

Multi-Source Data Fusion For Early Cervical Cancer Prediction: A Synergistic Ensemble Intelligence Framework

Dr.Avantika Mahadik

Assistant Professor ,Department of Computer Science and Information Technology
Somaiya School of Basic and Applied Sciences, Somaiya Vidyavihar University,
Mumbai , Maharashtra, India.

Abstract: In developed countries, due to the adoption of structured approaches in healthcare sectors by involving prediction, detection, and prevention, cervical cancer is now becoming a highly uncommon disease. The medical services sector receives significant advantages through the implementation of machine learning techniques for the prognosis of various diseases. Diseases such as heart disease, thyroid, chronic renal disease, and various kinds of cancer can now be easily detected in their earliest stages and prevented through early medication. I observed that most of the existing research is constrained by class disparities, limited generalizability, and inadequate interpretation. With the goal to determine the best features to aid in early cervical cancer prediction, this research study utilized an ensemble machine learning approach, integrating the datasets. By combining minority class oversampling with noise reduction, the ¹SMOTE-Tomek technique enhances class separability and addresses class imbalance. The framework predicts cervical cancer with an accuracy of 97.14% using the ²LightGBM model for prediction. The SMOTE-Tomek mechanism was incorporated into the proposed framework to solve the class imbalance. Through integrating noise reduction with minority class oversampling, I was able to successfully handle class separation. ³SHAP-based explainability is used to measure feature contributions and facilitate well-informed clinical decision-making in order to guarantee transparency and clinical interpretability. In accordance with experimental results, the proposed method performs superior in terms of prediction performance, resilience, and interpretability than traditional single-model approaches. These results suggest that explainable intelligence, balanced data processing, and synergistic learning works as an effective decision-support approach to early cervical cancer detection.

Keywords: Machine Learning, Cervical Cancer, SMOTE Tomek, SHAP, Cross Validation.

I. INTRODUCTION

Cervical cancer is becoming less common in many nations, especially developed ones, because of the implementation of predictive, detective, preventative measures and systematic techniques for therapy. Deaths from cervical cancer declined from 4% annually between 1996 and 2003 to 1% annually between 2010 and 2019. (Editorial Board for Cancer.Net). Clinicians can treat precancerous malignancies early on because of screening tests, preventing them from avoid developing into cancerous growths. However, the mortality rate in developing countries keeps rising because of a lack of funding, inadequate preventative measures, a lack of publicly accessible ⁴HPV vaccine programs, as well as a deficiency in awareness-raising campaigns. Human papillomavirus and other factors can cause cervical cancer, but early detection can help prevent it. The aim of this study is to find the main factors that predict cervical cancer and to handle the problem of unequal data using different machine learning models.

¹ SMOTE Tomek – Synthetic Minority Over Sampling Techniques with Tomek Links

² LightGBM- Light Gradient Boosting Machine

³ SHAP- SHapley Additive exPlanations

⁴ HPV- Human Papillomavirus

A predictive model for cervical cancer can be developed using a multitude of methods based on machine learning [3].

Despite advancements in screening and prevention, cervical cancer remains a critical public health concern. Cervical cancer is predicted to cause 4,200 deaths and 13,490 new cases in the US in 2026. Recent trends indicate that the incidence of cervical cancer has stabilized over the past decade, although the reality that incidence rates have significantly decreased over prior decades due to widespread Pap smear screening and HPV detection. Women under 50 considers a sizable percentage of cases, underscoring the necessity of ongoing preventative measures. In order to further decrease the incidence of cervical cancer, the study highlights the significance of HPV vaccination, routine screening, and early detection [15].

Early investigations into the diagnosis of cervical cancer have primarily depended on statistical methodologies and conventional screening modalities, including Pap smear assays, HPV assessments, and colposcopy procedures [13]. Several research investigations have examined at computational methods to support clinical decision-making processes as the capabilities of machine learning evolved. Previous investigations have mostly focused on using single traditional machine learning classifiers like Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RM), and neural networks to datasets related to risk factors for cervical cancer. To improve predictive performance, feature selection and class imbalance management techniques like Principal Component Analysis (PCA), Synthetic Minority Over-sampling Technique (SMOTE), and wrapper-based optimization methods have been widely used [10]. In addition, more recent studies have explored deep learning architectures, particularly convolutional neural networks and autoencoders, which have shown significant accuracy on recognized benchmark datasets.

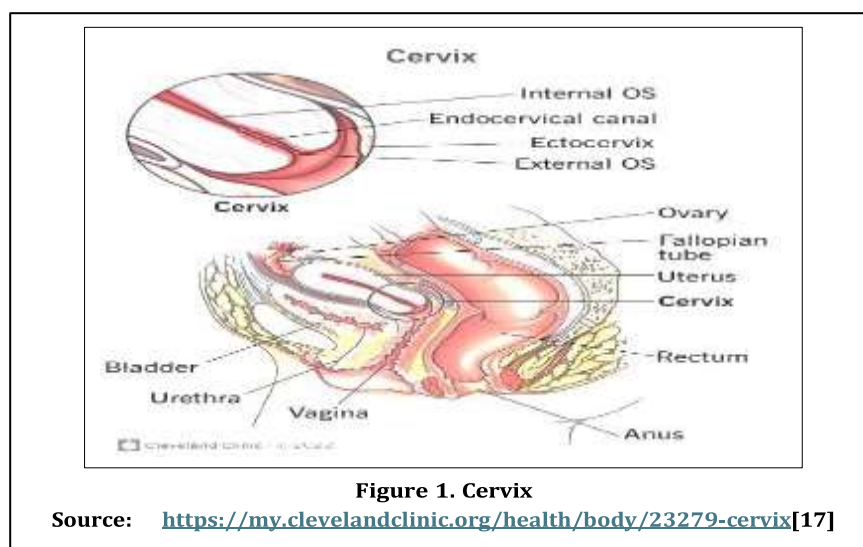
However, there are serious flaws in the present approaches. Many studies rely on a single publicly available dataset that is noticeably unbalanced, leading to an overestimation of accuracy and inadequate recall for cases that are positive. Many methods focus on feature selection despite neglecting the inherent missing values and inconsistencies in medical datasets. Furthermore, the majority of models rely on a single classifier, which makes them vulnerable to bias in datasets and the complexities of parameter selection. Even though biological and imaging data have been analysed separately, their combination with clinical risk factors has not yet received enough attention. These shortcomings highlight the need for ensemble-based frameworks incorporating multimodal extensions along with robust data correction techniques for enhancing clinical applicability, scalability, and reliability in the context of cervical cancer risk assessment.

A majority of the research is centred on a single, highly inconsistent public dataset that boosts accuracy while generating poor recall for cases that are

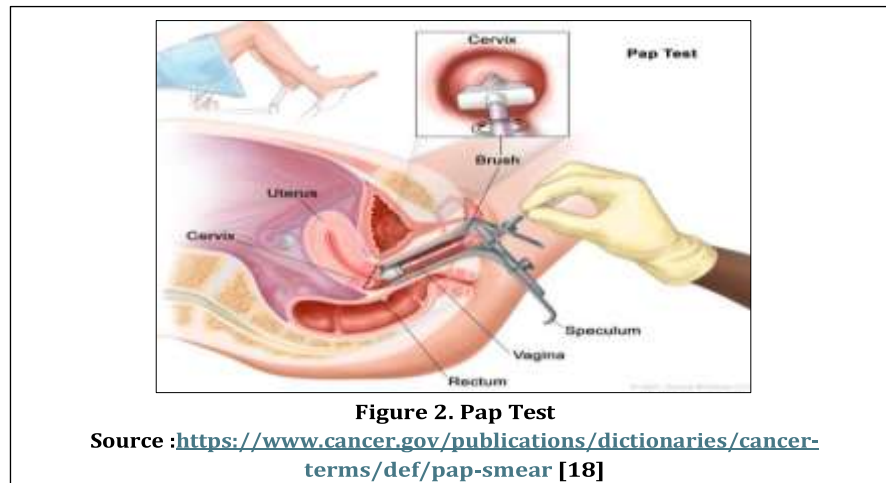
positive. A lot of techniques emphasize feature selection beyond the logical discrepancies and missing values observed in medical data. Further, although most models merely utilize single classifier, they are vulnerable to bias in the dataset and parameter selections. Whereas genetic and imaging data have been explored separately, there is still limited integration of these data with clinical risk factors. These deficiencies stimulate ensemble-based frameworks with reliable information correction strategies and multimodal expansions that would enhance clinical usability, scalability, and reliability of cervical cancer risk prediction.

