10.48047/jocaaa.2024.33.05.33

MACHINE LEARNING APPROACH FOR CERVICAL CANCER PREDICTION

C. Jayasundari and P. Arumugam
Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu
cjayasundari.krmmc@gmail.com; sixfacemsu@gmail.com

ABSTRACT

Cervical cancer remains a major health concern worldwide, especially in developing countries where screening programs may be less accessible. Early detection through accurate predictive models can significantly reduce mortality rates associated with this disease. The dataset features 36 variables, such as age, number of sexual partners, smoking habits, contraceptive use, and sexually transmitted diseases. Diagnostic variables include Hinselmann, Schiller, Cytology, and Biopsy results, with biopsy chosen as the target variable for its definitive diagnostic value. This work compares the performance of four machine learning models: Logistic Regression, Random Forest Classifier, Support Vector Machine (SVM), and XGBoost, in accurately predicting cervical cancer indicators. The findings suggest that XGBoost excels compared to other models in both accuracy and ROC AUC, underscoring its promise for early diagnosis in clinical settings. This research underscores the importance of robust model selection to address imbalanced datasets and demonstrates the significant impact that such predictive tools can have in the healthcare industry, potentially aiding in early diagnosis and improved patient outcomes

Keywords: ***Cervical Cancer, Machine Learning, Feature selection, XGBoost, Biopsy, Predictive Modeling, Healthcare.***

1. INTRODUCTION

Cervical cancer, primarily linked to human papillomavirus (HPV) infection, is the third most common cancer among female reproductive organs and the fourth most common cause of cause – related deaths in women worldwide. Despite being largely preventable through early screening, it remains asymptomatic until advanced stages, especially in developing countries where access to screening and medical care is limited. Consequently, the need for effective early detection methods is critical to improve survival rates and reduce the worldwide impact of cervical cancer.

Machine learning has become a formidable tool in healthcare, offering innovative approaches ffor the early detection and diagnosis of diseases. By analyzing massive datasets, ML models can predict the initial phase of various serious illnesses, including cervical cancer, with greater accuracy and sensitivity. In recent years, several studies have explored the application of ML algorithms to cervical cancer screening, leveraging diverse techniques such as logistic regression, support vector machine (SVM), K-nearest neighbors(KNN), and naïve Bayes classifiers.

The incorporation of ML in cervical cancer diagnosis offers promising potential to address the challenges of early detection, especially in resource-limited settings. The application of ML not only supports the development of computational approaches but also enhances existing clinical decision

10.48047/jocaaa.2024.33.05.33

support systems by leveraging electronic health records (EHRs) and other healthcare data. However, the implementation of ML in healthcare also poses risks and challenges, including system privacy, ethical concerns, and the need for interpretability and visualization of ML models.

The primary objective of this exploration is to compare several ML algorithms for prognosticating suggestions of cervical cancer. Using a dataset of 859 womanish cases, this study will assess how well colorful machine learning algorithms work when combined with class balancing strategies. The ideal is to produce an accurate and reliable prophetic model that will help with cervical cancer early opinion and enhance the prognosis.

2. DATA DESCRIPTION

The dataset, which focused on the prediction of markers and diagnosis pertaining to cervical cancer, was taken from the UCI machine learning repository. It includes the behavioural patterns, demographic information, and past medical data of 858 patients who were treated at the Hospitaal Universitario de Caracas in caracas, Venezula. Some participants chose not to share certain information out of privacy concerns, which led to missing data in some places in the dataset. The collection includes variables related to patient age, metrics related to sexual behavior such as the number of partners and the age at which the patient had their first sexual encounter, and reproductive history including the number of pregnancies, smoking habits, use of contraceptives, and the existence of sexually transmitted diseases with specific categories and counts.

3. METHODOLOGY

3.1 Logistic Regression

The Logistic Regression focuses on classifying tumors as either benign (non-cancerous) or malignant (cancerous) using machine learning techniques. The model relies on input features such as tumor size, texture, smoothness, radius, and perimeter to make predictions. The target variable is binary, where 1 indicates a malignant tumor and 0 represents a benign one. Accurate classification is crucial for early diagnosis and effective treatment planning. By analyzing these features, the model can assist healthcare professionals in making more informed decisions.

