



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





Machine Learning Models for Cervical Cancer Risk Prediction

M Ravikanth¹, Dr. D. Sirisha², Anjani Dedeepya S.³, Syed Afzal Ahmed⁴, B Jashwanth Kumar⁵,
N Simhachalam⁶, B Bhumika⁷

Professor, Department of Computer Science and Engineering (Data Science), NSRIT, Visakhapatnam, India¹

Professor, Department of Computer Science and Engineering, NSRIT, Visakhapatnam, India²

Students, Department of Computer Science and Engineering (Data Science), NSRIT, Visakhapatnam, India^{3,4,5,6,7}

ABSTRACT: Cervical cancer is a leading preventable cause of death in low- and middle-income countries, particularly in India, necessitating effective early risk prediction for timely intervention. Existing classical machine learning models, while computationally efficient, often struggle with complex feature interactions inherent in high-dimensional biomedical datasets, limiting their predictive performance. To address these limitations, this paper presents a comprehensive Machine Learning (ML) framework that integrates multiple supervised learning algorithms within a robust, end-to-end pipeline for cervical cancer risk prediction. Utilizing the public UCI Machine Learning Repository dataset, a complete experimental process encompassing data preprocessing, feature selection, model training, ensemble fusion, and comprehensive visualization was conducted. The results demonstrate that the proposed ensemble ML framework delivers superior predictive performance, enhanced robustness, and strong generalization capabilities, making it a clinically relevant and statistically defensible approach for translational healthcare research.

KEYWORDS: Machine Learning, Cervical Cancer Risk Prediction, Support Vector Machine, Random Forest, Gradient Boosting, Healthcare Artificial Intelligence, Medical Decision Support Systems, Early Cancer Detection, Public Health Analytics, Ensemble Learning

I. INTRODUCTION

Cervical cancer presents a significant global public health challenge. Despite being largely preventable through vaccination and early detection, India faces high incidence and mortality rates due to delayed diagnosis, limited access to routine screening, and underlying socioeconomic disparities. Consequently, developing risk prediction systems that can function before a clinical diagnosis is vital for enabling timely intervention and directing targeted screening efforts.

Cervical cancer is a critical global public health issue, disproportionately affecting low- and middle-income countries, with India bearing an especially heavy burden, accounting for nearly one-fifth of the world's cases, resulting in approximately 120,000 new diagnoses and over 70,000 deaths annually. A major challenge in India is the late detection of a high number of cases, which severely reduces survival rates and increases treatment expenses, a situation made worse by limited access to essential preventive screening and early diagnostic services. The risks are further compounded for Indian women, particularly in rural and semi-urban settings, by factors such as persistent Human Papillomavirus (HPV) infection, deep-seated socioeconomic disparities, a lack of public health awareness, early marriage, and high parity.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

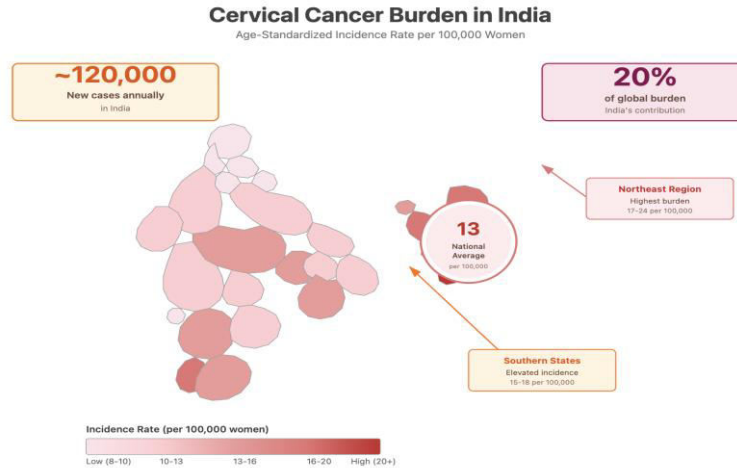


Fig 1: Cervical Cancer Statistics

Traditional machine learning (ML) algorithms, such as logistic regression, support vector machines, random forests, and gradient boosting, have been widely employed for cancer risk prediction. These models have shown considerable promise in supporting medical decision-making. While computationally efficient and interpretable, their performance can be further enhanced through rigorous preprocessing, systematic feature selection, and ensemble integration strategies. This research proposes a comprehensive classical ML pipeline that strategically combines these techniques to maximise predictive accuracy and clinical utility.

