INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 08 ISSUE: 04 | APRIL - 2024 SJIF RATING: 8.448 ISSN: 2582-3930

Predictive Model to Identify Higher Risk of Cervical Cancer

Mangali Sunil[1]
Final Year Student
Department of Information Technology
Malla Reddy College Of Engineering & Technology
Hyderabad, India.
mangalisunil3232@gmail.com

Neha Bollampally[2] ,
Final Year Student
Department of Information Technology
Malla Reddy College Of Engineering & Technology
Hyderabad, India.
nehabollampally357@gmail.com

Patnam Sakshit[3],
Final Year Student
Department of Information Technology
Malla Reddy College Of Engineering & Technology
Hyderabad, India.
patnamsakshit@gmail.com

Abstract

Cervical cancer is a malignancy that affects the cervix. Cervical cancer typically develops slowly over several years, progressing through precancerous changes in the cells of the cervix before becoming invasive cancer. In the relentless pursuit of enhancing healthcare outcomes, this cervical cancer risk prediction project aims to develop an innovative model that empowers both healthcare professionals and patients with proactive insights. Cervical cancer remains a significant global health challenge, but with the fusion of innovative technology and medical expertise, we strive to create a solution that can identify and communicate risk factors accurately. The main goal of this project is to create a model that, in addition to identifying risk indicators, also converts this information into useful insights. We set out on our journey by carefully undertaking system analysis, design, and implementation, and we coordinate our efforts with the accuracy of medical knowledge and the scalability of machine learning techniques. This documentation embodies our dedication to data privacy and serves as a comprehensive roadmap from conceptualization to deployment. The highest care and responsibility are used when handling patient information to maintain their trust. At the heart of this project lies the ability to provide understandable and accessible risk predictions. The system's output gives an accurate cervical cancer risk prediction and a deeper understanding of their health for proactive measures. Our objective is to provide a tool that not predicts the likelihood of cervical cancer but also instills trust, in users and the medical community through rigorous testing and validation. As we document this research, we envision a future where early detection and prevention serve as strategies, in combating cancer resulting in improved quality of life and medical outcomes.

I. INTRODUCTION

Cervical cancer is a prevalent malignancy that affects women all over the world. It is ranked as the fourth most common cause of death in women and typically remains asymptomatic during its initial stages, with the cancerous cells in the cervix progressing slowly. About 90% of the new cases and deaths worldwide in 2020 occurred in poor nations. It significantly affects the burden of cancer across all cultures and economics. Annually in India, 123,907 new instances of cervical cancer are identified, with 77,348 cases resulting in fatalities, contributing to

approximately 25% of the worldwide death toll is attributed to cervical cancer. The symptoms are contingent upon the tumor's size and the stage of the disease. Changes in the cervix are typically serendipitously detected during routine annual check-ups. In advanced stages, symptoms manifest in approximately 90% of cases. Cervical cancer is primarily caused by certain strains of the human papillomavirus (HPV), a sexually transmitted infection. Two strains of the human papillomavirus (HPV), specifically types 16 and 18, are accountable for nearly half of high-grade cervical pre-cancers. Women who have human immunodeficiency virus (HIV) face a six-fold higher risk of developing cervical cancer in comparison to those who do not have HIV. Other risk factors include smoking, a weakened immune system, multiple sexual partners, early sexual activity, and a family history of cervical cancer. These risk factors can be quantified and employed as features in machine learning models. The proposed methodology uses such features to train and test machine models with the help of various supervised machine-learning algorithms. Hinselmann, Schiller, Cytology, and Biopsy are the four target variables. The Hinselmann test, also known as

colposcopy, is a medical procedure utilized for the early detection of cervical abnormalities and cervical cancer. During this procedure, a mild acetic acid solution is applied to the cervix, which helps highlight any abnormal cells, making them more visible. Schiller's Iodine test, also known as Schiller's test, is a diagnostic medical procedure that involves the application of an iodine solution to the cervix to detect cervical cancer. Cytology, particularly in the form of the Pap smear or Pap test, is an important component of cervical cancer screening. During this procedure, a sample of cervical cells is collected and examined under a microscope to detect any abnormalities. A biopsy is a vital diagnostic procedure when it comes to cervical cancer. A biopsy involves collecting a sample of tissue from the cervix and examining it under a microscope to identify any cancerous cells. This procedure plays a role in confirming the presence of cancer determining its stage and informing treatment decisions.

II.LITERATURE REVIEW

1. Cervical Cancer Risk Prediction with Robust Ensemble and Explainable Black Boxes Method (Springer Link, May 2021):
The effects acquired from the test performed on this painting are twofold: firstly, it turned viable to construct an ensemble-kind classifier that reaches an accuracy of 94.5%. Using advanced tools of Explainable Machine Learning it was possible to obtain information on the black box models used and interpret them from the POV of clinical observation and also interpret the features understanding the relationships between them and the target variable, which is the presence or absence of cervical cancer.

