
An Explainable Longitudinal Machine Learning Framework for Early Cervical Cancer Risk Prediction

Sneha K. Deshmukh¹, Dr. Suvarna Eknath Pawar², Sachin Arun Thanekar³, Komal Munde⁴

Department of Computer Science and Engineering,

MIT Art, Design and Technology University, Loni Kalbhor, Pune, India

¹khupsesneha68@gmail.com, ²suvarna.pawar@mituniversity.edu.in,

³sachin.thanekar@mituniversity.edu.in, ⁴komal.munde@mituniversity.edu.in

Abstract

Early cervical cancer detection can reduce mortality by enabling timely intervention. This study develops an interpretable risk-prediction framework using SVM, Random Forest, XGBoost, and longitudinal LSTM models on clinical and behavioral data. Class imbalance is addressed through feature standardization and class-weighted learning. SHAP provides global and patient-specific explanations, with age, reproductive history, and contraceptive use identified as important risk factors. The results show strong ensemble performance, while LSTM effectively captures temporal risk patterns.

Keywords: Cervical Cancer, Early Risk Prediction, Explainable Artificial Intelligence, SHAP, Long Short-Term Memory

1. INTRODUCTION

Cervical cancer remains a major health concern, particularly in low-resource settings where screening and early diagnosis are limited [1]. Machine learning (ML) supports risk prediction using demographic, clinical, and behavioral data, while XAI techniques such as SHAP and LIME improve model transparency [4,6,8–10]. However, many existing methods overlook changes across clinical visits. This study integrates ML, LSTM-based temporal learning, threshold optimization, and SHAP for interpretable cervical cancer risk prediction.

2. RELATED WORK

Machine learning models such as Random Forest, Gradient Boosting, and SVM have shown promising results for cervical cancer risk prediction using clinical and behavioral data [1,8]. SMOTE and ADASYN improve minority-risk detection, while SHAP and XAI enhance prediction interpretability [1,8]. LSTM-based models capture longitudinal EHR information [12], while radiomics and multimodal approaches provide additional predictive capabilities [13]. However, missing data, preprocessing, external validation, and temporal modeling remain key challenges [11].

3. METHODOLOGY

Figure 1 shows the overall framework of the proposed explainable longitudinal framework, which involves preprocessing, baseline modeling, LSTM prediction, threshold optimization, and interpretability analysis



Figure 3.1. Architecture diagram

3.1. Dataset Description and Preprocessing

This study relies on structured cervical cancer data comprising of demographic, behavioral and clinical attributes. Median and Mode imputation were used to deal with missing values. N-score was used to normalize numerical attributes.

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

in which μ and σ are the mean and standard deviation, respectively. Class weighted learning was used to reduce the class imbalance and stratified sampling was used to ensure the data sets for training and testing were representative and had class proportions.

3.2. Baseline Machine Learning Models

The baseline machine learning models were compared to each other by cross validation, with RF, XGBoost, and SVM. They were evaluated with Accuracy, Recall, F1-score and ROC-AUC for a complete assessment.

3.3. Longitudinal LSTM Model

To capture the temporal risk progression, a Long Short-Term Memory (LSTM) system was needed to be developed that would serve as a Longitudinal LSTM Model A (LSTM). The data given as input were reorganized into a 3-D tensor

$$(\eta_{samples}, \eta_{timesteps}, \eta_{features})$$

3.4. Decision Threshold Optimization

Though the default probability threshold (0.5) is used, threshold tuning was carried out to maximise F1-score and Recall. The optimum threshold was found to be:

$$t^* = \arg \max_{t \in [0,1]} F1(t)$$

This guarantees a better sensitivity, which is essential in medical diagnosis to avoid false negativity.

3.5. Explainability Using SHAP

SHAP was applied to provide global and patient-level interpretation with the help of Summary plots and Force plots using TreeExplainer and DeepExplainer to make predictions more transparent and clinically understandable.

3.6. Evaluation Metrics

Model performance was assessed using F1-score, Accuracy, Precision, Recall, and ROC-AUC, measuring overall correctness, high-risk identification, sensitivity, balance, and class discrimination across thresholds.

4. RESULTS AND DISCUSSION

There is limited multicollinearity between some features as can be seen in Figure 2, as the moderate correlations between age and pregnancies (0.52) and age and first intercourse

(0.36) suggest. For the cervical cancer prediction model, Figure 3 presents the SHAP summary and the contributions of the model's features to the risk of cervical cancer. Age, first sexual intercourse, sexual partners and contraceptive duration are the most important predictors, and smoking is comparatively less important.

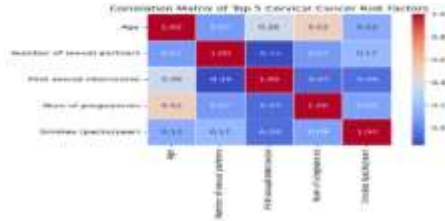


Figure 4.1: Correlation Matrix of Cervical Cancer Risk Factors

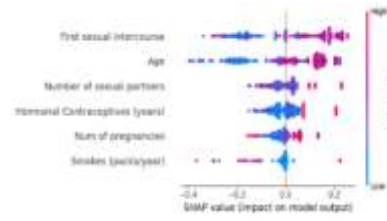


Figure 4.2: SHAP summary plot

Mean absolute SHAP values are displayed in Figure 4 by the features. Influence is strong for age at first intercourse and age, moderate for sexual partners and contraceptive use. In addition to this, pregnancy and smoking add to this. Temporal risk progression is illustrated in Figure 5, as patient 1 shows an increasing trend over the 4 visits and patient 2 consistently low risk.