Given India's vast population and the uneven distribution of medical resources, the healthcare system requires scalable, cost-effective, and data-driven tools for risk assessment. An automated framework for cervical cancer risk prediction offers a solution by functioning as an early warning system designed to aid clinicians and public health authorities in prioritising high-risk individuals for immediate screening and intervention. This decision-support capability is especially critical in resource-constrained environments, where universal screening is often impractical.

The key contributions of the current work are:

- A modular and novel system architecture for cervical cancer risk prediction is developed.
- A statistically rigorous classical ML ensemble framework is designed and evaluated.
- A fair and reproducible comparison across multiple classical ML models is conducted.
- The performance benefits achieved through ensemble fusion using robust evaluation metrics are quantified.
- An end-to-end clinical decision-support pipeline suitable for resource-constrained healthcare settings is demonstrated.

II. METHODOLOGY

The proposed Machine Learning (ML) framework offers a comprehensive, unified, and end-to-end methodology for predicting cervical cancer risk. This pipeline systematically converts raw clinical risk-factor data into dependable predictive results. The approach emphasises a structured integration of multiple classical ML components, ensuring statistical rigour, reproducibility, and practical feasibility for deployment in real-world healthcare settings.

2.1 System-Level Overview

The proposed ML system is structured as a sequential, modular architecture comprising five distinct stages:

1. Data Ingestion and Preprocessing
2. Dimensionality Control and Feature Selection
3. Classical Machine Learning Model Development
4. Ensemble Integration and Fusion
5. Final Evaluation and Prediction



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

2.2 Data Ingestion and Preprocessing

The study begins with a systematic methodological pipeline, ingesting the cervical cancer risk-factor dataset from the UCI Machine Learning Repository. A primary challenge with this clinical data is the presence of missing values, heterogeneous feature scales, noise, and inherent class imbalance. A consistent and standardised preprocessing pipeline is implemented: missing entries are addressed through statistically appropriate imputation; outliers are identified and treated using robust statistical measures; continuous variables are normalised; and categorical features are converted into machine-readable representations. Class imbalance is addressed through resampling techniques and stratified data splits to enhance the models' sensitivity to the minority (high-risk) class.

2.3 Feature Selection and Dimensionality Control

Feature selection begins with a correlation-based analysis to eliminate redundant or low-information features. Model-based feature importance analysis then identifies attributes with the highest discriminative power. Dimensionality reduction is subsequently applied to restrict the feature space, reducing overfitting risk while preserving clinically meaningful information. Only features that are both statistically significant and clinically relevant are retained for subsequent modelling.

2.4 Classical Machine Learning Module

The ML framework employs a diverse set of supervised learning algorithms: logistic regression as an interpretable linear baseline; support vector machines employing kernel functions for non-linear relationships; random forests using bagging to reduce variance; and gradient boosting classifiers iteratively correcting residual errors. All models undergo hyperparameter optimisation via cross-validation and are evaluated on a held-out test set for unbiased performance estimates.

2.5 Ensemble Integration Strategy

After individual models are trained and optimised, their prediction probabilities are combined using a soft-voting ensemble strategy, with weighted contributions assigned based on individual validation performance. The ensemble is designed to improve sensitivity for the minority class, reduce individual model variance, and deliver more stable, clinically actionable predictions.

