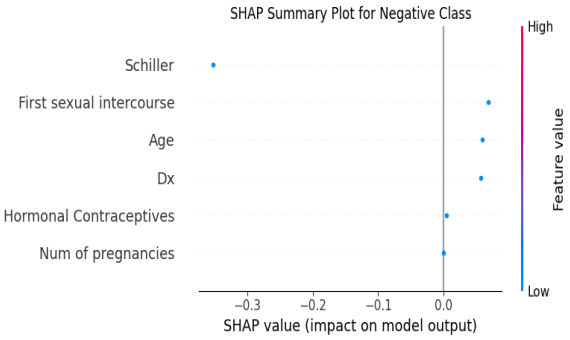

Figure 31: Result Interpretation Plot for Negative Case

Similarly, Figure 30 and Figure 31 represent the Testing Interface and Result Interpretation Plot respectively for a sample negative case. The model predicts low risk for cervical cancer. It is mainly driven by a negative Schiller test. It strongly supports a non-cancer outcome. Other features like age, sexual history, and diagnosis history have only minor influence. This approach results in an overall low-risk prediction.

4.7 Discussion

The experimental results showed that the ensemble based ML model performed better compared to other base classifiers. It combined robust preprocessing and class imbalance handling which could effectively predict cervical cancer risk from clinical and behavioral data. The incorporation of SHAP explainability enhanced transparency as well as facilitated informed clinical interpretation of model outputs. While previous studies relied on single classifiers or lacked interpretability, the proposed framework offers a comprehensive, explainable and deployment ready solution. However, this study has several limitations. The dataset used in this study is relatively modest and derived from a single source that may affect generalizability. For future work extension, external validation using multi-center datasets and prospective clinical data should be focused.

5. CONCLUSION AND FUTURE WORK

This study represented a comprehensive and explainable ML framework for cervical cancer risk prediction using a publicly available dataset from the UCI dataset repository. Several supervised ML algorithms: KNN, LR, DT, RF, NB, SVM and XGBoost were explored to figure out the most effective predictive model for the biopsy test result. Stratified k-fold cross-validation and hyperparameter tuning were incorporated into the model training pipeline to ensure robustness and generalization capability. Model performance was enhanced and redundancy was reduced via required feature preprocessing steps including missing value handling, scaling and correlation-based feature selection. The experimental analysis demonstrated that the ensemble model maintained a balanced performance in terms of all standard evaluation metrics compared to the trained base classifiers.

A key contribution of this study was the integration of SHAP for transparent and interpretable prediction. It improved clinical trust and supported informed decision making that is considered critical for healthcare applications. Furthermore, the ensemble model was designed to be deployable through a lightweight web based interface based on the top features. Overall, the proposed explainable ML model could effectively support early cervical cancer risk assessment.

The proposed model demonstrated promising performance but several extensions can be considered to enhance its clinical applicability and research contribution. For example, incorporation of real-world clinical data, handling class imbalance more effectively, integrating temporal and longitude analysis, combination of deep learning and hybrid models, exploring federated and privacy-preserving learning, integration of clinical decision support and involvement of prospective clinical validation are some possible future directions. By addressing these directions, the proposed model can be evolved into a more trustworthy decision making tool for cervical cancer risk prediction and early intervention.

6. ACKNOWLEDGMENTS

The authors are thankful to all individuals for their assistance during the research period and acknowledge the importance of all materials used in this research.

7. REFERENCES

[1] World Health Organization (WHO), “Cervical Cancer,” [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cervical-cancer [Accessed: Dec 02, 2025].

[2] H.N. Tahir *et al.*, “Artificial intelligence versus manual screening for the detection of diabetic retinopathy: a comparative systematic review and meta-analysis,” *Frontiers in Medicine*, vol. 12, May 2025, doi: 10.3389/fmed.2025.1519768.

[3] N.H. Alhumaidi *et al.*, “The Use of Machine Learning for Analyzing Real-World Data in Disease Prediction and Management: Systematic Review,” *JMIR medical informatics*, vol. 13, Jun 2025, Art. no. e68898, doi:10.2196/68898.

[4] N.A. Mudawi and A. Alazeb, “A Model for Predicting Cervical Cancer Using Machine Learning Algorithms,” *Sensors (Basel)*, vol. 22, no. 11, May 2022, Art. no. 4132, doi: 10.3390/s22114132.

[5] *UCI Machine Learning Repository*, "Cervical Cancer Risk Classification," [Online]. Available: https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors. [Accessed: Jan 27, 2026].

[6] A.S. Antonini *et al.*, "Machine Learning model interpretability using SHAP values: Application to Igneous Rock Classification task," *Applied Computing and Geosciences*, vol. 23, Sep 2024, Art. no. 100178, doi: 10.1016/j.acags.2024.100178.

[7] S. Devi, "Prediction and Detection of Cervical Malignancy Using Machine Learning Models," *Asian Pacific journal of cancer prevention: APJCP*, vol. 24, no. 4, pp: 1419–1433. Apr 2023, doi: 10.31557/APJCP.2023.24.4.1419.

[8] R. Abdulkareem and A. M. Abdulazeez, "A Comparative Study of Multi-Class Classification Based on Imbalanced Data: A Review," *The Indonesian Journal of Computer Science*, vol. 14, no. 5, Oct 2025, doi: 10.33022/ijcs.v14i5.5020.