Discovering the finest algorithm for machine learning among the numerous ones which can be utilized to predict cervical cancer with appropriate features is one of the primary objectives [11]. The main goal is to predict cervical cancer at early stage with all important features. Measuring the model's performance is one of the most important phases of developing an accurate machine learning model. The model's quality or performance is evaluated using a number of criteria; these metrics are also known as evaluation or performance metrics. With the help of these performance indicators, I can comprehend how effectively my model performed for the provided data. The confusion matrix, recall, accuracy, F-score, and AUC (Area Under the Curve) are indicators of performance for classification. However, the method known as regression includes the R2 score, modified R2, MAE (Mean Absolute Error), and MSE (Mean Squared Error) to measure performance [11].

Figure 1. illustrated cervical anatomy of the female reproductive system. The external and internal OS are joining the ectocervix and endocervical canal, that compose the cervix. As the main location for cervical intraepithelial neoplasia and cervical carcinogenesis, the squamocolumnar junction (transformation zone), which is situated between these regions, is very significant clinically. It has been demonstrated that nearby pelvic structures, such as the uterus, ovaries, fallopian tubes, bladder, urethra, rectum, and anus, offer both clinical and spatial context.

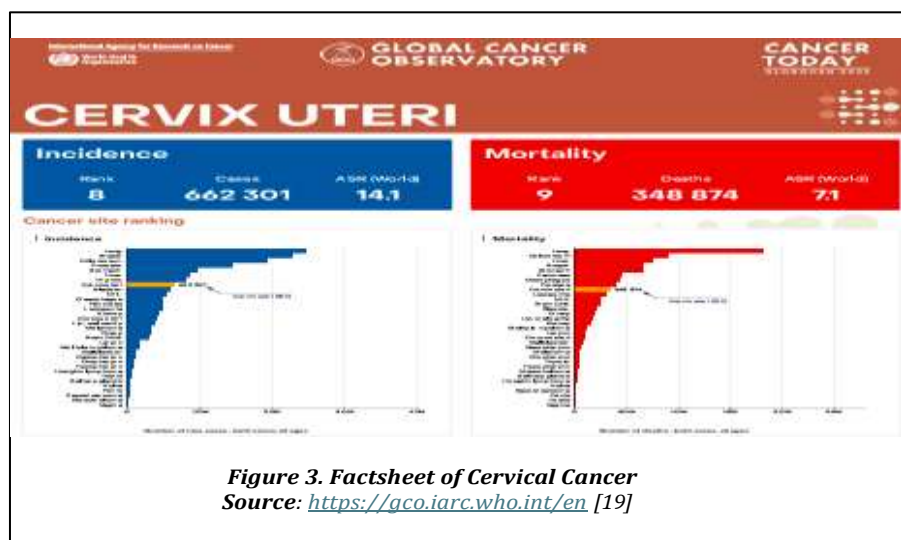


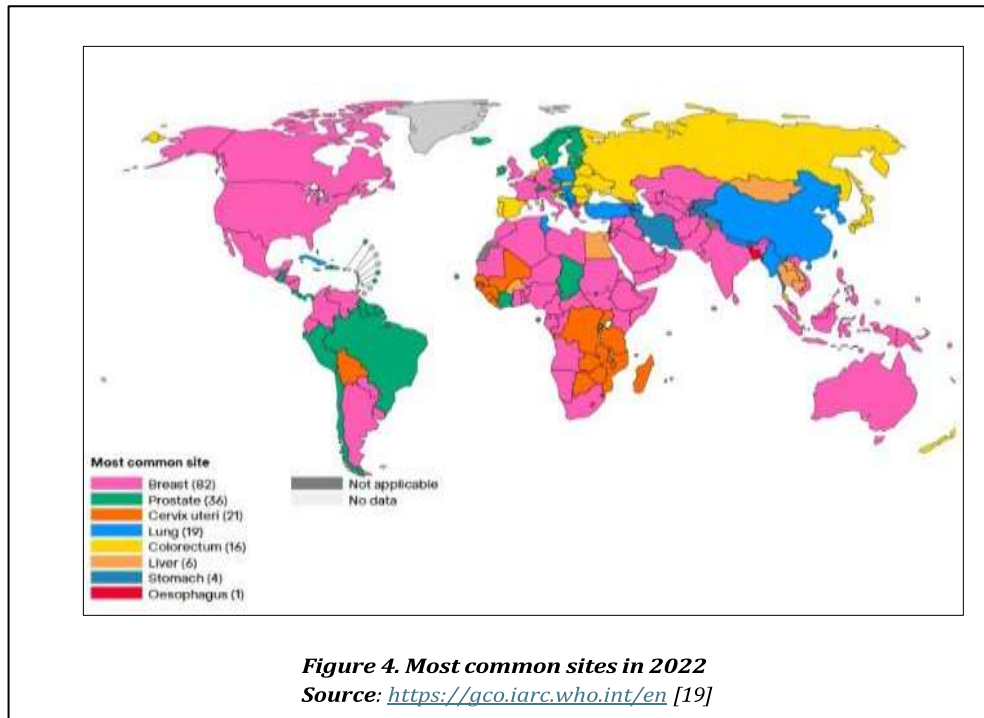
The Pap smear (Papanicolaou test) method used to screen for cervical cancer is shown schematically in Figure 2. After inserting a speculum into the vagina to examine the cervix, exfoliated epithelial cells from the ectocervix and endocervical canal are collected using a cytobrush, with a focus on the transformation zone.



To discover precancerous lesions, cervical intraepithelial neoplasia, and malignant abnormalities, the obtained cellular sample is further examined cytologically. By showing the uterus, vagina, and rectum, among other nearby pelvic tissues, the representation also offers anatomical context.

The global distribution of the most common cancer types by nation is illustrated in the fig. 3 and fig. 4, with each country color-coded according to the most common cancer occurrence. The map shows distinct regional trends in the prevalence of cancer worldwide. In most nations, especially in North America, South America, Europe, Oceania, and a significant portion of Africa and Asia, breast cancer is the most frequent type of cancer. Several South America and Northern European nations have high rates of prostate cancer [19].





Cervical cancer, on the other hand, continues to be the most common cancer type in several sub-Saharan African nations, underscoring ongoing inequalities in screening, vaccinations, and access to healthcare. Because of lifestyle and environmental risk factors, lung cancer is the most common cancer in several parts of East Asia. Additionally, in some regions, esophagus, stomach, liver, and colorectal malignancies predominate. Overall, the graph highlights significant regional variance in cancer incidence, which may be impacted by variations in lifestyle choices, environmental exposures, infection prevalence, socioeconomic status, and the accessibility of screening and prevention programs [19].