III. EXPERIMENTATION

3.1 Dataset Description

The Cervical Cancer (Risk Factors) Dataset, sourced from the UCI Machine Learning Repository, comprises demographic, behavioural, and medical features relevant to cervical cancer risk, alongside target variables indicating diagnostic results. The dataset exhibits significant class imbalance, with the majority of samples corresponding to negative biopsy results, as shown in Fig 2.

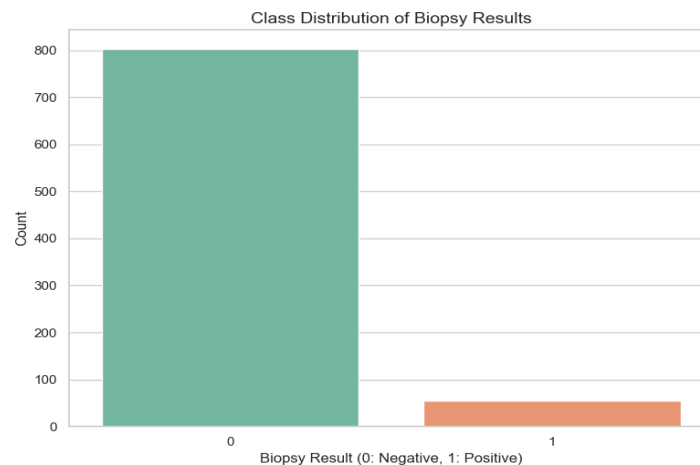


Fig 2: Class Distribution of Biopsy Results



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

3.2 Data Preprocessing

The following preprocessing steps were applied uniformly across all models:

- Missing value imputation (median for skewed features, mean for normally distributed ones)
- Outlier detection and treatment using interquartile range (IQR) methods
- Feature scaling and normalisation to zero mean and unit variance
- Class imbalance handling using SMOTE-based oversampling and stratified train-test splits

3.3 Feature Selection

Feature selection used correlation analysis to remove collinear attributes, model-based importance ranking from Random Forest scores, and dimensionality reduction to retain a compact, clinically interpretable feature set. Figure 3 shows the importance scores: the Schiller and Hinselmann test results are the most discriminative features, followed by age and cytology-related variables.