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}} \quad \ldots (1)$$

Where,

- $P(Y = 1|X)$: Probability that the output y is 1 given the input features X
- β_0: Intercept (bias term)
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$: Coefficients for the input features
- $x_1, x_2, \ldots, x_n$: Input features
- e: Euler's number (approximately 2.718)

3.2 Support Vector Machine

10.48047/jocaaa.2024.33.05.33

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm used for both classification and regression tasks. It works by finding the optimal hyperplane that best separates different classes in the feature space. The goal is to maximize the margin between the nearest data points of each class, known as support vectors. SVM can handle both linear and non-linear classification using kernel functions. Its effectiveness makes it popular in medical diagnosis, including tumor classification (Vladimir Vapnik and Alexey Chervonenkis in 1963).

The decision function is:

$$f(x) = \text{sign}(w.x + b) \quad \ldots (2)$$

Where,

- x is the input feature vector.
- w is the weight vector (defines the orientation of the hyperplane).
- b is the bias term (defines the offset from the origin).
- $sign(.)$ determines the class: +1 or -1.

3.3 Random Forest

Random Forest reduces overfitting by averaging multiple deep decision trees trained on different parts of the dataset. It introduces randomness through bootstrapping (sampling with replacement) and feature selection at each split. This leads to diverse trees, enhancing generalization. It performs well even with missing data and maintains accuracy on large datasets. Random Forest is widely used for classification, regression, and feature importance ranking *Leo Breiman (2001).*

For classification:

$$\hat{y} = \text{mode}(T_1(x), T_2(x), \ldots, T_B(x)) \quad \ldots (3)$$

For regression:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \quad \ldots (4)$$

Where,

$T_b(x)$ is the prediction from the b^{th} decision tree.

B is the total number of trees (boostrap aggregating or bagging is used to train each tree on a different random subset of the data).

3.4 XGBoost

10.48047/jocaaa.2024.33.05.33

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm that builds upon gradient boosting decision trees. Its formulation is deeply rooted in statistical learning theory, especially regularized regression and additive modeling. Below is the core XGBoost objective function, broken down and explained in terms of statistical concepts:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad \ldots (5)$$

Where,

- y_i : true table for instance i
- $\hat{y}_i^{(t-1)}$: prediction from the previous t-1 trees
- f_t : the new tree added at iteration t
- l : differentiable convex loss function
- $\Omega(f_t)$: regularization term

3.5 Chi-Square test for association

- Formulate Hypothesis
  H_0: No association between the variables.
  H_1: Association exists between the variables.
- Create a Contingency table:
  Organize data into a table showing frequencies of category combinations.
- Calculate expected frequencies

$$E_{ij} = \frac{(Row\ Total * Column\ Total)}{Grand\ Total} \quad \ldots (6)$$

- Calculate the chi – square statistics

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \ldots (7)$$

- Determine Degrees of freedom

$$df = (r-1)(c-1) \quad \ldots (8)$$

- Find the critical value and p-value
  Compare χ^2statistics to critical value or find the p-value
- Decision making
  Reject H_0 if χ^2> critical value or p<α

3.6 Model performance

Accuracy

Accuracy measures the proportion of correctly predicted instances out of the total instances in the dataset. It is determined by dividing the number of correct predictions by the total number of predictions made.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad \ldots (9)$$

10.48047/jocaaa.2024.33.05.33

Precision

Precision quantifies the proportion of correctly predicted positive instances among all instances predicted as positive. It emphasizes the accuracy of positive predictions.

$$\text{Precision} = \frac{TP}{TP+FP} \quad \ldots (10)$$

Recall

Recall, also known as sensitivity, measures the proportion of correctly predicted positive instances out of all actual positive instances. It focuses on the ability of the model to capture positive instances.

$$\text{Recall} = \frac{TP}{TP+FN} \quad \ldots (11)$$

Where, TP = True Positive, FP = False Positive, TN =True Negative, FN = False Negative.