2. Prediction and Detection of Cervical Malignancy Using Machine Learning Models (Asian Pacific Journal of Cancer Prevention, Vol 24, April 2023): The study demonstrates that machine learning algorithms can enhance cervical cancer predictions, with Logistic Regression and Decision Tree performing exceptionally well among six models. This research showcases accurate predictions of sensitivity and accuracy, even with imbalanced input and output criteria in the dataset.

3. Design and Development of an Efficient Risk Prediction Model for Cervical Cancer (IEEE Access, Vol 11, July 2023):
The proposed model provides results that can help in recognizing women who are at a higher risk of developing cervical cancer enabling healthcare providers to offer early screening and detection such as Pap smear and HPV testing. This can help to catch the disease in its early stages; when it is most treatable.

4. Data-Driven Cervical Cancer Prediction Model with Outlier Detection and Over-Sampling Methods (MDPI, May 2020):
In this study, a CCPM with feature extraction using Chi-square was proposed. They extracted ten features and used them in their study. The dataset is unbalanced. For missing values, they used the mean equation. The work proposed CCPM by joining DBSCAN and iForest for outlier detection, with SMOTE and SMOTETomek for class balancing and RF as a classifier. The CCPM can assist users in the early detection of cervical cancer risk.

5. A Machine Learning-Based Framework for the Prediction of Cervical Cancer Risk in Women (MDPI, September 2022):
4 To assess the accuracy of risks utilizing predictive models based on precision, recall, F1- score, and support, the authors profiled data and carried out extensive benchmarking. The dataset was chosen specifically to evaluate attributes such as smoking, STDs, STDs, and AIDS, which are the major risk factors of cervical cancer.

III . METHODOLOGY

Data collection

Collect necessary patient data such as age, medical records, patient history, and HPV status.

A. Data Preprocessing:

Data preprocessing is split into 3 important sections: information cleaning, information transformation, and information reduction. The importance of data preprocessing cannot be overstated as it directly affects the success of a project. Data impurity arises while attributes or characteristic values include noise, outliers, or redundant and lacking data [30]. In this dataset, we have got efficiently removed lacking values and outliers. The data transformation stage is crucial as it involves converting the data into suitable forms for the mining process. This research incorporates various techniques such as normalization, attribute selection, discretization, and concept hierarchy generation. Dealing with a large volume of data becomes more challenging when the data dimension is high. To address this, the research employs a data reduction approach to enhance storage efficiency and reduce the cost of data storage and processing. To mitigate overfitting in machine learning models, we have utilized the principal component analysis (PCA) technique as a dimension reduction technique.

Feature Selection:

Identify relevant features that contribute to prediction.

Model Selection: Choose suitable device mastering algorithms like:

Decision Trees:

Decision Trees are an easy-to-use but effective machine-learning technique that is used for both classification and regression problems. It recursively splits the dataset into subsets based on the most significant attribute at each node, aiming to maximize information gain (or minimize impurity in classification tasks). Each internal node in a decision tree represents a decision rule, and each leaf node represents the outcome or a class label. The process of constructing a decision tree involves selecting the best attribute to split the data into pure or homogeneous subsets. Decision timber is easy to apprehend and interpret, making it appropriate for explaining version decisions. However, they can be prone to overfitting, where the tree captures noise in the data, and they may not generalize well to new data.

Random Forest:

Random Forest is an ensemble gaining knowledge of technique that builds a couple of choice bushes and combines their predictions to enhance accuracy and decrease overfitting. It's based on the idea of bagging (Bootstrap Aggregating).

XGBoost:

XGBoost is an optimized and highly efficient gradient-boosting algorithm that has gained popularity for its performance and effectiveness in machine learning competitions. Gradient boosting is an ensemble method that builds decision trees sequentially, where each tree corrects the errors made by the previous ones. XGBoost optimizes this process using techniques like regularization, parallel processing, and a careful design of the objective function. XGBoost can be used for both classification and regression tasks and often outperforms other algorithms due to its ability to handle complex relationships in the data and manage imbalanced datasets effectively. It provides tools for early stopping to prevent overfitting and has a built-in capability for handling missing values in the data. XGBoost is highly customizable, with various hyperparameters to fine-tune for specific tasks.

Splitting the dataset into education and checking out sets:

The dataset is then cut up into parts: a training set and a testing set. The training set is used to train the machine learning model, and the testing set is used to evaluate the performance of the trained model.

