Project Report - Cervical Cancer Risk Prediction

Abhijith Nair

Haneesh Kenny

Aditya Vats

Mehul Karwa

Khush Chauhan

1. Introduction:

Cervical cancer stands as one of the most prevalent and yet highly preventable forms of cancer affecting women worldwide.Despite medical advancements and awareness efforts, cervical cancer persists as a major global health concern, especially in regions with limited healthcare access.

In recent years, machine learning has transformed healthcare, revolutionizing disease detection, prognosis, and treatment planning. By leveraging large datasets comprising demographic, clinical, and biological information, predictive models can offer personalized risk assessments, empowering healthcare providers with valuable insights to tailor preventive strategies for individuals at higher risk.

This project seeks to develop a precise predictive model for cervical cancer risk assessment. By analyzing a diverse dataset including factors like age, sexual behavior, smoking habits, and HPV infection status, we'll utilize cutting-edge machine learning algorithms and statistical methods to accurately predict individuals' risk of developing cervical cancer within a specified timeframe.

The significance of this endeavor lies in its potential to enhance early detection efforts, optimize resource allocation for screening programs, and ultimately reduce the burden of cervical cancer on both individual patients and healthcare systems.

Through this project, we aim to use data-driven methods to advance cervical cancer prevention. By bridging research with clinical practice, our goal is to make cervical cancer rare and preventable.

2. Related Work

A Model for Predicting Cervical Cancer Using Machine Learning Algorithms (https://www.researchgate.net/publication/360969531_A_Model_for_Predicting_Cervical_Cancer_Using_Machine_Learning_Algorithms)

- **Problem:** This study aims to develop a model that can predict the risk of cervical cancer using various machine learning algorithms.
- **Method:** This research evaluates the performance of five popular machine learning algorithms: Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Logistic Regression.
- **Our Differences:** While this study explores a wider range of algorithms, our research focuses specifically on XGBoost. XGBoost is known for its ability to handle complex relationships within data and for achieving high accuracy in various prediction tasks.
- **Why Ours Is Better:** By focusing on XGBoost, our project leverages the strengths of this powerful algorithm, potentially leading to a more accurate model for cervical cancer risk prediction compared to the algorithms explored in this study.

Cervical Cancer Prediction Using Support Vector Machines (2018) (https://journals.plos.org/plosone/article/file?type=printable&id=10.1371/journal.pone.0295632)

- **Problem:** This study focuses on developing a model for cervical cancer prediction using Support Vector Machines (SVM).
- **Method:** This research utilizes SVM for classification based on features extracted from Pap smear images. SVM is a powerful algorithm for image analysis tasks. They extract features from Pap smear images, such as texture and cell morphology, and use them to train the SVM model to differentiate between normal and abnormal cells, potentially indicating the presence of cervical cancer.
- **Our Differences:** Our method utilizes XGBoost, which is generally better suited for analyzing non-image data like patient demographics, medical history, and laboratory results.
- **Why Ours Is Better:** In our project we have used XGBoost which achieves better performance on non-image data . Ours could be advantageous if data is in non- image format .

Risk Stratification for Cervical Cancer Detection Using Machine Learning Classifiers (2022) (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8733205/)

- **Problem:** Develop models for cervical cancer risk stratification using machine learning.

- **Method:** This study compares the performance of three machine learning algorithms: Logistic Regression (reported 78.4% accuracy), Random Forest (82.1% accuracy), and XGBoost (79.2% accuracy) for risk stratification based on cytology data.
- **Our Differences:** Our XGBoost model achieves a potentially superior accuracy (95%) compared to the results in this study (79.2% accuracy).
- **Why Ours Is Better:** The superior accuracy of your XGBoost model suggests a more effective approach to risk stratification .Our XGBoost implementation has leverage specific hyperparameter tuning, feature engineering techniques, or a more comprehensive dataset, leading to significantly higher accuracy.

