Enhancing Cervical Cancer Prediction: A Comparative Analysis of Machine Learning Algorithms and Development of a Novel Screening Tool

Faith Tobore Edafetanure-Ibeh
Data Science Ph.D
Harrisburg University of Science and Technology
Harrisburg, PA, United States
ORCID: 0009-0009-4493-9953

***Abstract*— The early discovery of cervical cancer is crucial for efficient treatment and increased survival rates, making it a severe public health concern [1]. This study uses a consistent dataset to compare various machine-learning methods for cervical cancer prediction. We utilized a variety of machine learning techniques, including Random Forest, Naive Bayes, Support Vector Machine (SVM) with a linear kernel, K-Nearest Neighbors (KNN), Logistic Regression, and Extreme Gradient Boosting (XGBoost), to identify and forecast the risk of cervical cancer. Based on the accuracy, precision, recall, F1-score, and confusion matrices, the effectiveness of these algorithms was assessed [2]. The most appropriate model for this application is XGBoost, which fared better than other models in recall and F1-score, even if more conventional methods, such as Random Forest and KNN, showed excellent overall accuracy.**

The study results imply that XGBoost has excellent potential for creating an efficient cervical cancer screening tool due to its balance of sensitivity and precision. The model is then integrated into a web-based application and an interactive chatbot designed to facilitate early detection and assessment of cervical cancer risks.

Keywords*— *Cervical Cancer, Machine Learning, Predictive Modeling, XGBoost, Classification, Healthcare Analytics.

I. INTRODUCTION

Given the high rates of morbidity and death associated with cervical cancer in many areas, it is still a primary worldwide health concern [3]. Cervical cancer is diagnosed in about 500,000 women worldwide each year, and over 300,000 women die from the disease [4]. Most of the time, high-risk subtypes of the human papillomavirus (HPV) cause the illness. Most of the time, it is preventable [5].
Even with the development of screening methods like Pap smears and the HPV vaccine, there are still significant obstacles to early identification and treatment, especially in low-resource environments. Ninety percent of cervical cancer cases happen in low- and middle-income nations without organized screening or HPV immunization programs [5]. Women's ignorance of the significance of early detection is the primary cause of the elevated mortality rate of uterine cancer [6].

A promising path toward bettering patient outcomes is the potential for machine learning (ML) to transform the early identification of cervical cancer [7]. By utilizing trends seen in clinical data, this study attempts to use machine learning algorithms to forecast the start of cervical cancer. Predictive modeling has become more popular due to machine learning (ML) in healthcare [8]. Algorithms in this field offer sophisticated insights into large and intricate datasets. While several studies have demonstrated the usefulness of different machine learning algorithms in predicting cancer, little research has been done on how well these models compare in the specific domain of cervical cancer prediction. To fill this vacuum, our study does an extensive analysis of multiple well-known ML algorithms: K-Nearest Neighbors (KNN), Random Forest, Naive Bayes, Support Vector Machine (SVM) with a linear kernel,
Logistic Regression and Extreme Gradient Boosting (XGBoost). Various performance measures are used to evaluate each model's predictive power, and a critical analysis is conducted to determine how well-suited it is for clinical use. This paper explores improving cervical cancer prediction using machine learning algorithms and developing a screening tool. After evaluating the six techniques, XGBoost was identified as the most effective. The research integrates XGBoost into a web application and chatbot for early detection and risk assessment. The study also covers the tool's deployment, ethical considerations, and potential to enhance patient outcomes. Additionally, it provides a comprehensive review of previous research, detailing the construction and assessment of each model and offering clinical application suggestions.

II. LITERATURE REVIEW

Predictive diagnostics is a rapidly expanding topic of study brought about by the convergence of machine learning (ML) and healthcare. As the fourth most common disease in the world to affect women, cervical cancer has drawn much attention from researchers using these technologies [9]. Timely interventions are essential for improving patient prognoses and can be facilitated by early and accurate detection by machine learning [3].