F1 Score

The F1 score is the harmonic mean of precision and recall, offering a single metric that balances both aspects. It is especially valuable when dealing with imbalances between positive and negative instances.

$$\text{F1 Score} = 2 \times \frac{Precision * Recall}{Precision * Recall} \quad \ldots (12)$$

4. RESULTS

The number of older participants is too limited to provide significant conclusions. Additionally, the target variable, which represents cervical cancer diagnosis, is heavily skewed, with 803 instances of class 0(no cancer) and 55 instances of class 1 (cancer).

10.48047/jocaaa.2024.33.05.33

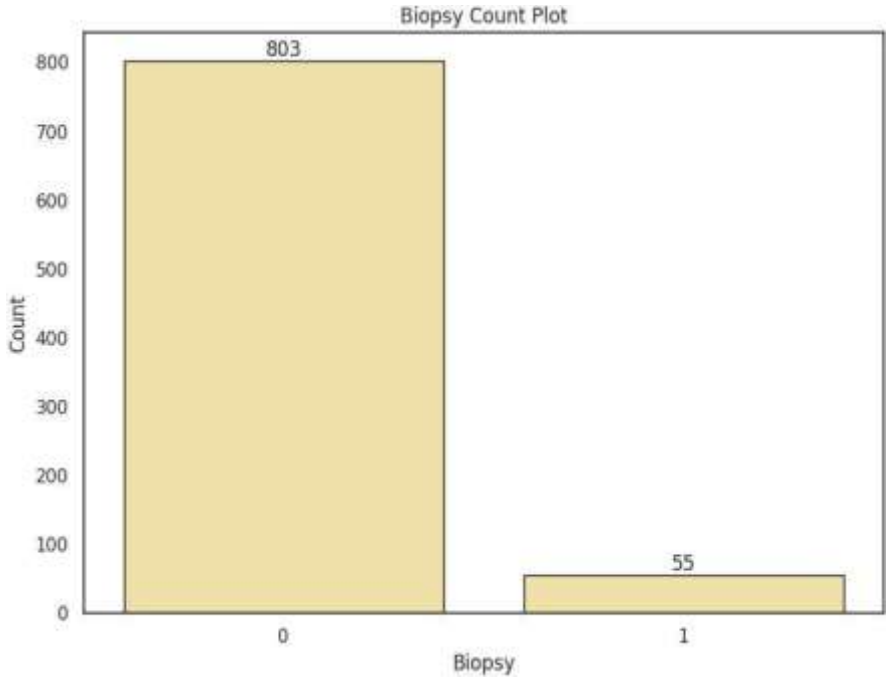
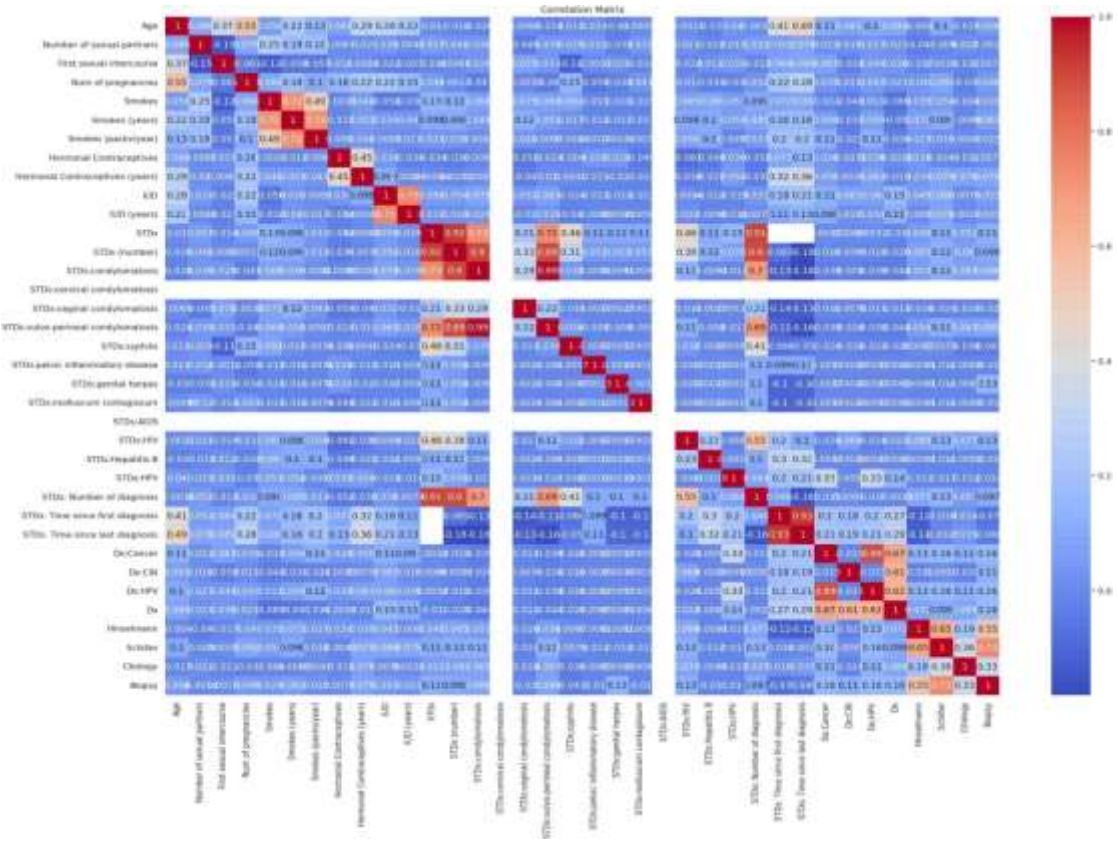