II. RELATED WORK

According to Kondo et al [1], By enhancing accessibility and privacy, machine learning techniques present a viable substitute for conventional cervical cancer screening. However, recent studies frequently use small datasets and inadequate preprocessing, which has an impact on the generalizability of the model. To create reliable as well therapeutically useful solutions, future research should prioritize ethical concerns, clarity, collaborative effort, and reliable data handling.

Luo et al (2021) [16] studied various methods for forecasting clinical outcomes in cervical cancer, such as machine learning techniques, statistical models, and medical imaging analysis. Tumor diameter, diseases stage, and histological subtype are indications of prognostic factors which are frequently identified using traditional statistical techniques like linear regression and Cox proportional hazards models. Through radiomic feature analysis, imaging modalities including CT, MRI, and PET scans additionally help in outcome prediction. Cervical cancer patients' survival, recurrence, metastasis, and treatment problems have recently been more accurately predicted using

machine learning and deep learning algorithms. Even with these improvements, problems including small datasets, overfitting, and poor interpretability still exist. All things considered, machine learning has great promise for improving outcome prediction and facilitating individualized treatment planning in the management of cervical cancer.

In [2], The authors intended to highlight that a single machine learning model is insufficient to produce a credible cervical cancer prediction. They propose that major issues including class imbalance, overfitting, and poor generalization are addressed by ensemble learning combined with appropriate data processing. The primary objective of this research is to prove that because ensemble-based approaches provide predictions that are more trustworthy, precise, and clinically significant, they have greater potential for enabling early cervical cancer screening and decision-making.

Rawat P. et al. constructed a cervical cancer prediction model in [3] using conventional machine learning algorithms. The research successfully illustrates the ability of ensemble methods and attains high accuracy, but it has a number of important constraints, such as the need to handle data imbalance and the diversity of the data. Exclusion of feature importance analysis as well. False negatives in cancer screening are crucial for assessing prediction accuracy. Srivastav

S. et al. employed a decision tree model in [4] to predict cervical cancer with 96.8% accuracy; however, the methodology primarily relied on overall accuracy rather than clinically significant measures like recall for positive cancer patients. The evaluation of model generalizability is constrained by the absence of external or real-world validation. Furthermore, model interpretability and feature importance—both crucial for therapeutic trust—receive little consideration. The study shows that machine learning is feasible for predicting cervical cancer, although it is still primarily experimental and needs more robust validation and clinically focused assessment for practical application.

Devi S. et al. developed a machine learning model in [5] using clinical screening data which had recently been collected from Indian women. The researchers investigated a few risk factors connected with cervical cancer along with a number of machine learning models, include Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, and Multi-layer Perceptron, to cope with the problem of extremely unbalanced datasets. The authors claim that Decision Trees and Logistic Regression are the most accurate and sensitive models. Important indications like smoking, sexual history, screening habits, and cervical infections have also been underlined. An investigation suggests that machine learning can improve early cervical cancer detection and awareness, even though its conclusions are challenged by a small number of confirmed cancer cases, a lack of external validation, and limited generalizability. Larger datasets and practical examination will be necessary for clinical application, despite the study's overall promising results [5].

The study performed in 2025 by Su et al. [6] shows that machine learning algorithms, especially XGBoost, may reliably predict cervical cancer using the results of traditional blood tests. The simple fact that blood tests are affordable, non-invasive, and easily available is a major advantage of this approach, making it practical for initial screening procedures. The addition of SHAP increases confidence in the model's dependability by clarifying the precise blood factors that affect the prediction results. Nevertheless, there are several limitations for

the research. Both retrospective nature and reliance on data from a single medical facility raise questions about how broadly the results may be applied to different populations. An inclination towards overfitting may be indicated by the training dataset's exceptionally high accuracy. Additionally, although platelet distribution width (PDW) has been found to be the most important characteristic, there is still insufficient evidence to support its direct biological link with cervical cancer. Before these findings are used in clinical settings, extensive, multi-institutional, and practical research is essential [6].

In [7], Lu J et al. proposed a machine-learning-based method for predicting the risk of cervical cancer using optional gene information and patient risk features. To better robustness on unbalanced datasets, the authors merge five classifiers (KNN, SVM, decision tree, logistic regression, and multilayer perceptron) using a voting-based ensemble. A logical data- correction technique is offered to address inconsistent and missing entries, and an auxiliary gene module (with PCA-based dimensionality reduction) is explored to enhance memory. Experiments on the public UCI Machine Learning Repository dataset and a private Chinese dataset show that the ensemble surpasses individual models in accuracy, precision, and F1- score. The gene module also marginally improves recollection, suggesting better identification of positive cases. The scientists point out that machine learning frameworks that incorporate many models with strong data management can more accurately predict the risk of cervical cancer than relying only on test results or individual algorithms. They emphasize the importance of accessible early screening methods, especially for populations with limited access to healthcare. By addressing data imbalance and missing values using a logical correction mechanism and a voting-based ensemble technique, the proposed method improves prediction stability [7].

The researchers in [8], Uddin K et al., used an openly available database that has a significant category of imbalance and missing values. They apply random oversampling, data preprocessing, and a combination of hybrid feature selection techniques, which include Principal Component Analysis (PCA), XGBoost, and SelectKBest, to resolve this problem. A hard-voting ensemble which includes Random Forest and Multilayer Perceptron achieves the best performance level with an accuracy of 99.19% and a sensitivity of 100% after a number of classifiers are progressively trained. Despite the high accuracy, the dataset is somewhat limited and has been deliberately balanced through Random Oversampling, which might result in overfitting lowering the reliability of the model in practical applications. The study's reliance on data from a single source limits the findings' applicability to other demographic groupings. Furthermore, a possible risk of data leaking may be indicated by the extremely high-performance levels observed. Thus, the before any practical implementation, clinical validation and testing utilizing larger and more diversified datasets continued to be necessary [8].

Ijaz et al. in [9], recommended a hybrid model for predicting cervical cancer that combines Random Forest, data balance, and outlier detection to increase accuracy. Despite the model's extremely high performance (~99%), the generalizability is limited because it only uses one dataset. Furthermore, overfitting and increased computational complexity could result from the intensive preprocessing. It is also less suitable for clinical decision-making due to its lack of explainability. In this research study, researchers achieved the high performance but lacks functional validation and interpretability—two