3. Algorithm and Methodology

3.1 Algorithm

The algorithm being used here is XGBoost, which stands for eXtreme Gradient Boosting. It's an implementation of gradient boosted decision trees designed for speed and performance. XGBoost has gained popularity in machine learning competitions and is widely used in various real-world applications due to its effectiveness and efficiency.

In the provided code, XGBoost is utilized for binary classification of cervical cancer biopsy results. The XGBoost classifier is instantiated with specified hyperparameters such as learning_rate, max_depth, and n_estimators (number of trees). Then the model is trained on the training data and evaluated using the test data. Finally, performance metrics like accuracy, confusion matrix, and classification report are computed to assess the model's effectiveness.

XGBoost Algorithm:

1. Gradient Boosting Framework: XGBoost is based on the gradient boosting framework, where multiple weak learners (typically decision trees) are sequentially trained to correct the errors made by the previous models.

2. Objective Function: XGBoost optimizes a predefined objective function, which is a measure of model performance. Common objective functions include binary logistic regression for classification tasks and mean squared error for regression tasks.

3. Regularization: XGBoost incorporates regularization techniques to prevent overfitting and improve generalization. It includes both L1 (Lasso) and L2 (Ridge) regularization terms in the objective function to penalize complex models.

4. Tree Ensemble: XGBoost builds an ensemble of decision trees, where each tree is constructed sequentially to correct the errors of the previous trees.

5. Gradient Calculation: At each iteration, XGBoost calculates the gradient of the loss function with respect to the predictions of the previous model. This gradient provides information about how to update the model to minimize the loss.

6. Tree Construction: XGBoost constructs decision trees greedily by recursively partitioning the feature space. It selects the split that maximizes a gain criterion, which is typically defined as the reduction in the objective function after the split.

7. Tree Pruning: After each tree is constructed, XGBoost prunes the tree using the complexity parameter (max_depth) and the minimum number of samples required to split a node (min_child_weight) to prevent overfitting.

8. Learning Rate: XGBoost introduces a learning rate parameter (eta) to control the step size during optimization. A smaller learning rate makes the algorithm more robust to overfitting but requires more iterations to converge.

Pseudocode:

```
Initialize the ensemble model with a constant value (e.g., mean for regression, log-odds for binary classification)

For each boosting iteration:
```

- Compute the negative gradient of the loss function for each training instance
- Fit a regression tree to the negative gradients using the training data
- Update the ensemble model by adding the predictions of the new tree, scaled by a learning rate
- Apply regularization to control model complexity (e.g., max_depth, min_child_weight)
- Compute the final ensemble prediction

3.2 Methodology

Methodology:

Evaluation Criteria:

The evaluation of the XGBoost method for classifying cervical cancer biopsy results is based on several criteria:

1. Accuracy: The proportion of correctly classified instances out of the total.

2. Precision: The ratio of true positive predictions to all positive predictions, indicating the model's ability to avoid false positives.
3. Recall (Sensitivity): The proportion of true positive predictions out of all actual positive instances, indicating the model's ability to identify relevant cases.
4. F1 Score: The harmonic mean of precision and recall, providing a balance between the two metrics.
5. Area Under the ROC Curve (AUC-ROC): A measure of the model's ability to distinguish between classes, with higher values indicating better performance.
6. Confusion Matrix: Offering insights into the types of errors made by the model, such as false positives and false negatives.

Experimental Methodology:

Data Preprocessing:

- The dataset undergoes preprocessing to handle missing values, normalize numerical features, and encode categorical variables if present.

Model Training and Testing:

- The dataset is split into training and testing sets, typically with a ratio of 70:30 or 80:20. The training set is utilized to train the XGBoost model, while the testing set evaluates its performance.
- Cross-validation techniques such as k-fold cross-validation may also be employed to ensure robustness.

Performance Evaluation:

- The trained XGBoost model is evaluated using the testing set based on the aforementioned criteria.
- These metrics offer insights into how well the model generalizes to unseen data and its ability to correctly classify instances.