Figure 1. Bar diagram of Cancer vs non Cancer

4.1 Correlation Analysis



10.48047/jocaaa.2024.33.05.33

Figure 2. Correlation matrix Heatmap for the features in cervical cancer

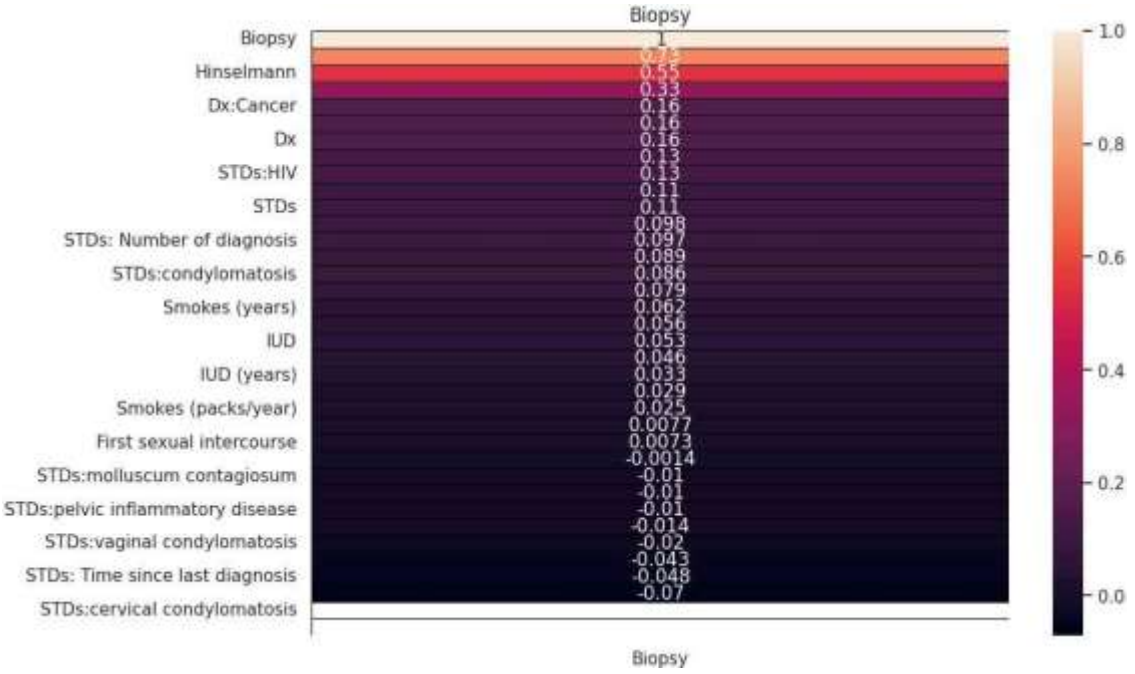


Figure 3. Correlation of individual features with the biopsy outcome

Strong correlations were found between biopsy and other screening tests: Schiller (0.733), Hinselmann (0.547), and Cytology (0.327).