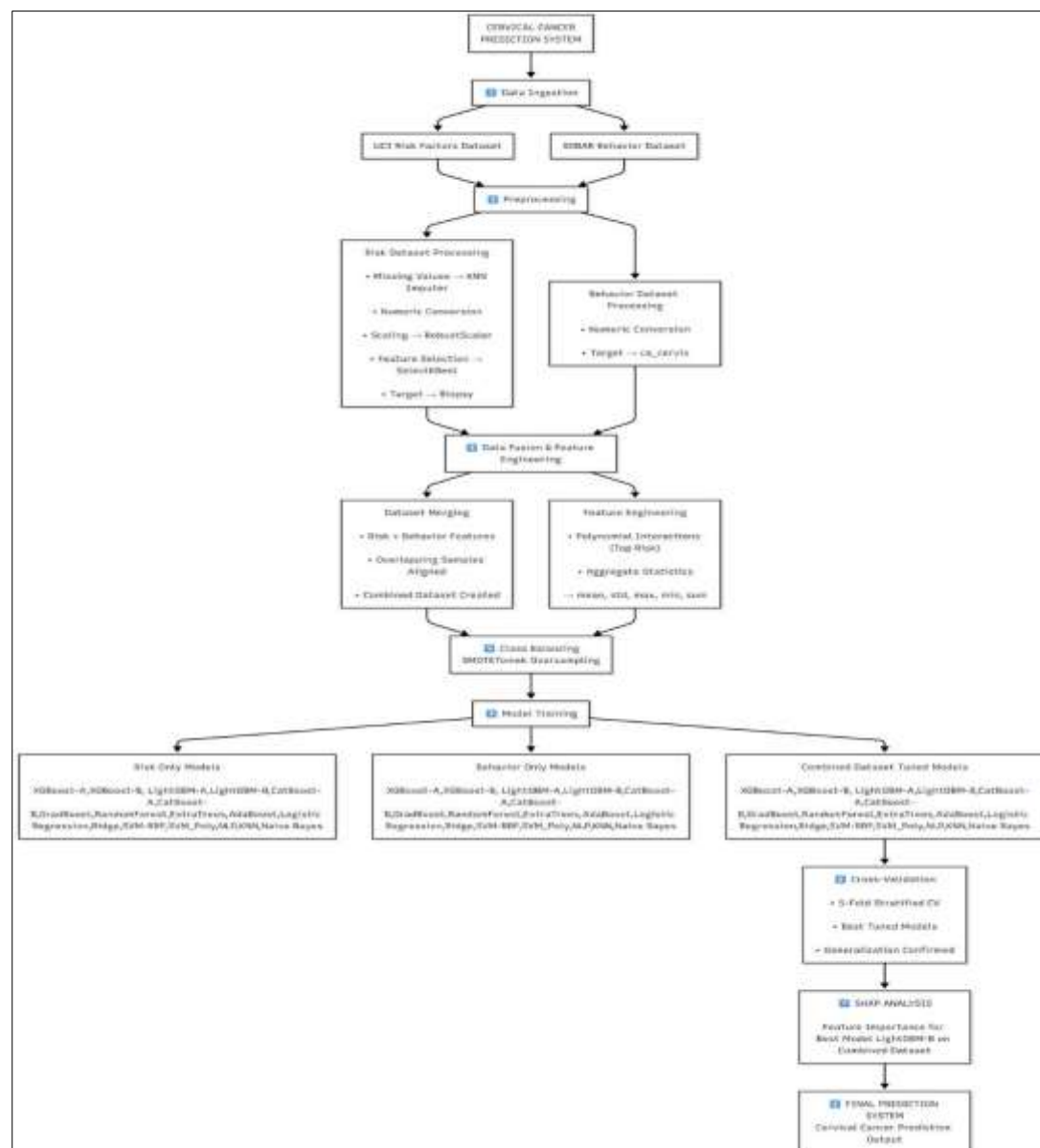
things that are necessary for healthcare applications.

III. PROPOSED MODEL

Overview

Considering the aim of promoting early and accurate cervical cancer detection, the suggested cervical cancer prediction system presents an extensive structure based on machine learning. The method integrates diverse clinical risk factor data with behavioral data from multiple sources. The system seeks to improve prediction accuracy for clinical decision assistance while preserving interpretability and resilience through feature engineering, class imbalance handling, ensemble learning, and methodical data preprocessing. Figure 5. below demonstrates the model I have been recommended.

Figure 5. The Proposed Model



A. Data Ingestion and Processing

Two separate data sources are used by the framework. Cervical cancer risk factors, such as demographics, medical history, and clinical examination results, are included in the first dataset. The second dataset includes lifestyle and behavioral characteristics linked to the risk for cervical cancer. Before undergoing additional processing, these datasets are consumed independently to maintain their inherent qualities. Table 1 and 2 showing the features of risk factors and behavioral dataset.

Table 1. Features of Risk Factor Dataset

No. of Features	Feature Name	Data Type
1	Age	Numerical
2	Number of sexual partners	Numerical
3	First sexual intercourse (age)	Numerical
4	Number of pregnancies	Numerical
5	Smokes	Categorical
6	Smokes (years)	Numerical
7	Smokes (packs/year)	Numerical
8	Hormonal Contraceptives	Categorical
9	Hormonal Contraceptives (years)	Numerical
10	Intrauterine Device (IUD)	Categorical
11	IUD (years)	Numerical
12	Sexually Transmitted Diseases (STDs)	Categorical
13	Number of STDs	Numerical
14	STDs: Condylomatosis	Categorical
15	STDs: Cervical condylomatosis	Categorical
16	STDs: Vaginal condylomatosis	Categorical
17	STDs: Vulvo-perineal condylomatosis	Categorical
18	STDs: Syphilis	Categorical
19	STDs: Pelvic inflammatory disease	Categorical
20	STDs: Genital herpes	Categorical
21	STDs: Molluscum contagiosum	Categorical
22	STDs: AIDS	Categorical
23	STDs: HIV	Categorical
24	STDs: Hepatitis B	Categorical
25	STDs: HPV	Categorical
26	STDs: Number of diagnoses	Numerical
27	STDs: Time since first	Numerical

	diagnosis	
28	STDs: Time since last diagnosis	Numerical
29	Diagnosis: Cancer	Categorical
30	Diagnosis: CIN	Categorical
31	Diagnosis: HPV	Categorical
32	Diagnosis (general)	Categorical
33	Hinselmann test result	Categorical
34	Schiller test result	Categorical
35	Cytology result	Categorical
36	Biopsy result	Categorical

i. Risk Factor Dataset Processing:

The clinical risk factor dataset undergoes several preprocessing steps to enhance data quality. Table 1., showing the thirty-six features before feature engineering. Missing values are addressed using the K-Nearest Neighbors (KNN) imputation technique, which estimates missing entries based on similarities among instances. To mitigate the influence of outliers, feature values are scaled using the Robust Scaler method. The SelectKBest technique then serves for feature selection with the objective to maintain the most relevant predictors. The biopsy result, which represents the diagnosis of cervical cancer, has been selected as the target variable.

Table 2. Features of Behavioral Dataset

No. of features	Feature Name	Data Type
1	behavior_sexualRisk	Numerical
2	behavior_eating	Numerical
3	behavior_personalHygine	Numerical
4	intention_aggregation	Numerical
5	intention_commitment	Numerical
6	attitude_consistency	Numerical
7	attitude_spontaneity	Numerical
8	norm_significantPerson	Numerical
9	norm_fulfillment	Numerical
10	perception_vulnerability	Numerical
11	perception_severity	Numerical
12	motivation_strength	Numerical
13	motivation_willingness	Numerical
14	socialSupport_emotional ity	Numerical
15	socialSupport_appreciati on	Numerical
16	socialSupport_instrumen tal	Numerical
17	empowerment_knowledg e	Numerical

18	empowerment_abilities	Numerical
19	empowerment_desires	Numerical
20	ca_cervix	Categorical

ii. Behavioral Dataset Processing:

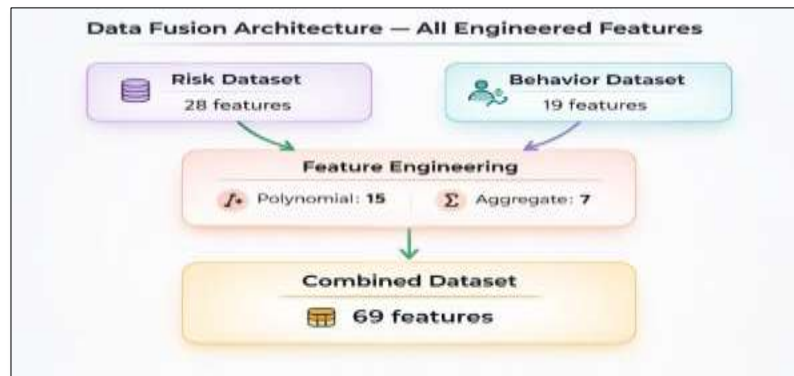
The behavioral dataset is processed independently. In order to ensure seamless integration with machine learning techniques, categorical features have been transformed into numerical representations, which is illustrated in table 2. The target labels correspond to the result of cervical cancer. In order to capture non-linear interactions among behavioral variables and enhance the dataset's expressive power, polynomial interaction features are additionally generated.