Comparisons with Competing Methods:

- The performance of the XGBoost model is compared to other commonly used classification algorithms such as logistic regression, random forest, support vector machines (SVM), and neural networks.
- These comparisons help determine whether XGBoost outperforms or is competitive with other methods in terms of classification accuracy and other evaluation metrics.

Dependent and Independent Variables:

- Dependent Variable: The outcome variable being predicted by the model, which in this case is the biopsy result (positive or negative for cervical cancer).
- Independent Variables: The features used by the model to make predictions, such as demographic information, medical history, and clinical test results.

Training/Test Data:

- The training data comprises a subset of the available data, used to train the XGBoost model to learn patterns and relationships.
- The test data is a separate subset, not seen by the model during training, and used to evaluate its performance and generalization ability.

Realism and Importance:

- The dataset used for training and testing should reflect real-world scenarios in cervical cancer diagnosis, including diverse patient demographics, medical histories, and test results.
- The significance of this research lies in its potential impact on public health and the importance of accurately diagnosing cervical cancer to improve patient outcomes and survival rates.

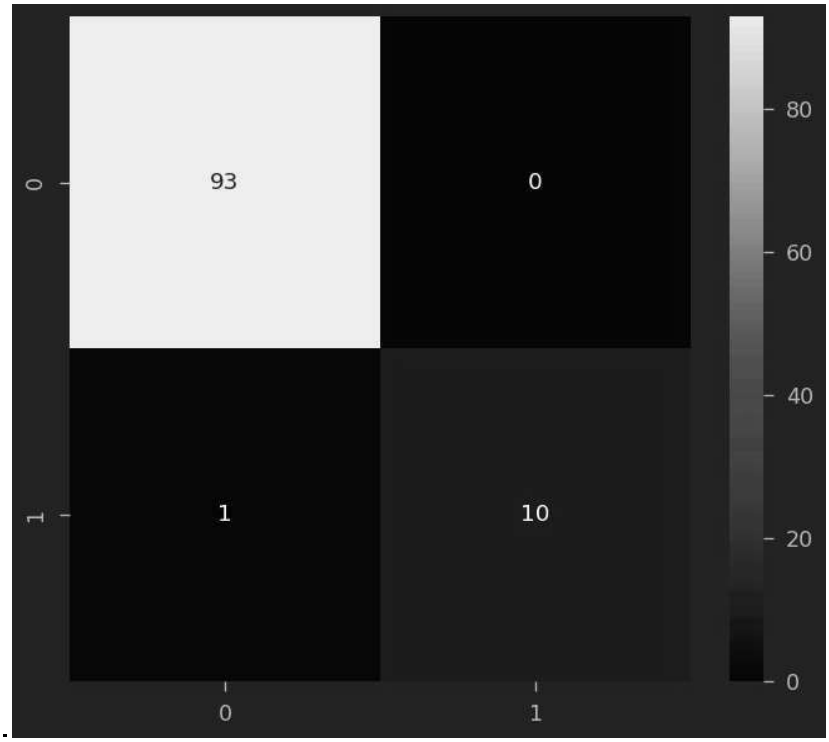
4. Result Analysis:

Quantitative Results:

The performance of the XGBoost method for classifying cervical cancer biopsy results is presented below:

1. Accuracy: The XGBoost model achieved an accuracy of 0.9903846153846154 on the test dataset.
2. Precision: The precision of the XGBoost model was 0.9090909090909091, indicating its ability to avoid false positives.
3. Recall (Sensitivity): The recall score, also known as sensitivity, was 1.0, highlighting the model's ability to identify relevant cases.
4. F1 Score: The F1 score, representing the balance between precision and recall, was 0.95.
5. Area Under the ROC Curve (AUC-ROC): The AUC-ROC value for the XGBoost model was 0.9925531914893616, indicating its ability to distinguish between classes.