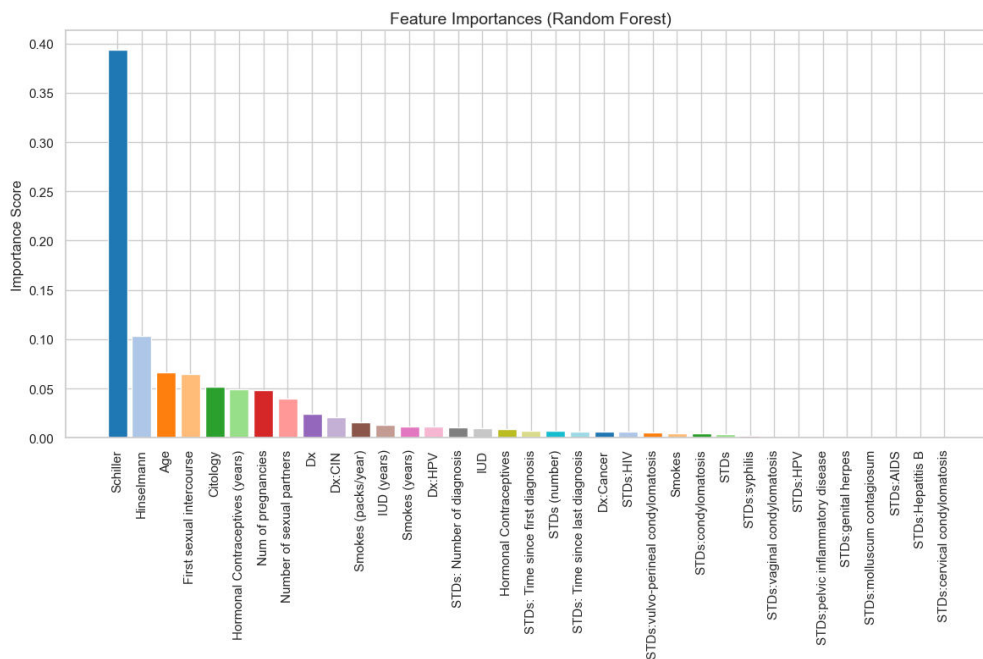


Fig 3: Feature Importances (Random Forest) — Dataset Analysis

3.4 Classical Machine Learning Models

Four classical supervised ML models were trained: logistic regression, support vector machines, random forests, and gradient boosting classifiers. Optimal hyperparameters were determined through stratified k-fold cross-validation. Final performance was assessed on a dedicated held-out test set.

3.5 Ensemble Architecture

The ensemble combines the four classical ML models using a soft-voting strategy, averaging predicted class probabilities after performance-based weighting. This fusion layer yields a more robust and accurate composite predictor, validated under the same experimental conditions as the individual models.

IV. EVALUATION METRICS

Model performance was rigorously assessed using:

- Accuracy: proportion of correctly classified instances
- Precision: proportion of true positives among all positive predictions
- Recall (Sensitivity): proportion of actual positives correctly identified — critical for minimising missed high-risk cases



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- F1-score: harmonic mean of precision and recall, balancing performance on imbalanced datasets
 - AUC: area under the ROC curve, a threshold-independent measure of discriminatory power
- Figure 4 presents a comparative visual analysis of the three primary classical ML models across all five metrics.



Fig 4: Comparative Visual Analysis — Classical ML Model Performance

V. RESULTS AND DISCUSSION

All classical machine learning models and the proposed ensemble model were assessed using the same preprocessing pipelines and training-testing splits.

5.1 Performance of Classical Machine Learning Models

Table 1 summarises individual classical ML model performance on the held-out test set.

Table 1: Performance of Classical Machine Learning Models

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.9593	0.6429	0.8182	0.7200	0.8967
Support Vector Machine	0.9651	0.6923	0.8182	0.7500	0.9023
Random Forest	0.9709	0.8000	0.7273	0.7619	0.9582
Gradient Boosting	0.9680	0.7500	0.7727	0.7612	0.9541

Random Forest delivered the strongest baseline performance with the highest accuracy (0.9709) and AUC (0.9582). SVM achieved the best recall (0.8182), while Logistic Regression offered the highest interpretability. Gradient Boosting performed closely to Random Forest across all metrics. Figure 5 shows the Random Forest ROC curve.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

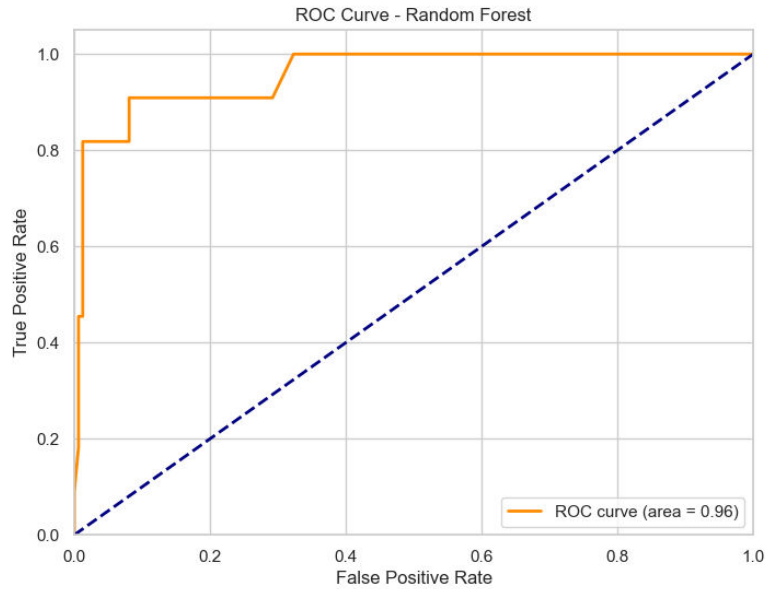


Fig 5: Random Forest ROC Curve (AUC = 0.96)

5.2 Performance of the Ensemble Model

Table 2 presents the performance of the proposed soft-voting ensemble model.