Table 1. Chi-square test results for categorical features and biopsy

Feature	**Chi-square statistic**	**Degrees of freedom**	**P-value**	**Association exist**
Smokers	0.4276	1	0.5131	False
Harmonal Contraceptives	0.0043	1	0.9472	False
IUD	1.4883	1	0.2224	False
STDs	7.6249	1	0.0057	True
STDs: condylomatosis	4.2721	1	0.0387	True
STDs: cervical condylomatosis	0.0	0	1.0	False
STDs: vaginal condylomatosis	0.0	1	1.0	False
STDs: vulvo-perineal condylomatosis	4.5478	1	0.0329	True
STDs: syphills	0.5116	1	0.4744	False

10.48047/jocaaa.2024.33.05.33

STDs: pelvic inflammatory disease	0.0	1	1.0	False
STDs: genital herpes	2.8245	1	0.0928	False
STDs: molluscum contagiosum	0.0	1	1.0	False
STDs: AIDS	0.0	0	1.0	False
STDs: HIV	9.0924	1	0.0025	True
STDs: Hepatitis	0.0	1	1.0	False
STDs: HPV	0.0	1	1.0	False
Dx: Cancer	17.866	1	2.3694	True
Dx: CIN	6.9219	1	0.0085	True
Dx: HPV	17.866	1	2.3694	True
Dx	17.588	1	2.7418	True
Hinselmann	245.93	1	1.9951	True
Schiller	450.64	1	5.2107	True

Significant associations were found between biopsy occurrence and variables related to general STD presence (STDs), specific types of STDs (STDs: condylomatosis, STDs: vulvo-perineal condylomatosis, STDs: HIV), as well as diagnostic outcomes (Dx: Cancer, Dx: CIN, Dx: HPV, Dx), and screening tests (Hinselmann, Schiller, Citology).

4.2 Evaluation Metrics

Table 2. Performance Evaluation Metrics of Different ML Models

	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression	0.959302	0.642857	0.818182	0.720000	0.901750
Random Forest	0.959302	0.750000	0.545455	0.631579	0.977414
SVM	0.848837	0.142857	0.272727	0.187500	0.792772
XGBoost	0.965116	0.727273	0.727273	0.727273	0.966685

- Logistic Regression: High recall (0.818) and good accuracy (0.959), suitable for identifying positive cases.
- Random Forest: Good precision (0.750) and accuracy (0.959), but lower recall (0.545).
- Support Vector Machine: Lowest performance across all metrics, indicating it less suitable for this dataset.
- XGBoost: Best overall performance with the highest accuracy (0.965), balanced F1 score (0.727), and highest ROC AUC(0.967).

10.48047/jocaaa.2024.33.05.33

4.3 Featrues Importance

Features related to screening tests (Hinselmann, Schiller, Citology) and diagnostic outcomes (Dx: CIN, Dx: HPV, Dx) were consistently identified as important by both Random Forest and XGBoost models. Lifestyle and sexual behavior factors, such as age of first sexual intercourse, count of sexual partners, and smoking habits, also played significant roles in assessing the risk of cervical cancer.