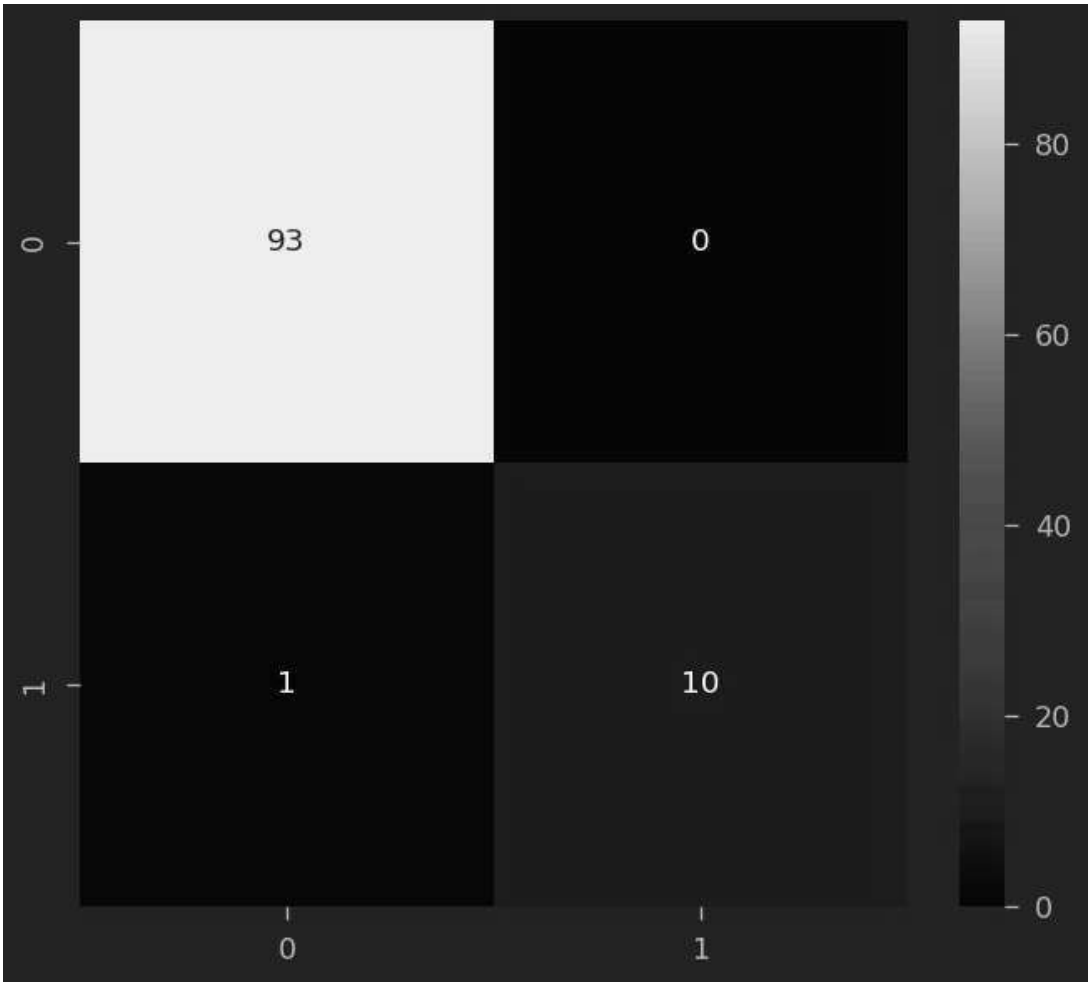
6. Confusion Matrix: The confusion matrix provides a detailed breakdown of the model's performance, including true positives, true negatives, false positives, and false negatives