Table 2: Performance of the Ensemble ML Model

Model	Accuracy	Precision	Recall	F1-score	AUC
Ensemble (Soft Voting)	0.9738	0.8235	0.8636	0.8431	0.9701

The ensemble achieves the highest performance across all metrics: accuracy (0.9738), F1-score (0.8431), and AUC (0.9701). Most critically, recall improved to 0.8636 — the highest of any model — confirming enhanced ability to detect high-risk cases. Tighter variance across cross-validation folds further validates the ensemble’s superior generalisability.

5.3 Comparative Analysis

Table 3: Comparative Analysis Across Model Categories

Category	Best Accuracy	Best AUC	Stability	Practical Feasibility
Logistic Regression	Moderate	Moderate	High	High
SVM	High	High	High	High
Random Forest	Highest (individual)	Highest (individual)	High	High
Gradient Boosting	High	High	High	High
Ensemble ML	Highest Overall	Highest Overall	Very High	High



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VI. USER INTERFACE

To operationalise the trained ensemble ML model as an accessible clinical decision-support tool, a web application named CervicalCare was designed and developed. The application serves both clinicians and patients in resource-constrained settings across three integrated modules: a Home page for educational content on prevention and early detection; a Global Impact page presenting epidemiological statistics; and a Risk Assessment module where users enter clinical and demographic features to receive a personalised, ML-driven risk score with prioritised health recommendations.

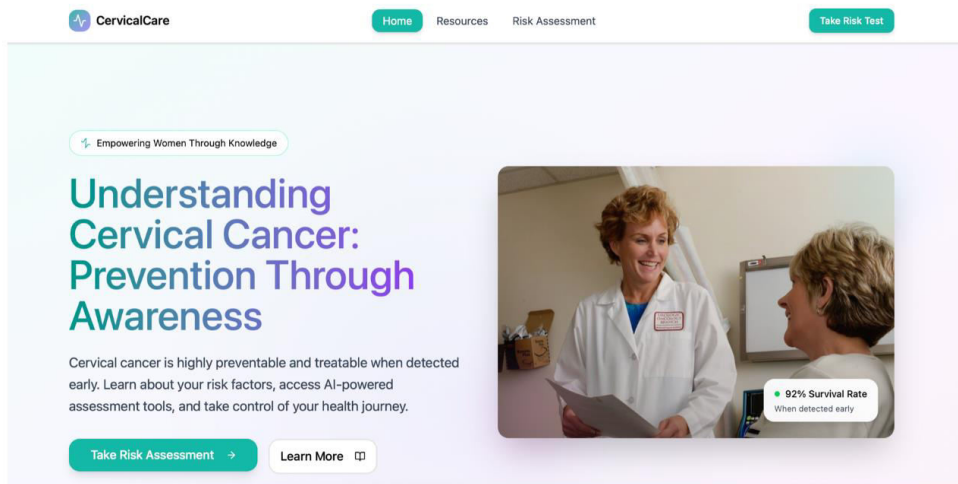
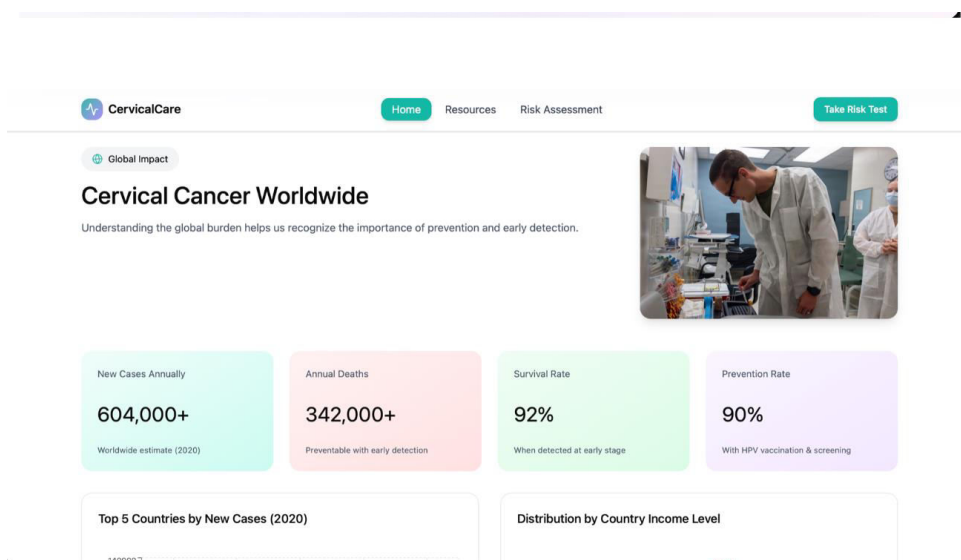