B. Data Fusion and Feature Engineering

i. Dataset Integration

The risk factor and behavioral datasets are merged in order to generate a single dataset that is depicted in Figure 6. following preprocessing and class balancing. To guarantee consistency and integrity across data sources, overlapping instances are meticulously aligned.

Figure 6. Combined Dataset



ii. Feature Engineering

Additional features are designed to enhance the model's performance substantially more. Statistical aggregation measurements like mean, standard deviation, maximum, minimum, and total are among them. To capture complex interactions between variables, polynomial interaction characteristics are also included. The capability of the model to identify distinguishing patterns from the combined dataset is improved by this feature engineering strategy. In the figure 7., shown the comparison of model performance based on multiple feature sets for the prediction of cervical cancer is displayed. With an accuracy of 97.09%, the risk-factor-based model employing CatBoost demonstrated a remarkable predictive capacity based only on clinical and demographic characteristics. The behavioral data-based CatBoost model, on the other hand, had slightly lower accuracy of 93.33%, indicating that although helpful, behavioral features may not be sufficiently discriminative when applied independently. Additionally, the combined

model that used LightGBM to integrate both risk and behavioral characteristics attained the best accuracy of 100.0%, demonstrating the efficacy of feature fusion and ensemble learning. This notable improvement implies that integrating diverse data sources improves the model's capacity to identify intricate patterns, increasing prediction accuracy and overall diagnostic dependability in the identification of cervical cancer.

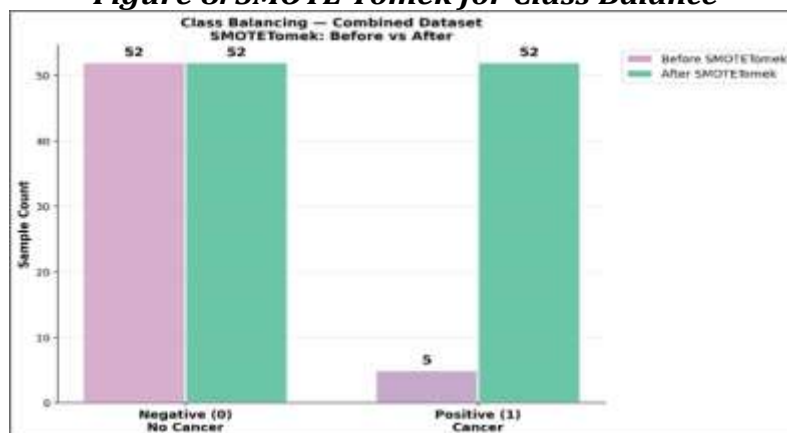
Figure 7. Dataset Performance Comparison



C. Class Imbalance Handling

The suggested approach uses the SMOTE Tomek technique to balance the distribution of classes since medical datasets frequently display an inherent class imbalance which are shown in the figure 8. By mixing Tomek link removal with synthetic minority oversampling, this technique increases the representation of minority classes while minimizing noise and overlapping data levels. Figure 8. represents the combined dataset's class distribution equally before and after the SMOTE Tomek resampling approach. There was a notable class imbalance in the dataset before resampling, involving 52 samples in the negative class (no cancer) while there were only 5 samples in the positive class (cancer). Such an imbalance might distort machine-learning algorithms in favour of the majority class and deteriorate predictive performance for the minority class.

Figure 8. SMOTE Tomek for Class Balance



The solution to this challenge has been SMOTE Tomek, which combines the Synthetic Minority Over-sampling Technique (SMOTE) with Tomek associations. Whereas SMOTE generates synthetic samples for the minority class, Tomek associations eliminate noisy and overlapping data from both classes to improve class separability. The dataset is fully balanced after resampling, with 52 samples in each class, as the image illustrates. This balanced distribution minimizes bias, strengthens model robustness, and generates more accurate cervical cancer case forecasts by guaranteeing that all groups have equal representation in the learning methods.

Nitesh V. Chawla and co-authors Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer initially put forward the SMOTE (Synthetic Minority Over-sampling Technique) in 2002. The approach was described in the article titled "SMOTE: Synthetic Minority Over-sampling Technique," which was published in the Journal of Artificial Intelligence Research [21]. In order to resolve the challenge of class imbalance in machine learning datasets, in [21], Chawla et al. recommend SMOTE (Synthetic Minority Over-sampling Technique). SMOTE generates synthetic instances by interpolating between a minority sample and its closest neighbors, so extending the minority decision region, as opposed to merely reproducing minority class samples. The study shows that classifier learning is enhanced when SMOTE and majority class under-sampling are combined. Experiments with classifiers like C4.5, Ripper, and Naïve Bayes demonstrate better minority class instance detection. When compared to conventional over-sampling or under sampling techniques, the findings assessed using ROC curves and AUC metrics show superior performance. The authors come to the conclusion that SMOTE is a useful preprocessing method for enhancing classification performance on datasets that are unbalanced [21].

Numerous data resampling approaches have been used to solve the issue of class imbalance in machine learning. Batista et al. [22] carried out a thorough investigation to assess several approaches for training data balancing. In addition to data cleaning techniques like Tomek Links, Edited Nearest Neighbor (ENN), and Neighborhood Cleaning Rule (NCL), their study contrasted conventional sample methods with synthetic sampling strategies like SMOTE. The authors also suggested hybrid approaches that balance the dataset while eliminating noisy and overlapping samples, such as SMOTE with ENN and SMOTE with Tomek Links. These hybrid approaches greatly enhance classifier performance, especially for severely imbalanced datasets, according to experimental results on datasets from the UCI Machine learning repository [22].

D. Model Training

Three experimental configurations—models trained basically on risk factor data, models trained solely on behavioral data, and models trained on the fused dataset—are employed by the framework to assess a number of machine learning models. In addition to conventional machine learning algorithms like Support Vector Machines, k-Nearest Neighbors, Naïve Bayes, Logistic/Ridge Regression, and Multilayer Perceptron, the assessed classifiers also include ensemble-based techniques like XGBoost, LightGBM, CatBoost, and Random Forest. To maximize the performance of the selected models, hyperparameter adjustment is done.

E. SHapley (SHAP) analysis

A model interpretability mechanism known SHapley Additive exPlanations (SHAP) is used to explain supervised machine learning model predictions. Shapley values from cooperative game theory, which properly distribute each feature's contribution to the model's prediction, serve as its foundation. SHAP gives both local explanations, which illustrate how individual factors affect a particular prediction, and global explanations, which indicate the overall value of features. Through the use of visualization tools including bar plots, beeswarm plots, scatter plots, and waterfall plots, which aid in the analysis of feature impact and linkages within the model, the method facilitates interpretation. Regression or classification algorithms, neural networks, and tree-based models are just a few of the machine learning models that SHAP can be used with. Researchers are able to more effectively understand the elements influencing predictions due to the analysis, which also increases model transparency. However, when evaluating SHAP results, constraints like processing complexity and possible problems with correlated features must be considered [20].

