DATA-DRIVEN INSIGHTS INTO CERVICAL CANCER RISK AND PREDICTION

Muhan Ding

Arcadia University, Glenside, PA
mding_02@arcadia.edu

1. MOTIVATION

Cervical cancer, although it is one of the most preventable malignancies, continues to represent a major global health challenge. Annually, over 300,000 women die from cervical cancer, with nearly 90% of these deaths occurring in low- and middle-income countries, where access to early screening programs and preventive education remains limited [1]. Although the widespread implementation of the HPV vaccine has led to a significant reduction in infection rates, a critical gap persists: many women remain insufficiently informed about the full spectrum of risk factors contributing to cervical cancer. These factors include not only HPV infection but also a range of behavioral, reproductive, and lifestyle factors, such as smoking, contraceptive use, and a history of sexually transmitted infections (STIs).

This study seeks to address this gap by transforming clinical data into actionable insights that can better inform healthcare strategies and empower women to make informed decisions about their health. We utilized a dataset of 858 patient records, each containing 36 variables representing demographic, behavioral, and diagnostic factors associated with cervical cancer risk.

The primary goal of this research is to provide actionable insights for young women to make informed health decisions. By deepening the understanding of the various risk factors for cervical cancer, we aim to shift from a vaccine-focused approach to a more comprehensive, lifelong prevention strategy. Ultimately, the goal is to empower women with the knowledge needed to take proactive control of their health and reduce the occurrence of preventable cervical cancer deaths.

2. DATASETS

2.1. Dataset Description

The dataset used in this project is titled *Risk Factors for Cervical Cancer* [2], and it was obtained from the UCI Machine Learning Repository.

It consists of medical and personal history records collected from 858 female patients, with a total of 36 features. These features cover a wide range of demographic, behavioral, reproductive, and diagnostic factors that may contribute to the risk of developing cervical cancer. The dataset is primarily intended for research and predictive modeling related to cervical cancer diagnosis and prevention.

In this project, **Biopsy** is selected as the target variable for prediction, as it is widely regarded as the most definitive diagnostic outcome among those included in the dataset.

2.2. Data Preparation

To prepare the data, all missing values represented by “?” were replaced with NaN, and columns with over 50% missing data were removed for quality. The dataset was then categorized into numerical and categorical variables. Numerical features included continuous data like age and contraceptive use, while categorical features contained binary indicators such as STD diagnosis or diagnostic tests.

For imputation, numerical values were filled using K-Nearest Neighbors (KNN) with 3 neighbors, after standardizing the data with StandardScaler[3]. Categorical values were imputed using mode imputation. After imputation, numerical data were inverse-transformed to restore the original scale.

A correlation matrix[4] (Fig. 1) was computed to identify relationships between features, and highly correlated pairs (above **0.9**) were reduced by removing one feature from each pair. A correlation analysis was also performed between features and the target variable (Biopsy result)(Fig.3), retaining features with an absolute correlation above **0.05**.

Lastly, the dataset was split into numerical and binary features based on cardinality, ensuring a clean and ready dataset for further analysis.

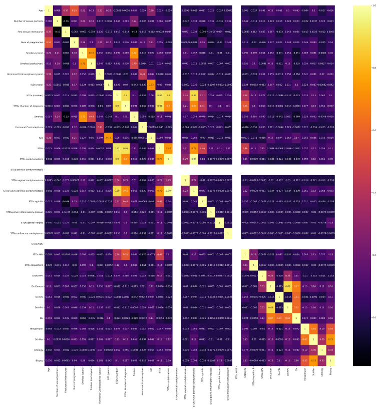

Fig. 1. Correlation matrix between features

3. RESEARCH QUESTIONS AND RELATED WORK

3.1. Research Questions

(1) What clinical and behavioral factors are significantly associated with cervical cancer risk, and how are these factors statistically related to biopsy-confirmed outcomes?

(2) Can machine learning models accurately predict high-risk cervical cancer profiles based on clinical and behavioral data, and how can these predictions inform more effective early screening strategies and suggestions?