[9] H. Park *et al.*, "Integrating Large Language Models with Deep Learning for Breast Cancer Treatment Decision Support," *Diagnostics*, vol. 16, no. 3, Jan 2026, Art. no. 394, doi:10.3390/diagnostics16030394.

[10] S.U. Hassan *et al.*, "Local interpretable model-agnostic explanation approach for medical imaging analysis: A systematic literature review," *Computers in Biology and Medicine*, vol. 185, Feb 2025, Art. no. 109569, doi: 10.1016/j.compbiomed.2024.109569.

[11] R. Chauhan *et al.*, "Predictive modeling and web-based tool for cervical cancer risk assessment: A comparative study of machine learning models," *MethodsX*, vol. 12, Jun 2024, Art. no. 102653, doi: 10.1016/j.mex.2024.102653.

[12] G.S. Collins *et al.*, "Clinical prediction models using machine learning in oncology: challenges and recommendations," *BMJ oncology*, vol. 4, no. 1, Oct 2025, Art. no. e000914, doi: 10.1136/bmjonc-2025-000914.

[13] M.H. Kabir, "Study on the Performance of Classification Algorithms for Data Mining," *IOSR Journal of Computer Engineering (IOSR-JCE)*, vol. 21, no. 3, pp. 23-30, Jul 2019, doi: 10.9790/0661-2103062330.

[14] A.S.M.S. Hassan *et al.*, "A Machine Learning Approach for Optimized Heart Disease Diagnosis with SMOTE and Voting Classifiers," *International Journal of Computer Applications*, vol. 187, no. 64, pp. 30-36, Dec 2025, doi: 10.5120/ijca2025926079.

[15] C. Bunkhumpornpat *et al.*, "Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem," in *Proc. Advances in Knowledge Discovery and Data Mining*, Bangkok, Thailand, vol. 5476, pp. 475-482, Apr 2009, doi: 10.1007/978-3-642-01307-2_43.

[16] M. Afkanpour *et al.*, "Identify the most appropriate imputation method for handling missing values in clinical structured datasets: a systematic review," *BMC medical research methodology*, vol. 24, no. 1, Aug 2024, Art. no. 188, doi: 10.1186/s12874-024-02310-6.

[17] S. Matharaarachchi *et al.*, "Enhancing SMOTE for imbalanced data with abnormal minority instances Author links open overlay panel," *Machine Learning with Applications*, vol. 18, Dec 2024, Art. no. 100597, doi: 10.1016/j.mlwa.2024.100597.

[18] N. Pudjihartono *et al.*, "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction," *Frontiers in bioinformatics*, vol. 2, Jun 2022, Art. no. 927312, doi: 10.3389/fbinf.2022.927312.

[19] W. Zhai *et al.*, "A Bagging-SVM field-road trajectory classification model based on feature enhancement," *Computers and Electronics in Agriculture*, vol. 217, Feb 2024, Art. no. 108635, doi: 10.1016/j.compag.2024.108635.

[20] N. Alamsyah *et al.*, "XGBoost hyperparameter optimization using randomizedsearchcv for accurate forest fire drought condition prediction," *Journal Pilar Nusa Mandiri*, vol. 20, no. 2, pp. 103-110, Sep 2024, doi: 10.33480/pilar.v20i2.5569.

[21] P. Chithuloori and JM. Kim, "Soft voting ensemble classifier for liquefaction prediction based on SPT data," *Artificial Intelligence Review*, vol. 58, May 2025, Art. no. 228, doi: 10.1007/s10462-025-11230-w.

[22] S. Swaminathan and B. R.Tantri, "Confusion Matrix-Based Performance Evaluation Metrics," *African Journal of Biomedical Research*, vol. 27, pp. 4023-4031, Nov 2024, doi: 10.53555/AJBR.v27i4S.4345.

[23] J. H. Cabot and E. G. Ross, "Evaluating prediction model performance," *Surgery*, vol. 174, no. 3, pp. 723–726, Jul 2023, doi: 10.1016/j.surg.2023.05.023.

[24] V. V. Kumar *et al.*, "The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification," *Healthcare Analytics*, vol. 4, no. 7, Sep 2023, Art. no. 100247, doi: 10.1016/j.health.2023.100247.

[25] P. Shu *et al.*, "SHAP combined with machine learning to predict mortality risk in maintenance hemodialysis patients: a retrospective study," Frontiers in medicine, vol. 12, Jul 2025, Art. no. 1615950, doi: 10.3389/fmed.2025.1615950.

[26] P. Roy *et al.*, "Interpretable artificial intelligence (AI) for cervical cancer risk analysis leveraging stacking ensemble and expert knowledge," *Digital health*, vol. 11, Mar 2025, Art. no. 20552076251327945, doi: 10.1177/20552076251327945.

[27] A. A. Wani, "Comprehensive review of dimensionality reduction algorithms: challenges, limitations, and innovative solutions," *PeerJ. Computer science*, vol. 11, Jul 2025, Art. no. e3025, doi: 10.7717/peerj-cs.3025.

[28] Y. Li *et al.*, "An interpretable machine learning model using SHapley Additive exPlanations for preoperative cervical lymph node metastasis risk stratification in tongue squamous cell carcinoma: a multicenter study," *BMC oral health*, vol. 26, no. 1, Dec 2025, Art. no. 185, doi: 10.1186/s12903-025-07528-4.