IV. MODEL EVALUATION AND RESULTS

Validation Strategies

A percentage of the available data (usually between 20–30% of the total dataset) had been selected at random to produce a single endurance test set. This test set was used exclusively for the final model evaluation and was maintained completely distinct from the training procedure. The accuracy of the test set has been estimated as:

$$\text{Test Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Test Samples}} \times 100\%$$

Cross-Validation

K=5 or k=10 folds (precise value to be determined dependent on dataset size and computational restrictions) were used to implement K-fold cross-validation. For each model:

- I. K equal-sized folds have been established from the training data (without considering the test set).
- II. Using k-1 folds for training and one-fold for validation, the model was trained k times.
- III. Each fold's accuracy was recorded.
- IV. For each of the k folds, the mean accuracy and standard deviation were determined. While the standard deviation measures model stability, the cross-validation mean accuracy offers a more reliable performance estimate than a single holdout test:

$$\text{CV Mean Accuracy} = \frac{1}{k} \sum_{i=1}^k \text{Accuracy}_i$$

$$CV\ Std = \sqrt{\frac{1}{k-1} \sum_{i=1}^k (Accuracy_i - CV\ Mean\ Accuracy)^2}$$

Performance Metrics

For each model, I recorded:

1

$k-1$

k

$\sum_{i=1}^k (Accuracy_i$

i

$- CV\ Mean\ Accuracy)^2$

The efficiency of the model has been assessed using an independent test dataset. Accuracy, precision, recall, F1-score, and the area under the Receiver Operating Characteristic curve (ROC-AUC) are evaluation metrics. A comprehensive assessment of classification performance is ensured by the use of several evaluation measures, especially when it comes to minority class detection. The best-performing model, test and cross-validation accuracy scores, and comprehensive performance data for all assessed models are among the outputs that the system produces and saves. Furthermore, SHAP-based feature importance analysis is used to improve transparency and interpretability for clinical stakeholders by offering explainable insights into model predictions.

The suggested framework produces a reliable and comprehensible cervical cancer prediction system that can produce precise early-stage forecasts. The system shows great promise as a helpful tool for early cervical cancer screening and well-informed clinical decision-making by combining multi-source data fusion, class imbalance handling, ensemble learning, and explainable artificial intelligence methodologies.

The patient's static risk profile was represented by a subset of eighteen features that were retrieved. These characteristics function as the fundamental structural components of the model's decision-making procedure. Fifteen behavioral characteristics were chosen to record dynamic activity. By giving the model a time context, these variables enable it to adjust to shifting user behavior and past trends. The pipeline incorporates a special feature engineering layer to account for intricate, non-linear connections that raw variables are unable to capture on their own. To describe the synergistic relationship between risk indicators and behavioral patterns, five features were constructed for polynomial interactions, one of the two types of synthetic variables used in this layer.

The model can identify correlations that only emerge when multiple factors are considered simultaneously according to these interactions. To summarize long-term behavioral patterns, two more features have been developed. Through providing a smoothed representation of activity, these aggregates act as statistical anchors and decrease the model's sensitivity to short term volatility.

Twenty machine learning models are compared in this evaluation using a variety of metrics, as shown in Table 3 below. F1-score, accuracy (mean and standard deviation), recall, precision, and AUC. The primary objectives are to discover the best performers overall and highlight any significant patterns in model behavior.

Figure 9. Working Flow of Cervical Cancer Prediction System

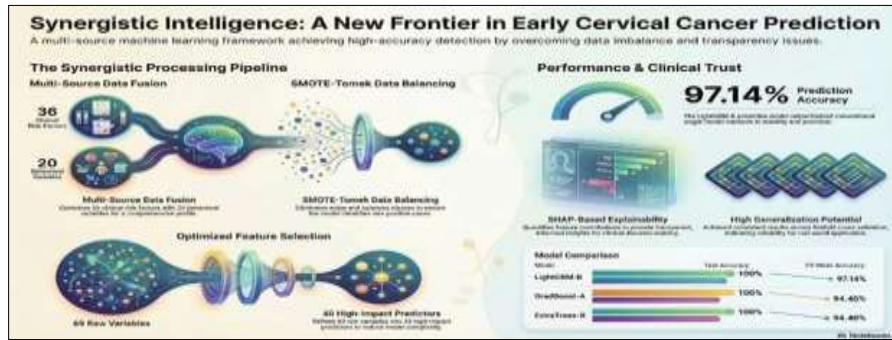


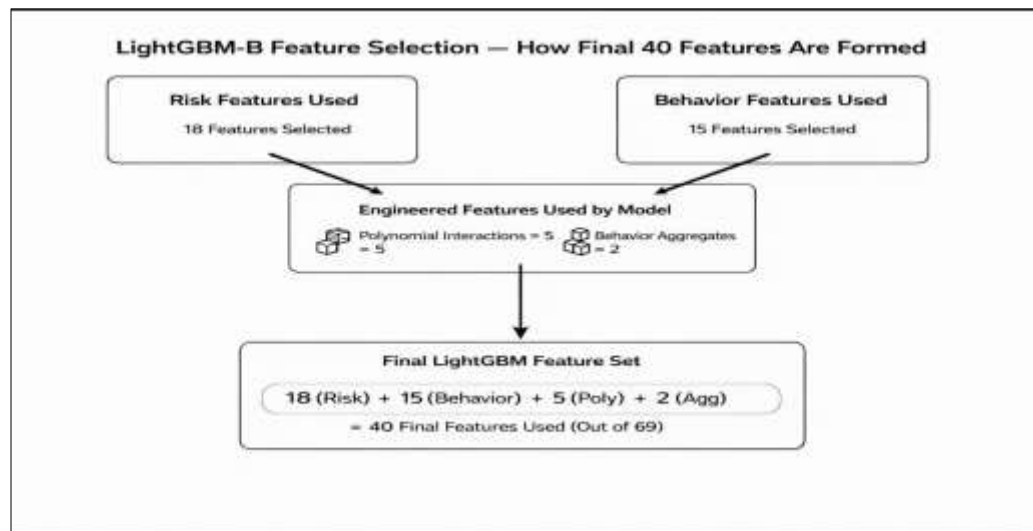
Table 3. Cross Validation Result of Twenty Models

No	Model	CV_Accuracy_Mean	CV_Accuracy_Std	CV_Precision_Mean	CV_Recall_Mean	CV_F1_Mean	CV_AUC_Mean
1	LightGBM-B	97.14285714	3.499271061	0.7	0.8	0.7333333333	0.984615385
2	LightGBM-A	95.80952381	3.425924904	0.5	0.6	0.5333333333	0.984615385
3	XGBoost-C	94.38095238	2.815570571	0.3	0.4	0.3333333333	0.984615385
4	XGBoost-A	94.38095238	2.815570571	0.3	0.4	0.3333333333	0.984615385
5	Gradient Boosting-B	94.38095238	5.323119471	0.4	0.4	0.4	0.984615385
6	CatBoost-A	94.38095238	5.323119471	0.4666666667	0.6	0.5	0.984615385
7	CatBoost-B	94.38095238	5.323119471	0.4666666667	0.6	0.5	0.984615385
8	Gradient Boosting-A	94.38095238	5.323119471	0.6	0.8	0.66	0.984615385
9	ExtraTree-B	94.38095238	5.323119471	0.4	0.4	0.4	0.984615385
10	AdaBoost	93.04761905	6.183879269	0.4	0.4	0.4	0.884615385
11	XGBoost-B	92.95238095	4.521553322	0.2666666667	0.4	0.3	0.984615385
12	RandomForest-A	91.71428571	2.530539096	0	0	0	0.888461538
13	ExtraTree-A	91.71428571	2.530539096	0	0	0	0.931868132
14	RandomForest-B	91.71428571	2.530539096	0	0	0	0.894505495