3.2. Related work

A similar cervical cancer risk prediction task exists on Kaggle, where researchers primarily rely on a single UCI dataset due to the limited availability of large-scale, authoritative clinical datasets. While these public datasets provide a useful starting point, most existing approaches apply only basic preprocessing steps, such as simple mean or median imputation for missing values and use standard classification models like Logistic Regression or Random Forest without in-depth tuning or evaluation.

In addition, many existing notebooks focus solely on prediction accuracy and lack deeper exploration of the underlying clinical or behavioral factors that may influence cervical cancer risk.

In contrast, our approach involves a more comprehensive and nuanced pipeline. First, we perform advanced data imputation by comparing both statistical (mode) and algorithmic (KNN imputation)[5] methods after visually(Fig. 2) and numerically analyzing missingness patterns. Furthermore, we address class imbalance using SMOTE[6] and go beyond accuracy by evaluating multiple performance metrics (precision, recall, F1-score) under adjusted classification thresholds, which is especially important in a medical context where false negatives carry high cost.

Most importantly, our project aims not only to improve predictive performance, but also to identify and interpret the most influential clinical and behavioral factors through permutation-based feature importance[7] and correlation analysis with biopsy results. This dual emphasis on interpretability and methodological rigor sets our work apart from the largely prediction-focused efforts found on Kaggle.

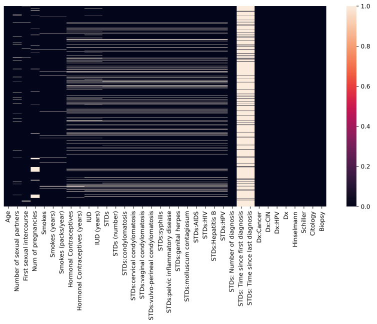

Fig. 2. Missing values in the dataset

4. RESEARCH METHODS

To address the research objectives of identifying significant clinical and behavioral risk factors for cervical cancer result beyond HPV infection and accurately detecting high-risk patient profiles, we implemented a supervised machine learning framework focused on classification modeling.

Initial exploratory data analysis (EDA) involved descriptive statistics, distribution plots, and correlation heatmaps to assess data structure, detect outliers, and uncover preliminary variable relationships.

Given the substantial proportion of missing values in clinical features, we employed **K-Nearest Neighbors (KNN)**[5] imputation to estimate missing numerical data. This approach preserved multivariate relationships and reduced imputation bias compared to simpler techniques.

For predictive modeling, we utilized the **Random Forest classifier**[8] due to its robustness to noise, nonlinear relationships, and feature interactions. To mitigate the issue of class imbalance—a common challenge in medical datasets, we applied the **Synthetic Minority Over-sampling Technique (SMOT)** [6], thereby enhancing representation of the minority class during training.

To ensure the model's robustness and optimal predictive capability, we systematically optimized its configuration through hyperparameter tuning using **GridSearchCV**[9], focusing on key parameters such as the number of trees, tree depth, and minimum samples per leaf in the Random Forest model. **Cross-validation** was employed to mitigate overfitting and validate model performance across different data folds. We assessed the model using multiple evaluation metrics– accuracy, precision, recall, and F1-score, to ensure balanced performance, especially given the class imbalance in the cervical cancer dataset. This comprehensive evaluation not only facilitated reliable identification of high-risk profiles but also enhanced the model's ability to generalize across varying clinical data patterns.

We also quantified the contribution of individual features to the model's predictive outcomes to identify the most influential risk factors. Using **permutation feature importance**, we ranked features based on their impact on model performance, providing insights into the relative importance of both clinical and behavioral variables. This analysis supported the identification of key determinants associated with biopsy-confirmed cervical cancer risk and aligned with the broader objective of enhancing understanding of the underlying risk factors.

5. FINDINGS AND INTERPRETATION

We began by summarizing key clinical features in the dataset. All the analysis are based on the information about the patients whose biopsy test results are positive. 65.45% of patients reported the use of hormonal contraceptives, with an average duration of 3.32 years. 16.36% of patients reported the use of intrauterine devices (IUDs). Furthermore, 21.8% of patients had a recorded history of sexually transmitted diseases (STDs), and 12.7% had a diagnosis documented under the general "Dx" category.