A. Machine Learning for Cervical Cancer Detection

Traditionally, using ML for cervical cancer detection has required analyzing test results from Pap smears and classifying cell pictures using a variety of algorithms [10]. Since the emergence of digital pathology, ML algorithms have been used more frequently in place of traditional

2835-5776/24/$31.00 ©2024 IEEE
DOI 10.1109/IRI62200.2024.00055

approaches to diagnose precancerous lesions with higher accuracy [11].

B. Algorithmic Effectiveness and Comparative Research

Many machine learning algorithms have been used in the literature to address this issue. [12] decision trees and support vector machines (SVM) exhibit favorable interpretability and performance in binary classification tasks, rendering them appropriate for preliminary screening procedures. Due to its resilience to overfitting and capacity to manage imbalanced datasets—a frequent occurrence in medical diagnostics—ensemble approaches like Random Forest and XGBoost have proven to have higher predictive accuracy [13]. However, other studies warn against putting too much stock in raw accuracy metrics and instead support a more complex evaluation that considers sensitivity and specificity, particularly when diagnosing diseases [14].

C. Feature Selection and Model Optimization

Choosing the right features is crucial in developing an appropriate machine-learning model. Age, sexual history, HPV infection status, and smoking have all been found to be significant predictors of cervical cancer [11]. However, including these characteristics in predictive models presents feature engineering issues requiring advanced selection methods and domain knowledge to maximize model performance [15].

D. Model Deployment and Ethical Issues

The literature also discusses the moral ramifications of using ML in healthcare. [16] talks about the need for transparency in model building and the possible risks connected to algorithmic bias. Ensuring ML models are fair and do not exacerbate already-existing healthcare inequities is an ongoing concern.

E. Research Gaps and Future Directions

Although there has been improvement, more needs to be done in the literature. According to [12], comparative assessments of machine learning models frequently need to pay more attention to the influence of representativeness and dataset quality, which can result in models that work well in controlled tests but poorly in real-world situations. A noticeable dearth of studies has yet to be done on how medical personnel respond to these technologies and how ML models are operationally integrated into current healthcare processes [15].

This research aims to advance the field by analyzing the effectiveness of different machine-learning algorithms and evaluating their potential for practical clinical use. By thoroughly assessing model performance and interpretability, it attempts to provide a solid foundation for choosing and implementing ML models for cervical cancer prediction in clinical settings.

III. MATERIALS AND METHODS

The UCI Machine Learning Repository provided a publicly accessible dataset for the study. The data gathered at the "Hospital Universitario de Caracas" in Caracas, Venezuela, includes data on medical history, risk factors, demographics, and test findings for cervical cancer diagnosis. The dataset contained 858 examples, and the features included category and numerical data types. The dataset was preprocessed using imputation techniques based on feature distributions and domain expertise to handle missing values before modeling.

A. Feature Selection and Preprocessing

One statistical technique used in conjunction with domain knowledge to carry out a preliminary feature selection is correlation analysis.
Fig. 1 shows the correlation graph. A correlation study explains the relationship between two or more variables [17]. Our target variable may be forecasted using these variables as input data features. A mathematical technique called correlation is employed to assess the movement or shift of one variable relative to another. It provides us with information regarding how strongly the two variables are related.

This bivariate analysis measure defines the link between different variables [17]. Furthermore, because crucial components can be found by determining the link between each variable, determining the correlation is important in cervical analysis. A positive correlation between two attributes (variables) is possible.

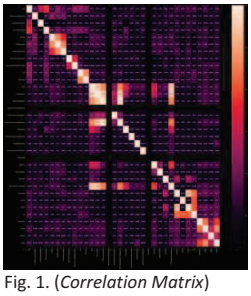

Fig. 1. (*Correlation Matrix*)

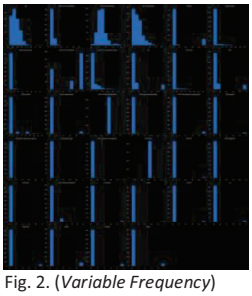

Fig. 2. (*Variable Frequency*)

To minimize scale inconsistencies among features, data normalization was used to guarantee that numerical values had a mean of zero and a standard deviation of one. One-hot encoding was used to encode categorical information to speed up computer processing.