MODEL TRAINING AND EVALUATION:

- Train the model
- Tune Hyperparameters
- Evaluate the model based on Accuracy, Precision, Recall, F1 Score.

Deployment:

Deploy the trained model

Model Evaluation:

Evaluate the model's overall performance by the usage of accuracy, precision, recall, f1 score, and confusion matrix. The confusion matrix table shows model performance by comparing actual values of the data with predicted values. It gives a more detailed picture of how well the model is performing by showing the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). True positives (TP) are the cases where the model correctly predicted a positive outcome (the student dropped out) when the case was positive (the student dropped out of college).

True negatives (TN) are the cases where the model correctly predicted a negative outcome (the student was not dropped out) when the actual case was negative (the student was not dropped out of college).

False positives (FP) are the instances wherein the version expected a fine outcome (the scholar dropped out) whilst the real case becomes negative.

False negatives (FN) are the cases where the model predicted a negative outcome (the student did not drop out) when the actual case was positive.

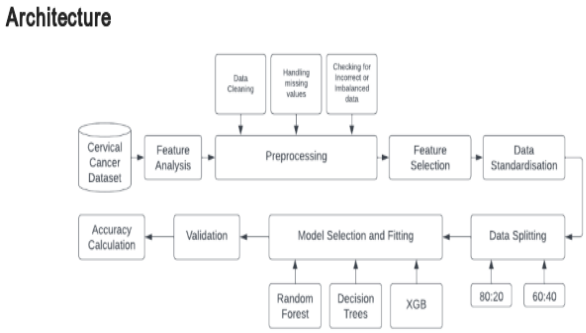

Fig 1: Flow chart of Methodology

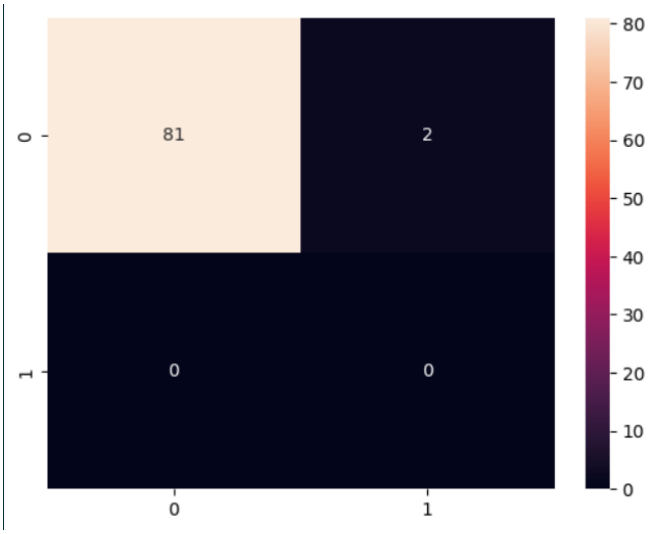

Fig2: Confusion Matrix for Random Forest

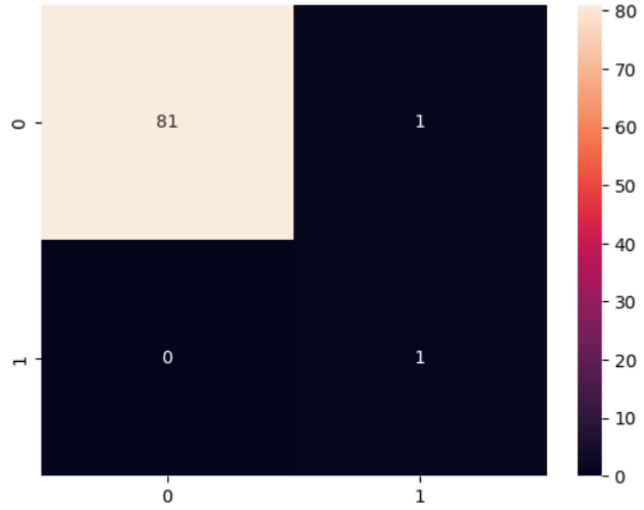

Fig 3:Confusion Matrix for decision tree

IV.RESULT AND ANALYSIS

Target Variable: Hinselmann

Models	Accuracy	Precision		Recall		F1 Score		Support	
		0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0
Random Forest	0.9759	0.98	0.00	1.00	0.00	0.99	0.00	81	2
Decision Trees	0.9879	0.99	1.00	1.00	0.50	0.99	0.67	81	2
XGBoost	1.0000	1.00	1.00	1.00	1.00	1.00	1.00	81	2