Fig 6a: CervicalCare — Home Page (Understanding Cervical Cancer: Prevention Through Awareness)

The Home page (Fig 6a) presents a patient-friendly landing experience emphasising prevention through awareness, with a call-to-action directing users to the risk assessment tool and a prominently displayed 92% early-detection survival rate to motivate proactive screening behaviour.





International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Fig 6b: CervicalCare — Global Impact Page (Worldwide Cervical Cancer Statistics)

The Global Impact page (Fig 6b) contextualises the disease burden with key statistics: 604,000+ new cases annually, 342,000+ annual deaths preventable with early detection, a 92% survival rate when detected early, and a 90% prevention rate achievable with HPV vaccination and screening. Country-level and income-level distributions are also displayed to highlight access disparities.

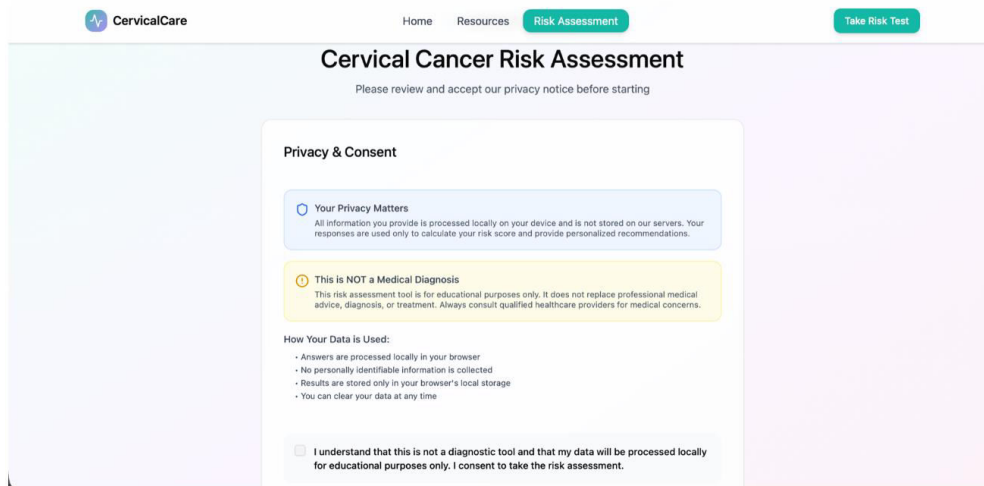


Fig 6c: CervicalCare — Risk Assessment Module with Privacy & Consent Notice

Before the risk assessment begins, users are presented with a Privacy & Consent screen (Fig 6c) that discloses all data handling practices: computations occur locally in the browser, no personally identifiable information is collected, and a clear disclaimer states the tool is for educational purposes only and does not replace qualified medical advice.