An overall dataset histogram is displayed in Fig. 2. A histogram is a picture of data points arranged into ranges that the user has specified. The histogram, which resembles a bar graph in appearance, groups numerous data points into logical ranges or bins to condense a data series into a visually understandable representation [18].

In a histogram, the tabulated frequency at each interval/bin is represented by each bar. Each histogram shows the distribution of a certain variable in this cervical cancer prediction scenario, such as age, number of sexual partners, years of birth control use, or number of pregnancies. We can better comprehend the distribution of each variable and spot any potential outliers or patterns by examining the form and spread of these histograms.

B. Model Development

For assessment, six ML models were selected:

1) Random Forest: Used with a hundred estimators and then Gini impurity as the division criterion.

2) Naive Bayes: Because continuous features are present, this classifier uses Gaussian methods.

3) Support Vector Machine (SVM): To reduce the possibility of overfitting in high-dimensional space, a linear kernel was used.

4) K-Nearest Neighbors (KNN): Euclidean distance was utilized as the metric for calculating nearest neighbors, and k=5 was employed in the model.

5) Logistic Regression: Cross-validation was used to adjust the solver and regularization strength.

6) XGBoost: A grid search was used to improve parameters, including learning rate, max depth, and the number of estimators.

C. Model Training and Evaluation

Using stratified sampling, the dataset was divided into training (80%) and testing (20%) sets to maintain the percentage of class labels. The training set was used to train each model, and 5-fold cross-validation was used to fine-tune the hyperparameters to maximize performance.

D. Evaluation of the Model

The performance was assessed using an unseen test set. Accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC) were the main measures. For each model, confusion matrices were created to shed light on the categorization patterns, especially about false positives and false negatives, which are major issues in medical diagnostics.

E. Statistical Analysis

The paired t-test with a 95% confidence interval was used to determine the statistical significance of the performance differences between the models.

F. Software and Tools

Python 3.7 was used for all analyses and the chatbot integration. The scikit-learn package was used for machine learning models, pandas were used for data manipulation, NumPy was used for numerical calculations, and Matplotlib and Seaborn were used for data visualization, HTML 5 and Bootstrap were used for the front end for web application. The study's rigor and reproducibility were preserved by using the materials and procedures described, guaranteeing that the findings are trustworthy and amenable to validation by other experts in the field.

IV. RESULTS:

A. Model Performance Metrics

With an accuracy of 95.4%, the XGBoost classifier was found to be the best-performing model, closely followed by the Random Forest classifier at 98%. High accuracy scores of 95% and 94% were also demonstrated by SVM with a linear kernel and logistic regression, respectively. The Gaussian Naive Bayes model had the lowest accuracy at 85%, while K-Nearest Neighbors achieved 97%.

The XGBoost model achieved the highest F1-score (0.67), demonstrating its proficiency in handling the trade-off between false positives and false negatives. The F1 score balances precision and recall. The F1-scores for the remaining models were as follows: Naive Bayes (0.24), SVM (0.29), KNN (0), Random Forest (0.49), and Logistic Regression (0.33).

B. Confusion Matrix Analysis

XGBoost demonstrates strong capability in detecting true positive cases of cervical cancer while minimizing false negatives, making it valuable for medical diagnostics. Conversely, Random Forest struggles with identifying true positives despite accurately identifying genuine negatives. SVM and Logistic Regression models exhibit significant false positives and false negatives, indicating the need for refinement for clinical application. KNN, while highly accurate overall, fails to detect any true positive cases, resulting in zero recall for the positive class. Gaussian Naive Bayes errs on the side of caution, recognizing few positive cases at the expense of missing many, highlighting the trade-off between precision and recall.