Regarding diagnostic screening tools on the patients whose biopsy test results are positive, 87.3% of patients also tested positive on the Schiller test, 45.5% on the Hinselmann test, and 32.7% on the cytology test. Among individuals who yielded positive biopsy results, 18.18% reported smoking, with an average smoking duration of 2.15 years and a mean consumption of 0.65 packs per year. The mean age among patients with abnormal biopsy findings was 28.64 years.

These findings underscore several potential risk indicators associated with cervical abnormalities. The high positivity rate of the Schiller test suggests its utility as a sensitive initial screening method. The observed prevalence of STDs and hormonal contraceptive use among patients may point to relevant behavioral or physiological risk factors contributing to cervical dysplasia or neoplastic changes. Importantly, the relatively young mean age of patients with positive biopsy outcomes highlights the necessity of early and proactive screening efforts, particularly in populations under 30 years of age. Although the percentage of smokers was comparatively low, the presence of cervical abnormalities in this subgroup reinforces the documented association between tobacco exposure and cervical epithelial changes, even in cases of low-intensity smoking.

Next, we assessed the correlation between clinical features and the biopsy result (Fig. 3), which served as the target variable in our analysis. Among all evaluated variables, the highest correlation coefficients were associated with diagnostic screening tests, specifically: Schiller test (r = 0.73), Hinselmann test (r = 0.55), and Cytology test (r = 0.33).

These results indicate a strong positive association between abnormal test outcomes and the presence of cervical pathology as confirmed by biopsy. Notably, the Schiller test demonstrated the strongest correlation, suggesting it may serve as a particularly informative feature in predictive modeling. Overall, these diagnostic indicators are expected to be among the most influential predictors in the context of machine learning-based classification for cervical cancer risk stratification.

Following this, we constructed a Random Forest Classifier to predict cancer risk based on patient features. To ensure optimal performance, we applied GridSearchCV[9] to tune key hyperparameters. Since the dataset was imbalanced, we employed SMOTE to oversample the minority class (positive biopsy cases). The model achieved an overall accuracy of **97%**. A confusion matrix was show (Fig. 4) and Performance metrics for the minority class (positive biopsy) were: Precision: 0.65, Recall: 0.87, F1-score: 0.74.

The high recall (0.87) indicates that 87% of true positive cases were correctly identified, which is critical in a clinical context where missing a positive diagnosis could lead to delayed treatment or adverse outcomes.

However, the moderate precision (0.65) suggests that approximately 35% of positive predictions were false positives. While this may lead to some over-referral or unnecessary follow-up testing, such a trade-off is often acceptable in medical diagnostics—especially in screening scenarios where maximizing sensitivity (recall) is prioritized over precision.

The F1-score of 0.74, which balances both precision and recall, confirms that the model maintains an overall good level of performance.

To gain deeper insights into the model's decision-making process, we applied permutation importance (Fig. 5) to assess the relative contribution of each feature to the model's predictions. Among the features evaluated, the Schiller test exhibited the highest importance, with a mean importance score of 0.0.092606, indicating that it plays a pivotal role in predicting cervical cancer risk. In contrast, other features demonstrated a relatively modest influence, suggesting that it has a limited effect on the model's classification decisions.