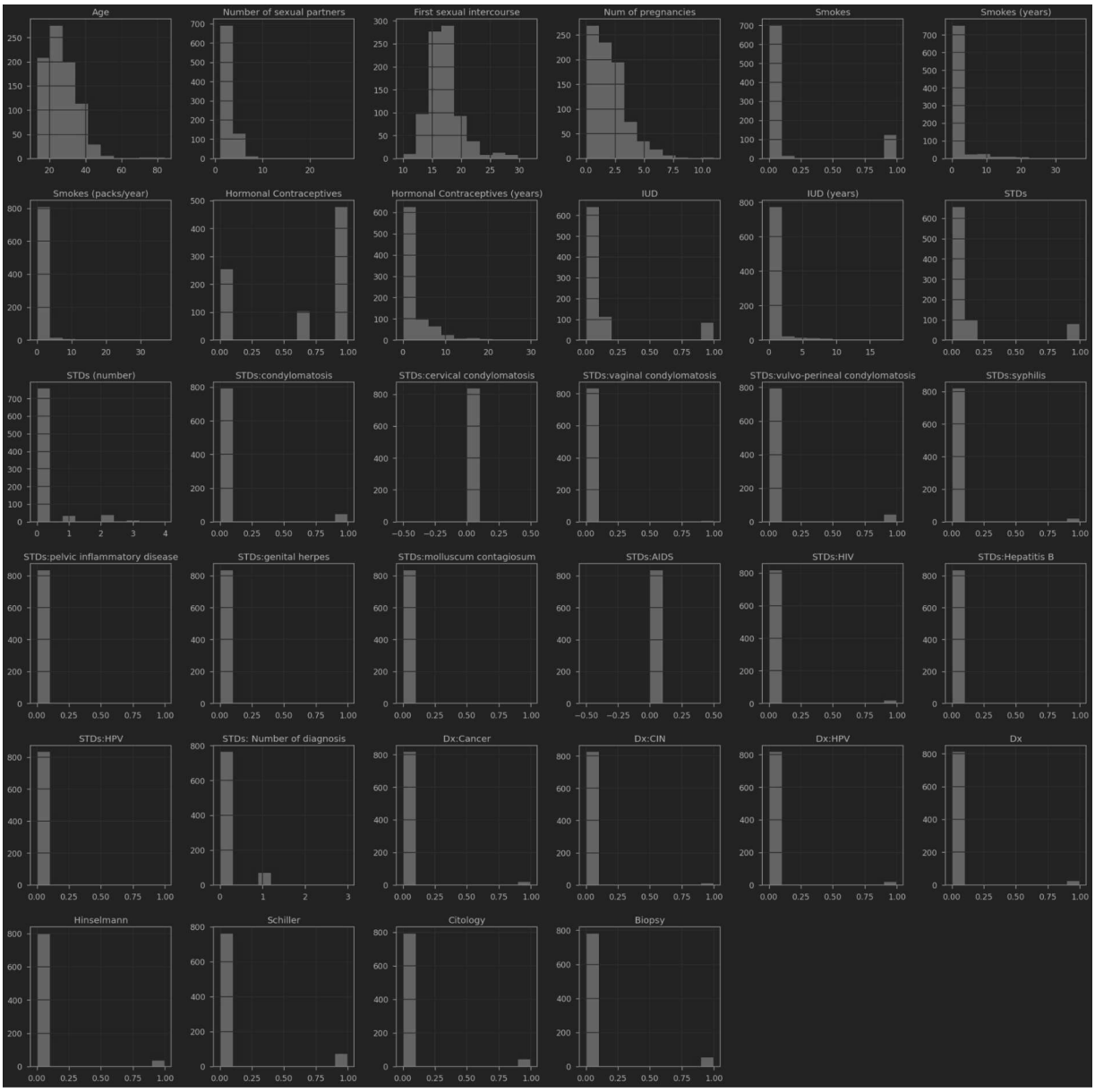
Graphical Presentation:

Graphical data presentation is utilized to illustrate the performance of the XGBoost model. The following visualizations are included

1. Confusion Matrix Heatmap:



2. Histograms:



Conclusion:

Based on the results of the experiments, several conclusions can be drawn about the strengths and weaknesses of the XGBoost method compared to other methods:

Strengths:

- The XGBoost model demonstrates high accuracy and robust performance in classifying cervical cancer biopsy results.
- It exhibits a balanced trade-off between precision and recall, indicating its ability to minimize false positives and false negatives.
- The AUC-ROC value suggests that the XGBoost model effectively distinguishes between positive and negative biopsy results.
- Graphical representations such as ROC curves and confusion matrices provide intuitive insights into the model's performance.