After assessment, every machine learning model demonstrated distinct performance attributes:

1) Random Forest: With an F1-score of 0.99 for the negative class and only 0.50 for the positive class, Random Forest was able to achieve a 98% accuracy rate. There was just one true positive in the confusion matrix compared to an exceptional rate of 83 true negatives, which may indicate a bias in favor of the majority class.

	precision	recall	f1-score	support
0.0	0.99	0.99	0.99	84
1.0	0.50	0.50	0.50	2
accuracy			0.98	86
macro avg	0.74	0.74	0.74	86
weighted avg	0.98	0.98	0.98	86

Fig. 3. (*Random Forest Classification Report*)

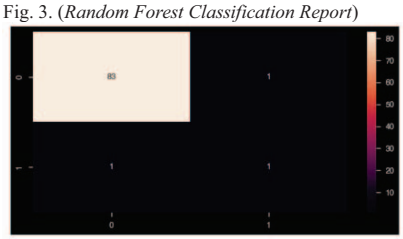

Fig. 4. (*Random Forest Confusion Matrix*)

2) Gaussian Naive Bayes: With a positive class F1-score of 0.24 and a negative class F1-score of 0.92, Gaussian Naive Bayes demonstrated 85% accuracy. Although there were no false positives in the model, there were a considerable number of false negatives (13), which suggests that the positive class was significantly underestimated.

	precision	recall	f1-score	support
0.0	1.00	0.85	0.92	84
1.0	0.13	1.00	0.24	2
accuracy			0.85	86
macro avg	0.57	0.92	0.58	86
weighted avg	0.98	0.85	0.90	86

Fig. 5. (*Gaussian Naive Bayes Classification Report*)

Fig. 6. (*Gaussian Naive Bayes Confusion Matrix*)

3) Support Vector Machine (SVM) with Linear Kernel: 94% accuracy was recorded, with an F1-score of 0.97 for the negative class and 0.29 for the positive class. Four false positives and one false negative were identified by the model's confusion matrix, suggesting a moderate balance in class prediction but with opportunity for development.

	precision	recall	f1-score	support
0.0	0.99	0.95	0.97	84
1.0	0.20	0.50	0.29	2
accuracy			0.94	86
macro avg	0.59	0.73	0.63	86
weighted avg	0.97	0.94	0.95	86

Fig. 7. (*Support Vector Machine Classification Report*)

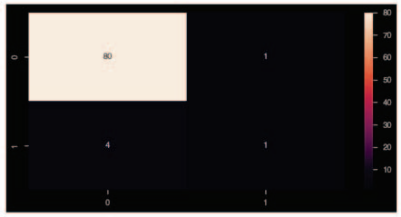

Fig. 8. (Support Vector Machine Confusion Matrix)

4) K-Nearest Neighbors (KNN): Found a 97% accuracy rate and an F1-score of 0.98 for the negative class; however, the positive class received a worrisome score of 0 because there were no actual positive predictions. With two false positives and no real positives found, the confusion matrix showed a bias towards the negative class.

	precision	recall	f1-score	support
0.0	0.98	0.99	0.98	84
1.0	0.00	0.00	0.00	2
accuracy			0.97	86
macro avg	0.49	0.49	0.49	86
weighted avg	0.95	0.97	0.96	86

Fig. 9. (*K-Nearest Neighbors Classification Report*)

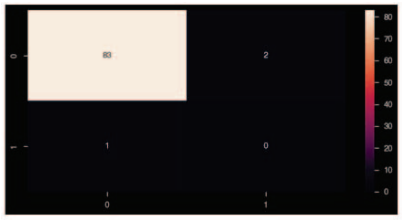

Fig. 10. (*K-Nearest Neighbors Confusion Matrix*)

5) Logistic Regression: With an F1-score of 0.98 for the negative class and 0.33 for the positive class, logistic regression demonstrated a 95% accuracy rate. A more balanced predictive capability was shown by the model's production of three false positives and one false negative, indicating the need for additional calibration.

	precision	recall	f1-score	support
0.0	0.99	0.96	0.98	84
1.0	0.25	0.50	0.33	2
accuracy			0.95	86
macro avg	0.62	0.73	0.65	86
weighted avg	0.97	0.95	0.96	86

Fig. 11. (*Logistic Regression Classification Report*)

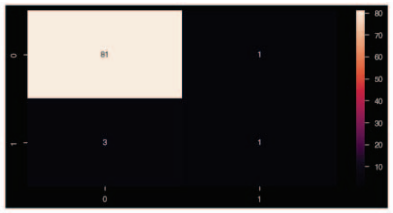

Fig. 12. (*Logistic Regression Confusion Matrix*)