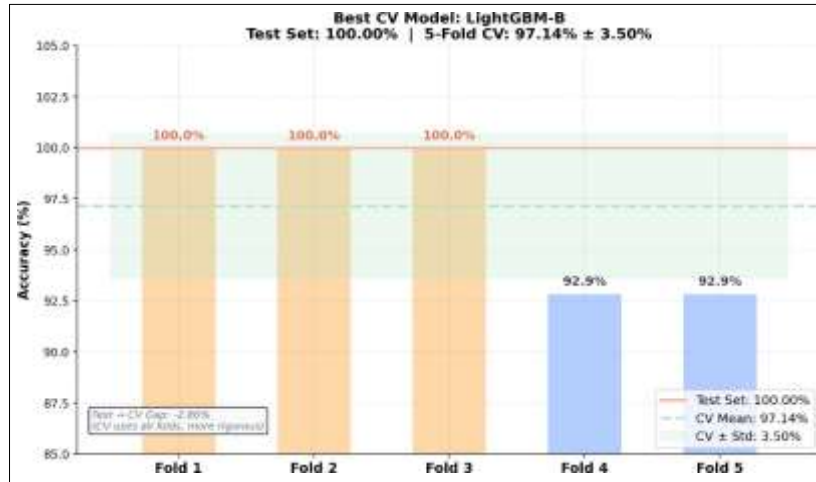
15	SVM-Poly	91.71428571	2.530539096	0.2	0.4	0.2666666667	0.908791209
16	MLP-B	89.04761905	5.119878116	0.05	0.2	0.08	0.701098901
17	MLP-A	88.95238095	5.297498661	0.0666666667	0.2	0.1	0.728571429
18	SVM-RBF	87.52380952	6.743771797	0.1	0.2	0.1333333333	0.807692308
19	KNN-5	82	8.250060125	0.14	0.3	0.1666666667	0.624175824
20	KNN-3	82	7.091242083	0.05	0.2	0.08	0.565934066

Figure 10. depicted the final feature set, which consists of 40 variables, a considerable decrease from the initial pool of 69 candidates. The model achieves a more favourable balance between complexity and generalizability by using fifty eight percentage of the given data.

Figure 10. Features set of LightGBM



A number of benefits of this streamlined approach include moderate parsimony, latency inference, and overfitting reduction. LightGBM-B performed additional fivefold cross validation test after collecting forty of the sixty-nine greatest features. Figure 11. displays LightGBM-B, the better classification model. Five-fold cross-validation and an independent test set have been used for evaluating this. With a flawless test accuracy of 100%, the model demonstrated remarkable predictive power on unknown data. All five cross-validation folds showed constant excellent performance from the model. Folds 1, 2, and 3 all achieved 100% accuracy, while Folds 4 and 5 yielded significantly lower but still remarkable accuracies of 92.9%. With a mean cross-validation accuracy of 97.14% and a standard deviation of $\pm 3.50\%$, the model exhibited both high accuracy and acceptable variability across folds.

Figure 11. Fivefold Cross Validation of LightGBM

The LightGBM-B model depends upon a hierarchical feature selection pipeline to achieve its predictive performance. Using this method, an extensive number of candidate variables are transformed into a high-impact, refined feature set that promotes information gain while reducing stochastic noise. Figure 9. shows the transformation from the raw data pool to the final model input. To present a comprehensive picture of the target variable, the model architecture aggregates data from two separate subject domains.

The modest discrepancy (-2.86%) between test accuracy and cross-validation mean, which shows minimal overfitting, supports the robustness of the estimated framework. Because cross-validation evaluates the model on many data partitions, it offers a more comprehensive and accurate assessment of generalization performance than a single train-test split. These results demonstrate that the given LightGBM-based model predicts cervical cancer in a reliable, accurate, and stable manner, making it appropriate for practical screening and decision-support applications. In displays each model's cross-validation score. LightGBM-B is the model with the highest prediction accuracy (97.14%). Figure 12. shows the twenty models cross validation score on combined dataset.

Figure 12. Cross Validation Score

CV RESULTS TABLE - Combined Dataset				
Model	CV Acc Mean	Std	F1	AUC
LightGBM-B	97.14%	3.50%	0.7333	0.9846
LightGBM-A	95.81%	3.43%	0.5333	0.9846
XGBoost-C	94.38%	2.82%	0.3333	0.9846
XGBoost-A	94.38%	2.82%	0.3333	0.9846
GradBoost-B	94.38%	5.32%	0.4000	0.9846
CatBoost-A	94.38%	5.32%	0.5000	0.9846
CatBoost-B	94.38%	5.32%	0.5000	0.9846
GradBoost-A	94.38%	5.32%	0.6600	0.9846
ExtraTrees-B	94.38%	5.32%	0.4000	0.9846
AdaBoost	93.05%	6.18%	0.4000	0.8846
XGBoost-B	92.95%	4.52%	0.3000	0.9846
RandomForest-A	91.71%	2.53%	0.0000	0.8885
ExtraTrees-A	91.71%	2.53%	0.0000	0.9319
RandomForest-B	91.71%	2.53%	0.0000	0.8945
SVM-Poly	91.71%	2.53%	0.2667	0.9088
MLP-B	89.05%	5.12%	0.0000	0.7011
MLP-A	88.95%	5.30%	0.1000	0.7256
SVM-RBF	87.52%	6.74%	0.1333	0.8077
KNN-5	82.00%	8.25%	0.1667	0.6242
KNN-3	82.00%	7.09%	0.0000	0.5659

BEST CV MODEL: LightGBM-B
Test Acc: 100.00% (fixed hold-out test set)
CV Acc: 97.14% ± 3.50% (5-fold generalization)
Interpretation: 100% on test split; -97.1% average across all folds

SHAP ANALYSIS - LightGBM-B on Combined Dataset
(Using best CV model: LightGBM-B - CV Acc: 97.14%)

Individual Model Performance 1 LightGBM-B

- Test Set Accuracy: 100%
- CV Mean Accuracy: 97.1%
- CV Standard Deviation: $\pm 3.6\%$
- Performance Gap: 2.9%

GradBoost-A

- Test Set Accuracy: 100%
- CV Mean Accuracy: 94.4%
- CV Standard Deviation: $\pm 5.0\%$
- Performance Gap: 5.6%

GradBoost-B

- Test Set Accuracy: 100%
- CV Mean Accuracy: 94.4%
- CV Standard Deviation: $\pm 5.0\%$
- Performance Gap: 5.6%

ExtraTrees-B

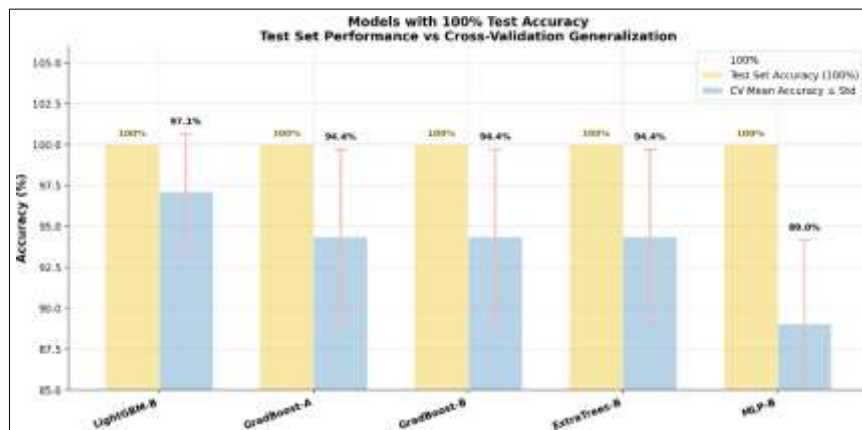
- Test Set Accuracy: 100%
- CV Mean Accuracy: 94.4%
- CV Standard Deviation: $\pm 5.0\%$
- Performance Gap: 5.6%

MLP-B

- Test Set Accuracy: 100%
- CV Mean Accuracy: 89.0%
- CV Standard Deviation: $\pm 5.4\%$
- Performance Gap: 11.0%

Top five models depicted in the following fig. 11. Which shown the LightGBM is the best model among five for prediction of cervical cancer at early stage.

Figure 11. Top five models



After selecting top five best models which are shown in the Table 4 below, I concluding that LightGBM-B giving 97.14% accuracy in the prediction of cervical cancer.