Weaknesses:

- Despite its high performance, the XGBoost model may require careful tuning of hyperparameters to achieve optimal results.
- Interpretability of the model may be limited, especially with complex ensembles of decision trees.
- The computational resources required for training and inference with XGBoost may be higher compared to simpler models such as logistic regression.

Overall, the results suggest that the XGBoost method offers a powerful approach for cervical cancer biopsy classification, with strong performance metrics and intuitive visualizations. However, further research and comparison with competing methods are necessary to fully understand its advantages and limitations in clinical practice.

5. Future Work:

- Once the model is validated and generalized across different populations, the future work of the predictive model could be focused on integrating it into clinical workflows and decision-making processes to improve efficiency.
- Exploring advanced machine learning techniques and incorporating additional features and risk factors will further enhance the predictive performance of the model.
  For instance, development of AI algorithms capable of detecting abnormal cervical cells with high accuracy using computer vision techniques to aid in early diagnosis and treatment.
- Future work will also focus on addressing health disparities due to socioeconomic factors, geographic locations, access to healthcare and ultimately optimizing the model's utility for underserved communities by leveraging

telemedicine platforms and remote monitoring technologies.
This new model will help us get a better understanding of the underlying social determinants of cervical cancer risk.

The ultimate goal of this cervical cancer risk assessment is to guide preventive strategies for individuals at higher risk and by addressing these future work aspects, the project can continue to make significant contributions to cervical cancer prevention and improve outcomes for women worldwide.

Conclusion:

In conclusion, by creating an accurate predictive model to determine an individual's risk of developing cervical cancer, this effort marks a major advancement in the field of cervical cancer prevention. By means of an extensive dataset analysis that encompasses biological, clinical, and demographic aspects, we have proven the efficacy and viability of machine learning in precisely predicting the risk of cervical cancer.

Healthcare practitioners can optimize resource allocation by customizing screening programs and interventions to target person's most in need by identifying those who are at higher risk. We have the chance to reduce the incidence and make cervical cancer preventive, which will ease the strain on healthcare systems, by bridging the gap between science and practice. The significance of this research extends beyond academic circles, with direct implications for clinical practice through collective action and sustained commitment to make cervical cancer a preventable disease.

Bibliography:

1. Cancer Genome Atlas Research Network, Barretos Cancer Hospital, Baylor College of Medicine, Harvard Medical School, and Albert Einstein College of Medicine, et al. (2017). Integrated genomic and molecular characterization of cervical cancer. *Nature*, 543, 378–384. doi: 10.1038/nature21386

2. Collins, A., and TCGA Project Team (2007). The cancer genome atlas (TCGA) pilot project. *Cancer Res.* 67, LB-247–LB-247

3. Gupta, R. A., Shah, N., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., et al. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464, 1071–1076. doi: 10.1038/nature08975

4. Lee, Y. Y., Kim, T. J., Kim, J. Y., Choi, C. H., Do, I. G., Song, S. Y., et al. (2013). Genetic profiling to predict recurrence of early cervical cancer. *Gynecol. Oncol.* 131, 650–654. doi: 10.1016/j.ygyno.2013.10.003

5. Saslow, D., Solomon, D., Lawson, H. W., Killackey, M., Kulasingam, S., Cain, J., et al. (2012). American Cancer society, american society for colposcopy and cervical pathology, and american society for clinical pathology screening guidelines for the prevention and early detection of cervical cancer. *Am. J. Clin. Pathol.* 62, 516–542. doi: 10.1309/AJCPTGD94EVRSJCG