Table 1: Hinselmann as the target variable

Target Variable: Schiller

Models	Accuracy	Precision		Recall		F1 Score		Support	
		0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0
Random Forest	0.9759	0.99	0.83	0.99	0.83	0.99	0.83	77	6
Decision Trees	0.9879	0.99	1.00	1.00	0.83	0.99	0.91	77	6
XGBoost	1.0000	1.00	1.00	1.00	1.00	1.00	1.00	77	6

Table 2: Schiller as the target variable

Target Variable: Cytology

Models	Accuracy	Precision		Recall		F1 Score		Support	
		0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0
Random Forest	0.9638	0.96	0.00	1.00	0.00	0.98	0.00	80	3
Decision Trees	0.9518	0.96	0.00	0.99	0.00	0.98	0.00	80	3
XGBoost	0.9638	0.96	0.00	1.00	0.00	0.98	0.00	80	3

Table 3: Cytology as target variable

Target Variable: Biopsy

Models	Accuracy	Precision		Recall		F1 Score		Support	
		0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0
Random Forest	0.9879	0.99	1.00	1.00	0.75	0.99	0.86	79	4
Decision Trees	0.9638	0.99	0.60	0.97	0.75	0.98	0.67	79	4
XGBoost	0.9759	0.98	1.00	1.00	0.50	0.99	0.67	79	4

Table 4:Biopsy as the target variable

Target Variable	Random Forest	Decision Trees	XGBoost
Hinselmann	97.59%	98.79%	100%
Schiller	97.59%	98.79%	100%
Cytology	96.38%	95.18%	96.38%
Biopsy	98.79%	96.38%	97.59%

Table 5: Accuracy comparison table

V.Conclusion

This study focuses on categorizing research based on the prediction of cervical cancer risks. We conducted data profiling and comprehensive benchmarking to assess risk prediction performance using precision, recall, F1-score, and support metrics. Our proposed model was developed using the Python programming language along with the previously mentioned packages and libraries. The cervical cancer dataset from UCI with 36 features and 835 instances was utilized for initial data analysis, followed by data standardization and visualization. Subsequently, the model was trained to accurately predict cervical cancer, and we assessed both its accuracy and performance. The dataset was specifically chosen to evaluate attributes such as smoking, STDs, AIDS, and first sexual intercourse, which are significant risk factors for cervical cancer. The computational results revealed that, among the four target variables (Hinselmann, Schiller, Cytology, and Biopsy), Hinselmann and Schiller exhibited the highest accuracy, demonstrating the effectiveness of our proposed model in handling the cervical cancer dataset. In the future, this model can be upgraded for real-time risk assessment, personalized risk assessment, and integration of multi-modal and early detection technologies.

VI.REFERENCES

1. Cervical Cancer: World Health Organization, https://www.who.int/news-room/fact-sheets/detail/cervical cancer

2. Glučina, M.; Lorencin, A.; Anđelić, N.; Lorencin, I. Cervical Cancer Diagnostics Using Machine Learning Algorithms and Class Balancing Techniques.Appl. Sci. 2023, 13, 1061. https://doi.org/10.3390/app13021061

3. Curia, F. Cervical cancer risk prediction with robust ensemble and explainable black boxes method. Health Technol. 11, 875–885 (2021). https://doi.org/10.1007/s12553-021-00554-6

4. Devi, S., Gaikwad, S. R., & R, H. (2023). Prediction and Detection of Cervical Malignancy Using Machine Learning Models. Asian Pacific journal of cancer prevention. APJCP, 24(4), 1419–1433. https://doi.org/10.31557/APJCP.2023.24.4.1419

5. R. Hariprasad, T. M. Navamani, T. R. Rote and I. Chauhan, "Design and Development of an Efficient Risk Prediction Model for Cervical Cancer," in IEEE Access, vol. 11, pp. 74290-74300, 2023, doi: 10.1109/ACCESS.2023.3296456.

6. Ijaz, M.F.; Attique, M.; Son, Y Data-Driven Cervical Cancer Prediction Model with Outlier Detection and Over-Sampling Methods. Sensors 2020, 20, 2809. https://doi.org/10.3390/s20102809

7. Kaushik, K.; Bhardwaj, A.; Bharany, S.; Alsharabi, N.; Rehman, A.U.; Eldin, E.T.; Ghamry, N.A. A Machine Learning-Based Framework for the Prediction of Cervical Cancer Risk in Women. Sustainability 2022, 14, 11947. https://doi.org/10.3390/su141911947

8. NumPy Documentation: https://numpy.org/doc/ 59

9. Pandas Documentation: https://pandas.pydata.org/docs/

10.MatplotlibDocumentation: https://matplotlib.org/stable/index.html

11. Seaborn Documentation: https://seaborn.pydata.org/

12. Scikit-learn Documentation: https://scikit-learn.org/stable/