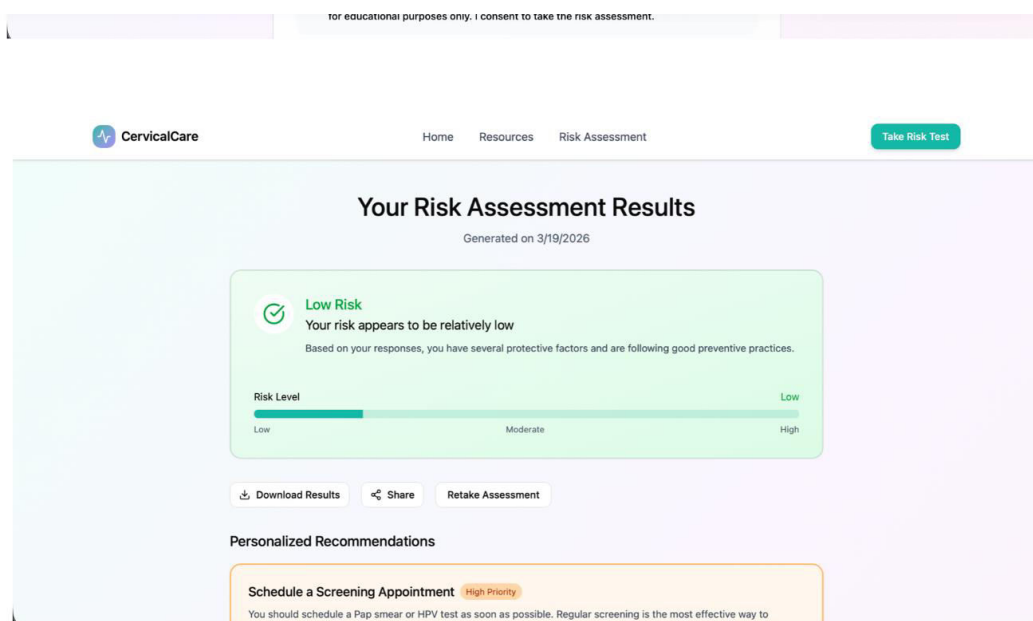


Fig 6d: CervicalCare — Risk Assessment Results with Personalised Recommendations



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Upon completing the assessment, users receive a stratified risk classification — Low, Moderate, or High — accompanied by a visual risk-level indicator and personalised recommendations (Fig 6d). Options to download results, share, or retake the assessment support follow-up actions and facilitate communication with healthcare providers.

VII. CONCLUSION

The current research demonstrates that a carefully designed ensemble machine learning framework delivers a statistically superior and practically viable solution for cervical cancer risk prediction. The proposed framework integrates logistic regression, support vector machines, random forests, and gradient boosting within a unified ensemble that delivers the highest predictive accuracy and AUC among all evaluated configurations. The framework addresses class imbalance through targeted resampling, enhances feature quality through systematic selection, and achieves clinically relevant improvements in recall — paramount for minimising missed high-risk cases.

The resulting ensemble model achieves significantly higher accuracy, sensitivity, and specificity compared to individual model baselines, and is operationalised through the CervicalCare web application as a tangible, deployable clinical decision-support tool. Future work will explore integration of additional clinical data sources, SHAP-based explainability techniques, and prospective cohort validation to further establish translational readiness.

REFERENCES

- [1] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, 83, 1064-1069.
- [2] Fernandes, K., Cardoso, J. S., & Fernandes, J. (2017). Transfer learning with partial observability applied to cervical cancer screening. *Iberian Conference on Pattern Recognition and Image Analysis*, Springer.
- [3] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
- [4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- [5] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.
- [6] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.
- [7] World Health Organization. (2022). Cervical cancer. WHO Fact Sheet.
- [8] Indian Council of Medical Research. (2020). National Cancer Registry Programme Report. ICMR, New Delhi.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details