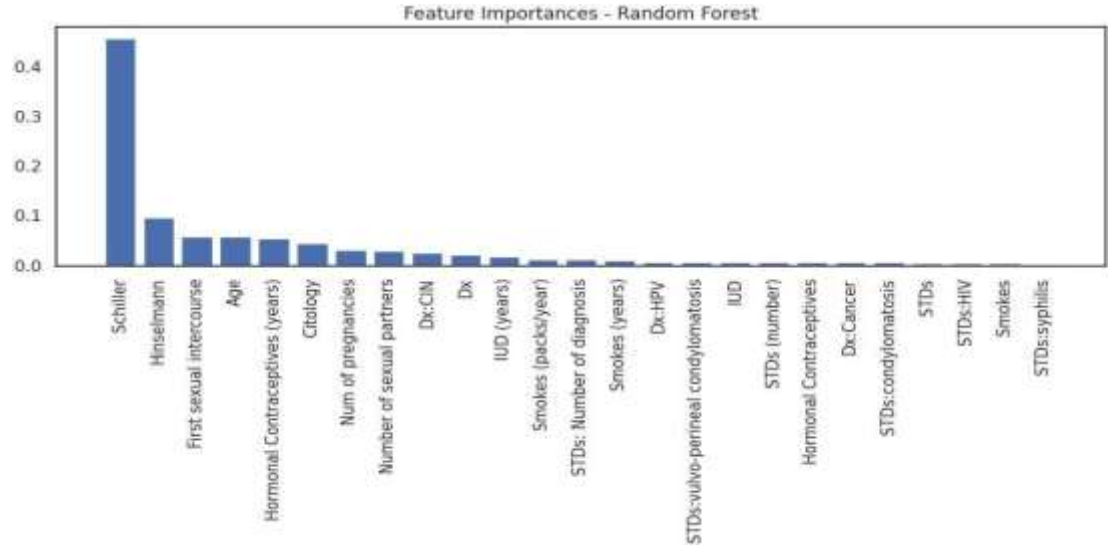


Figure 4. Feature importance scores from the Random Forest model

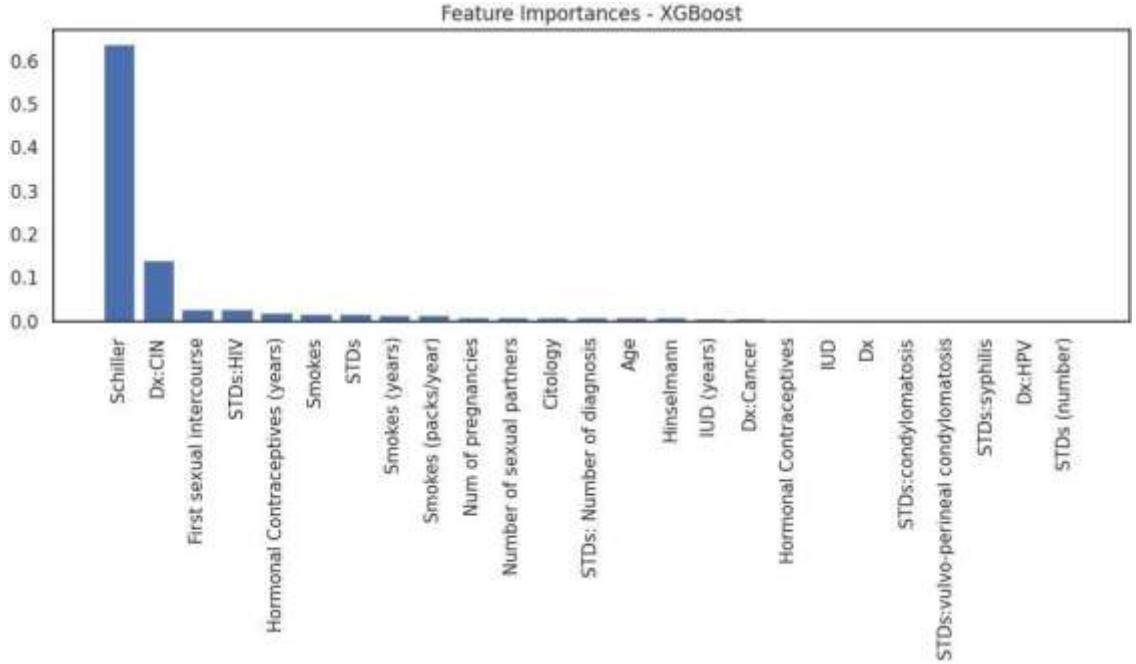


Figure 5. Feature importance scores from the XGBoost model

10.48047/jocaaa.2024.33.05.33

Given its high accuracy, balanced F1 score, and superior ROC AUC, XGBoost is recommended as the most effective model for evaluating the risk of cervical cancer in this dataset. Logistic Regression, with its high recall, is also suitable for applications where identifying positive cases is crucial.