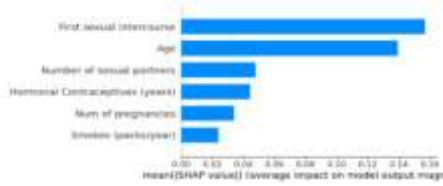


Figure 4.3: Importance of Global Features by the Mean Absolute SHAP Values to predict the risk of cervical cancer.

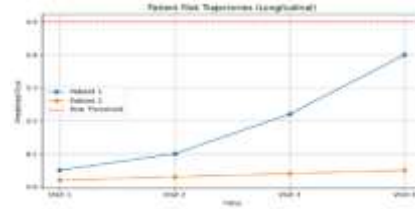


Figure 4.4: Patient risk Trajectories

TABLE I. COMPARISON OF PERFORMANCE

Model	Precision	Recall	F1-score	Accuracy	AUC
SVM	0.81	0.74	0.77	0.89	0.88
Random Forest	0.85	0.79	0.82	0.92	0.91
XGBoost	0.87	0.81	0.84	0.93	0.93
LSTM	0.80	0.89	0.84	0.90	0.92

At the 0.5 threshold, LSTM has a high accuracy of 0.90 and a high recall of 0.84, as shown in Table 1. With optimization, this recalled 0.89, enhancing sensitivity to early risk detection. It has good predictive ability with 0.92 ROC-AUC and 0.84 F1-score.

5. CONCLUSION

This study introduces a longitudinal match learning approach to the early prediction of cervical cancer risk based on clinical and behavioral features. Several machine learning models were tested (SVM, Random Forest, XGBoost, LSTM) and the SHAP model gave

interpretable predictions. The results show good ensemble results and LSTM can well reflect temporal risk patterns and enhance the sensitivity of early screening.

7. REFERENCES

- [1] R. Shakil et al., "A precise machine learning model: Detecting cervical cancer using feature selection and explainable AI," *J. Pathol. Inform.*, vol. 15, Dec. 2024.
- [2] P. Karthikeyan et al., "Explainable AI based cervical cancer prediction using FSAE feature engineering and H2O AutoML," *Sci. Rep.*, vol. 15, 40439, 2025.
- [3] S. Prusty et al., "Predicting cervical cancer risk probabilities using advanced H2O AutoML," *PeerJ Comput. Sci.*, vol. 10, e1916, 2024.
- [4] W. Chansanam et al., "An analytics-based framework for early detection of cervical cancer," *Sci. Dir.* (2025).
- [5] "An ensemble classification approach for cervical cancer prediction using behavioral risk factors," *Healthc. Anal.*, vol. 5, 100324, 2024.
- [6] R. Ashtagi et al., "Cervical Cancer Prediction Using Machine Learning," *J. Electr. Syst.*, vol. 20, No. 1s, 2024.
- [7] S. Abrar et al., "Recent advances in applications of machine learning in cervical cancer research," *Obstet. Gynecol. Sci.*, 2025.
- [8] P. Roy et al., "Interpretable artificial intelligence (AI) for cervical cancer risk analysis leveraging stacking ensemble and expert knowledge," *Digit. Health*, vol. 11, 2025.
- [9] K. Toffaha et al., "Enhancing cervical cancer risk prediction using machine-learning and explainable models," *J. Med. Data Health*, 2025.
- [10] J. Su et al., "Cervical cancer prediction using routine blood analysis and SHAP interpretation," 2024.
- [11] Y. Wu and A. Wang, "A review on the applications of artificial intelligence for diagnosing and treating cervical cancer," *Int. J. Womens Health*, 2025.
- [12] R. Weegar and K. Sundström, "Using machine learning for predicting cervical cancer from Swedish electronic health records by mining hierarchical representations," *PLoS One*, vol. 15, p. e0237911, 2020.
- [13] W. Lin et al., "Preoperative prediction of cervical cancer recurrence using explainable MRI radiomics," *Abdom. Radiol.*, 2025.

Biographies



Sneha Deshmukh is an Assistant Professor at Department of Computer Science and Engineering, MIT Art, Design and Technology University, Loni Kalbhor, Pune, India. with over 12 years of teaching experience. She is currently pursuing her Ph.D. and has research interests in Machine Learning, Artificial Intelligence, and Internet of Things (IoT).



Dr. Suvarna Pawae With 19+ years of teaching experience, she has served as a Head of Department for 15+ years, mentoring students and guiding UG/PG projects and research. She holds a Ph.D. in Artificial Intelligence and Machine Learning with expertise in IoT, prototype development, patents, copyrights, and industry-oriented solutions.



Dr. Sachin Arun Thanekar is an academician and researcher. His research interests include Artificial Intelligence, Machine Learning, Natural Language Processing, Explainable AI, and Air Quality Prediction. He has contributed to research publications, patents, academic projects, and industry-academia collaborations.



Prof. Komal Sunil Munde. She is currently working as an Assistant Professor at the Department of Computer Engineering, MITADT University. She has 9+ year experience. Her research areas include Machine learning, deep learning.