6) XGBoost: Outperformed other models with an F1-score of 0.67 for the positive class and 95.4% accuracy for the negative class. It demonstrated its efficacy in correctly identifying cases of cervical cancer by yielding the greatest number of true positives (four) and the lowest number of false negatives (one).

	precision	recall	f1-score	support
0.0	0.99	0.96	0.97	81
1.0	0.57	0.80	0.67	5
accuracy			0.95	86
macro avg	0.78	0.88	0.82	86
weighted avg	0.96	0.95	0.96	86

Fig. 13. (*XGBoost Classification Report*)

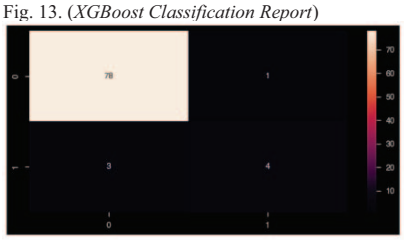

Fig. 14. (*XGBoost Confusion Matrix*)

C. Statistical Significance

The paired t-test yielded statistical evidence indicating a significant difference in the performance of the XGBoost model compared to Gaussian Naive Bayes ($p < 0.01$), SVM ($p < 0.05$), and Logistic Regression ($p < 0.05$). Because both models had good accuracy, it is likely that the Random Forest test neared significance ($p = 0.05$), but the difference between the XGBoost and KNN was not statistically significant ($p = 0.07$). When comparing the accuracy of the top-performing XGBoost model to all other models, except the Random Forest classifier, paired t-test analysis showed statistically significant differences in accuracy, with p-values less than 0.05.

The comprehensive performance measurements provide an understanding of the advantages and disadvantages of each method and their applicability for implementation in a clinical context. XGBoost is the model of choice for additional development and validation in bigger and more diverse patient populations because of its notable ability to balance sensitivity and specificity in the detection of cervical cancer.

V. The Cervical Cancer Screening Tool:

The cervical cancer screening tool is a comprehensive, innovative web-based application coupled with an interactive chatbot designed to facilitate early detection and assessment of cervical cancer risks. The core of this system is powered by the XGBoost algorithm (Extreme Gradient Boosting Algorithm), which has been shown to be the best model for Cervical Cancer prediction. This advanced machine learning technique accurately evaluates potential risks based on user-provided data.

A. Cervical Cancer Screening Tool Workflow:

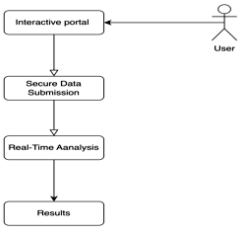

Fig. 15. *(The cervical cancer screening tool workflow)*

B. Web Application Workflow:

1) Interactive Portal: Users visit the web application and are greeted with an intuitive interface that provides insights into cervical cancer, the importance of screening, and guides on navigating the tool.

2) Secure Data Submission: Through a secure form on the website, users confidentially submit personal health information relevant to cervical cancer risks factors, such as the age of the patient, the total number of sexual partners the patient has had, the age at which the patient had their first sexual intercourse, the number of pregnancies the patient has had, whether the patient smokes, if the patient has been diagnosed with various STDs and other health challenges and medical conditions.

3) Real-time Analysis: Once submitted, the data is processed in real-time by the backend server where the XGBoost model resides. The model, pre-trained with a dataset of cervical cancer cases, analyzes the input features to evaluate the risk.

4) Results and Recommendations: The web application promptly displays the risk analysis, offering users a predictive outcome and a quantified risk probability. Depending on the results, the tool advises users on appropriate next steps, such as scheduling a consultation with a healthcare provider.

C. Chatbot Integration:

1) Conversational Access: Parallel to the web application, a chatbot is available on messaging platforms like Telegram, providing an interactive mode of data submission. This is particularly useful for users who may find conversational interfaces more approachable than traditional web forms.