5. CONCLUSION

Cervical cancer continues to be an important global health concern, particularly in low-resource environments where screening access is constrained. This work showed that machine learning (ML) models, especially XGBoost, can successfully predict cervical cancer by utilizing clinical and demographic information. Among the various models evaluated, XGBoost outperformed the others thanks to its capability to manage missing data, implement regularization, and recognize complex patterns. The main risk factors uncovered included age, smoking habits, number of pregnancies, and contraceptive usage, which correspond with established epidemiological patterns. Incorporating ML models into clinical decision-support systems could improve early detection and the allocation of resources, particularly in underprivileged regions. Although the results are promising, the study recognized limitations such as a small dataset and the necessity for external validation. Future work should involve deep learning, larger datasets, and multi-modal data integration to improve accuracy. This study supports the growing role of AI in preventive healthcare and highlights its potential to reduce cervical cancer mortality through early intervention and informed screening strategies.

References

1. Akinyelu, A. A., & Blount, M. (2020). Machine learning approaches to breast cancer detection: A systematic review. *Heliyon*, *6*(11), e04983. https://doi.org/10.1016/j.heliyon.2020.e04983
2. Al-Mousa, D. S., & Alakhras, M. (2020). Cervical cancer detection and classification using machine learning techniques: A systematic review. *Applied Sciences*, *10*(18), 6474. https://doi.org/10.3390/app10186474
3. American Cancer Society. (2022). *Cervical cancer*. https://www.cancer.org/cancer/cervical-cancer.html
4. Arbyn, M., Weiderpass, E., Bruni, L., de Sanjosé, S., Saraiya, M., Ferlay, J., & Bray, F. (2020). Estimates of incidence and mortality of cervical cancer in 2018: A worldwide analysis. *The Lancet Global Health*, *8*(2), e191–e203. https://doi.org/10.1016/S2214-109X(19)30482-6
5. Bhardwaj, A., Sharma, A., Singh, P., & Mittal, M. (2020). Predictive analysis for cervical cancer using hybrid ensemble model. *Materials Today: Proceedings*, *33*, 3435–3438. https://doi.org/10.1016/j.matpr.2020.06.551
6. Chorley, A. J., Marlow, L. A. V., Forster, A. S., Haddrell, J., & Waller, J. (2017). Experiences of cervical screening and barriers to participation in the context of an organized programme: A systematic review and thematic synthesis. *Psycho-Oncology*, *26*(2), 161–172. https://doi.org/10.1002/pon.4126

10.48047/jocaaa.2024.33.05.33

7. Das, D., & Turkoglu, M. (2021). Cervical cancer classification using hybrid machine learning techniques. *Computer Methods and Programs in Biomedicine*, *200*, 105880. https://doi.org/10.1016/j.cmpb.2021.105880
8. Denny, L. (2021). Cervical cancer: Prevention and treatment. *Discovery Medicine*, *31*(167), 1–6.
9. Dhivya, R., & Vasanth, S. (2022). Cervical cancer diagnosis using XGBoost classifier. *ICT Express*, *8*(1), 121–126. https://doi.org/10.1016/j.icte.2021.05.004
10. Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., & Bray, F. (2020). *Global cancer observatory: Cancer today*. International Agency for Research on Cancer. https://gco.iarc.fr/today
11. Ghosh, S., & Ghosh, S. (2021). Application of machine learning in early detection of cervical cancer. *Procedia Computer Science*, *185*, 59–66. https://doi.org/10.1016/j.procs.2021.05.007
12. Gonçalves, A. M., & Diniz, A. J. (2019). Comparative analysis of machine learning algorithms for cervical cancer classification. *Informatics in Medicine Unlocked*, *17*, 100244. https://doi.org/10.1016/j.imu.2019.100244
13. Kaur, T., & Sharma, A. (2020). Predictive modeling for cervical cancer diagnosis using hybrid machine learning approach. *Materials Today: Proceedings*, *28*, 321–325. https://doi.org/10.1016/j.matpr.2020.05.457
14. Li, Y., & Li, J. (2019). Hybrid model for cervical cancer detection using XGBoost and SMOTE. *BMC Medical Informatics and Decision Making*, *19*(1), 1–9. https://doi.org/10.1186/s12911-019-0827-1
15. Sharma, A., & Dey, N. (2019). Cervical cancer detection using hybrid machine learning techniques: A review. *Computer Methods and Programs in Biomedicine*, *174*, 95–107. https://doi.org/10.1016/j.cmpb.2019.03.020