Table 4. Comparative Performance Metrics

Rank	Model	Test Accuracy	CV Mean Accuracy	CV Std Dev	Performance Gap
1	LightGBM-B	100%	97.10%	±3.6%	2.90%
2	GradBoost-A	100%	94.40%	±5.0%	5.60%
3	GradBoost-B	100%	94.40%	±5.0%	5.60%
4	ExtraTrees-B	100%	94.40%	±5.0%	5.60%
5	MLP-B	100%	89.00%	±5.4%	11.00%

In fig. 12 shown the final summary of all five best models among twenty models. Cross validations are done on the combined dataset.

Figure 12. Final Summary

=====		
FINAL SUMMARY		
=====		
Models achieving 100% on Test Set (Combined):		
LightGBM-B	Test: 100.00%	CV: 97.14% ± 3.50%
GradBoost-A	Test: 100.00%	CV: 94.38% ± 5.32%
GradBoost-B	Test: 100.00%	CV: 94.38% ± 5.32%
ExtraTrees-B	Test: 100.00%	CV: 94.38% ± 5.32%
MLP-B	Test: 100.00%	CV: 89.05% ± 5.12%
Best CV Model: LightGBM-B		
Test Set: 100.00%		
CV (5-fold): 97.14% ± 3.50%		
Features used by model: 40 / 69		

VI. CONCLUSION

Through the integration of diverse clinical and behavioral datasets, this study developed a comprehensive machine learning framework for the prediction of cervical cancer. A more robust description of the variables affecting cervical cancer risk is made possible by the proposed approach, which combines data from the SOBAR behavioral Risk Dataset and the UCI Cervical Cancer Risk Factors Dataset. Predictive performance was greatly enhanced by the integration of several data sources as well as extensive preprocessing and feature engineering approaches. During the preprocessing phase, the hybrid resampling technique SMOTE Tomek was implemented to resolve the challenge of class imbalance frequently found in medical datasets. By balancing the majority and minority classes and eliminating noisy observations, this method improved the robustness of the model. In addition, by reducing the feature space from 69 to 40 significant features through feature selection and engineering qualities, computational efficiency and model interpretability were enhanced. This study analysed a composite dataset to evaluate twenty machine learning models. Merging the behavioral with risk factors datasets yields the best prediction accuracy. We can use LightGBM-B to predict cervical cancer at an early stage with an accuracy of 97.14% due to the integration of the

dataset, the ensemble technique, SMOTE Tomek, and SHAP analysis. Several machine learning models achieved 100% classification accuracy on the test dataset, in accordance with the experimental evaluation, demonstrating the excellent predictive capability of the recommended framework. With a cross-validation accuracy of $97.14\% \pm 3.50\%$, which indicates excellent generalization potential over numerous validation folds, the LightGBM-based model performed the best overall among the examined models. Competitive performances were also attained by other ensemble models, demonstrating the efficiency of tree-based algorithms and gradient boosting in managing intricate linkages in the combined dataset.

Furthermore, by identifying the most significant clinical and behavioral aspects influencing cervical cancer risk, the integration of feature importance analysis utilizing SHAP-based interpretation optimizes the prediction process's transparency. In healthcare applications, where comprehending the underlying logic of predictive models is essential for clinical adoption, this interpretability is especially beneficial. The suggested approach shows how combining diverse datasets, sophisticated feature engineering, and reliable machine learning algorithms can significantly improve the precision of cervical cancer prediction. The findings demonstrate how data-driven predictive models can facilitate risk assessment and early identification, helping medical practitioners make well-informed decisions. With the aim to improve early screening and preventive healthcare techniques to detect cervical cancer, future study will concentrate on constructing deployable clinical decision support systems, investigating deep learning-based models, and testing the suggested strategy using larger real-world clinical datasets.

Funding Statement: The author received no specific funding for this study.

Conflicts of Interest: The author declare that no conflicts of interest to report regarding the present study.

REFERENCES

1. Kondo, T. S., Ngondya, D., & Rusheke, H. (2025). Application of artificial intelligence in cervical cancer diagnosis using risk factors: A systematic review. *Telematics and Informatics Reports*, 20, 100250.
2. Uddin, K. M. M., Al Mamun, A., Chakrabarti, A., Mostafiz, R., & Dey, S. K. (2024). An ensemble machine learning-based approach to predict cervical cancer using hybrid feature selection. *Neuroscience Informatics*, 4(3), 100169.
3. Rawat, P., Bajaj, M., Mehta, S., Sharma, V., & Vats, S. (2023, March). A study on cervical cancer prediction using various machine learning approaches. In *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)* (pp. 1101-1107). IEEE.
4. Srivastav, S., Guleria, K., & Sharma, S. (2023, March). Predictive Machine Learning Approaches for Cervical Cancer Detection: An Analytical Comparison. In *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 951-956). IEEE.
5. Devi, S., & Gaikwad, S. R. (2023). Prediction and detection of cervical malignancy using machine learning models. *Asian Pacific journal of cancer prevention: APJCP*, 24(4), 1419.

6. Su, J., Lu, H., Zhang, R., Cui, N., Chen, C., Si, Q., & Song, B. (2025). Cervical cancer prediction using machine learning models based on routine blood analysis. *Scientific Reports*, 15(1), 22655.
7. Lu, J., Song, E., Ghoneim, A., & Alrashoud, M. (2020). Machine learning for assisting cervical cancer diagnosis: An ensemble approach. *Future Generation Computer Systems*, 106, 199-205.
8. Uddin, K. M. M., Al Mamun, A., Chakrabarti, A., Mostafiz, R., & Dey, S. K. (2024). An ensemble machine learning-based approach to predict cervical cancer using hybrid feature selection. *Neuroscience Informatics*, 4(3), 100169.
9. Ijaz, M. F., Attique, M., & Son, Y. (2020). Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. *Sensors*, 20(10), 2809.
10. Mahadik, A., Sharma, P., & Narawade, V. (2023, October). Breast cancer prediction with gradient boost and xgboost. In *International Conference on Computer & Communication Technologies* (pp. 19-37). Singapore: Springer Nature Singapore.
11. Mahadik, V. N. A., & Sharma, P. (2023). Comparison based on examination of machine learning techniques for minimising the amount of features utilised in breast cancer forecast. *European Chemical Bulletin*, 12(Special Issue 4), 20623–20627.
12. Hanzala, A., Akter, T., & Rahman, M. S. (2025). A hybrid approach for cervical cancer detection: Combining D-CNN, transfer learning, and ensemble models. *Array*, 27, 100434. doi.org
13. Schiffman, M., Castle, P. E., Jeronimo, J., Rodriguez, A. C., & Wacholder, S. (2007). Human papillomavirus and cervical cancer. *The lancet*, 370(9590), 890-907.
14. Chen, M., Shi, X., Zhang, Y., Wu, D., & Guizani, M. (2017). Deep feature learning for medical image analysis with convolutional autoencoder neural network. *IEEE transactions on big data*, 7(4), 750-758.
15. Siegel, R. L., Kratzer, T. B., Wagle, N. S., Sung, H., & Jemal, A. (2026). *Cancer statistics, 2026*. Ca, 76(1), e70043.
16. Luo, W. (2021). Predicting cervical cancer outcomes: statistics, images, and machine learning. *Frontiers in Artificial Intelligence*, 4, 627369.
17. <https://my.clevelandclinic.org/health/body/23279-cervix>
18. <https://www.cancer.gov/publications/dictionaries/cancerterms/def/pap-smear>
19. <https://gco.iarc.who.int/today/en/fact-sheets-cancers>
20. Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and translational science*, 17(11), e70056.
21. 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.
22. Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1), 20-29.