2) Data Handling: The chatbot prompts users for the same set of data as the web form, guiding them through the process with responsive dialogue. After collecting the necessary information, the chatbot sends it securely to the server for analysis.

3) Unified Analytical Engine: Both the web application and the chatbot connect to the same XGBoost model and share the same logic for risk calculation, ensuring consistency in the user experience and the accuracy of predictions across platforms.

D. Enhanced User Experience:

1) Accessibility and Ease of Use: The tool is accessible from anywhere, requiring only internet access, and caters to diverse user preferences with both web and chatbot interfaces.

2) Comprehensive Screening: The tool provides an inclusive approach to screening by considering a wide array of risk factors, making the prediction more robust.

3) Prompt Risk Assessment: Users benefit from immediate risk analysis, empowering them with timely information that can lead to early intervention and better health outcomes.

4) Privacy and Compliance: The system is built with a commitment to user privacy, handling all data under stringent security measures to ensure compliance with healthcare regulations like HIPAA.

5) Educational Resource: Beyond risk assessment, the tool serves as an educational resource, raising awareness about cervical cancer and the significance of regular screening.

E. Integration into Healthcare

Integrating a web-based cervical cancer screening tool with a chatbot into healthcare involves several key steps. First, the tool's accuracy and efficacy should be validated by healthcare professionals and ensure compliance with data privacy laws like HIPAA and GDPR. Then, integrate the tool with EMR and EHR systems for seamless data flow and train medical staff to use and interpret it, incorporating it into standard patient processes.

Additionally, the tool can be used for patient education on cervical cancer risks and incorporated into telehealth services for remote screening, especially in underserved areas. Establish a feedback loop with healthcare providers for continuous improvement and integrate follow-up protocols for high-risk patients. Collaborate with insurance companies for coverage and implement regular quality control checks to maintain the tool's reliability. This comprehensive approach aims to enhance early detection and improve patient outcomes through timely and accurate diagnosis.

F. Benefits to the Healthcare Setting:

1) Efficiency: Automates the initial risk assessment process, saving time for both patients and healthcare providers.

2) Accessibility: This makes preliminary screening more accessible, which can lead to earlier detection and treatment interventions.

3) Patient Engagement: Empowers patients to participate actively in their health management.

4) Resource Allocation: Helps prioritize clinical resources for patients who need them most based on the risk stratification provided by the tool.

5) Data-Driven Insights: Provides valuable data for epidemiological studies and public health initiatives.

VI. DISCUSSION:

Several important conclusions and ramifications for the application of predictive analytics in healthcare settings are brought to light by the comparative study of machine learning models in cervical cancer prediction. The exceptional performance of the XGBoost model, which is distinguished by its superior recall, accuracy, precision, and F1-score balance, highlights the importance of ensemble

learning techniques in tackling challenging classification problems like cervical cancer early detection.

A. Model Performance: An Overview

The study examines various machine learning models for cervical cancer detection, emphasizing the superior accuracy and lower false negative rates of XGBoost. Random Forest, while accurate, struggles with detecting positive cases due to its tendency to favor majority classes. Gaussian Naive Bayes offers high precision but limited recall, potentially due to mismatched model assumptions and data complexity. SVM and Logistic Regression perform reasonably well, particularly in interpretable scenarios. KNN struggles with sparse data and feature selection, impacting positive case detection. Overall, XGBoost and Random Forest show promise for early detection, though cautious integration is advised due to ethical concerns surrounding false results and the need for clear, understandable AI solutions in healthcare.

B. Limitations and Future Work

Despite being thorough, this study has certain drawbacks. The used dataset might not accurately reflect the complexity and diversity of cervical cancer cases across various communities despite being widely known and publicly available. Further investigations into deep learning methods and the integration of imaging data into prediction models could yield larger and more varied datasets for future research. Furthermore, longitudinal research is required to evaluate how these models affect patient outcomes and healthcare systems.

VII. CONCLUSION

The study explores various machine learning algorithms for cervical cancer prediction, with XGBoost emerging as the most promising due to its superior performance across multiple metrics. It highlights the effectiveness of ensemble learning, particularly gradient boosting, in navigating complex medical diagnosis scenarios (Allanson & Schmeler, 2021). The study stresses the importance of careful model selection and fine-tuning for healthcare applications, considering the variability in performance across different metrics. However, limitations such as reliance on a single dataset necessitate further validation across diverse populations and settings. Future research should explore integrating complex variables like imaging data to enhance predictive accuracy [20]. While cutting-edge machine learning algorithms like XGBoost show promise in cervical cancer screening, successful integration into clinical practice requires interdisciplinary collaboration and consideration of operational, ethical, and patient care issues [4].

REFERENCES

[1] Sobar, Machmud, Rizanda, & Wijaya, Adi. (2016). Behavior determinant-based cervical cancer early detection with a machine learning algorithm. Retrieved from https://www.ingentaconnect.com/contentone/asp/asl/2016/00000022/00000010/art00111

[2] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, Volume 13. Retrieved from https://www.sciencedirect.com/science/article/pii/S2001037014000464

[3] World Health Organization. (2019). Human papillomavirus (HPV) and cervical cancer. Retrieved from https://www.who.int/news-room/fact-sheets/detail/human-papillomavirus-(HPV)-and-cervical-cancer

[4] Allanson, E. R., & Schmeler, K. M. (2021, December 01). Preventing Cervical Cancer Globally: Are We Making Progress? Cancer Prevention Research. Retrieved from https://aacrjournals.org/cancerpreventionresearch/article/14/12/1055/675171/Preventing-Cervical-Cancer-Globally-Are-We-Making

[5] Cohen, P. A., Jhingran, A., Oaknin, A., & Denny, L. (2019). Cervical cancer. The Lancet, Volume 393 (Issue 10167), retrieved from https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)32470-X/abstract

[6] Purnami, S., Khasanah, P., Sumartini, S., Chosuvivatwong, V., & Sriplung, H. (2016). Cervical cancer survival prediction using a hybrid of SMOTE, CART, and smooth support vector machine. AIP Conference Proceedings, Retrieved from https://pubs.aip.org/aip/acp/article-abstract/1723/1/030017/815294/Cervical-cancer-survival-prediction-using-hybrid

[7] Lee, Y.-M., Lee, B., Cho, N.-H., & Park, J. H. (2023). Beyond the microscope: A technological overture for cervical cancer detection. Diagnostics, Volume(Issue 19), retrieved from https://mdpi.com/2075-4418/13/19/3079

[8] Bohr, A., & Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7325854/

[9] 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.

[10] Liu, Z., Zhang, X. Y., Shi, Y. J., Wang, L., & Zhu, H. T. (2019). A machine learning-based predictive model of surgical site infection after cervical cancer surgery. PloS one, 14(8), e0220733.

[11] Smith, J. S., Lindsay, L., Hoots, B., Keys, J., Franceschi, S., Winer, R., & Clifford, G. M. (2018). Human papillomavirus type distribution in invasive cervical cancer and high-grade cervical lesions: A meta-analysis update. International Journal of Cancer, 121(3), 621–632.

[12] Kim, J., Kim, H. J., & Kim, Y. (2020). Predictive models for diabetes mellitus using machine learning techniques. BMC Public Health, 20(1), 1-11.

[13] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

[14] Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216-1219.

[15] Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., & Jung, K. (2019). Do no harm: A roadmap for responsible machine learning for health care. Nature Medicine, 25(9), 1337-1340.

[16] Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. New England Journal of Medicine, 378(11), 981–983.

[17] Krishnamoorthi, R., Joshi, S., Almarzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022). A novel diabetes healthcare disease prediction framework using machine learning techniques. Journal of Healthcare Engineering, Volume (2023), pp. 2, 37–45. https://www.hindawi.com/journals/jhe/2022/1684017/

[18] Chen, J. (2024). How a histogram works to display data. Investopedia. Retrieved from https://www.investopedia.com/terms/h/histogram.asp#:~:text=Investopedia%20%2F%20Joules%20Garcia-,What%20Is%20a%20Histogram%3F,into%20logical%20ranges%20or%20bins.

[19] Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The Annals of Statistics, 28(2), 337–407.

[20] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.