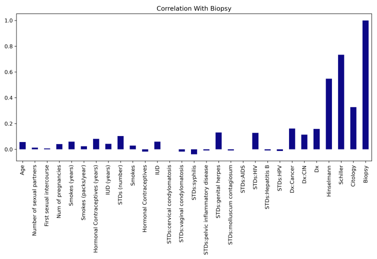

Fig. 3. Correlation between features and biopsy result

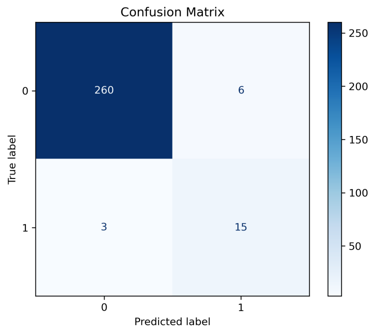

Fig. 4. Confusion matrix of model prediction

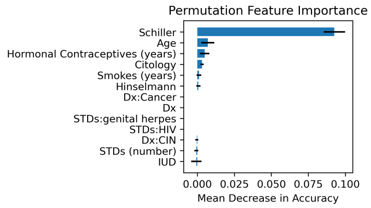

Fig. 5. Permutation feature importance

6. CONCLUSIONS

In this project, we used the UCI Cervical Cancer Risk Factors dataset to investigate significant clinical and behavioral contributors to cervical cancer risk, and to develop a predictive model for early detection.

Specifically, we aimed to (1) identify features most strongly associated with abnormal biopsy outcomes, and (2) assess whether machine learning can effectively predict risk based on some clinical data.

We conducted data cleaning, correlation analysis, and feature selection. One of the two highly correlated features was removed to reduce multicollinearity, enhance model stability, and improve the reliability of feature importance analysis. Exploratory analysis highlighted strong associations between diagnostic tests—Schiller, Hinselmann, and Citology and Biopsy results. Behavioral and physiological factors, such as a history of sexually transmitted infections, hormonal contraceptive use, and tobacco exposure also contribute to elevated risk. Notably, abnormal biopsy results were also observed among young individuals, emphasizing the importance of initiating proactive screening efforts earlier.

A Random Forest Classifier was trained using SMOTE to address class imbalance, and hyperparameters were optimized with GridSearchCV. The final model achieved a **97%** accuracy and **87%** recall for detecting positive biopsy cases, indicating high effectiveness in identifying at risk patients.

Permutation importance further confirmed that Schiller and Hinselmann tests were the most predictive features. These findings support the integration of data-driven models into cervical cancer screening workflows, especially in environments where early detection is critical.

7. ACKNOWLEDGEMENT AND REFLECTION ON THE USE OF LLMS

For this project, we received valuable assistance from ChatGPT, primarily during data preparation, modeling, and writing stages. It helped us understand advanced techniques like K-Nearest Neighbors (KNN) for missing value imputation. Initially, we used average values to handle missing numerical data, but this yielded low accuracy. Seeking a more effective approach, ChatGPT suggested KNN imputation, which significantly improved results.

When addressing the challenge of data information imbalance (with most data showing negative biopsy results), ChatGPT introduced us to SMOTE (Synthetic Minority Over-sampling Technique), a method previously unfamiliar to us. After researching and implementing this technique, we found it substantially enhanced model accuracy.

ChatGPT also assisted in refining our report by suggesting clearer phrasing, improving transitions between sections, and helping articulate technical findings more concisely in an academic tone. When we encountered LaTeX formatting challenges, particularly with citations and paragraph structure, it provided efficient solutions to syntax and formatting issues.

Our experience using ChatGPT throughout this course has been beneficial. We consistently verified and ensured us understood all suggestions, using it as a tool for learning, ideation, and editing rather than as a replacement for critical thinking or original work. The assistance enhanced our ability to understand and express complex technical concepts effectively. When used responsibly, LLMs like ChatGPT can serve as valuable companions in the research and writing process.

8. REFERENCES

[1] W. H. Organization, "Cervical cancer," *WHO Fact Sheets*, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cervical-cancer

[2] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Transfer learning with partial observability applied to cervical cancer screening," in *Iberian Conference on Pattern Recognition and Image Analysis*, 2017, pp. 243–250. [Online]. Available: https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors

[3] Scikit-learn Developers, "Standardization of features," https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[4] pandas development team, "pandas.dataframe.corr — pandas 2.2.3 documentation," https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html.

[5] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for dna microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html

[6] G. Lemaître, F. Nogueira, and C. K. Aridas, "imbalanced-learn: Smote implementation," https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html.

[7] Scikit-learn Developers, “Permutation feature importance,” https://scikit-learn.org/stable/modules/permutation_importance.html.

[8] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: https://link.springer.com/article/10.1023/A:1010933404324

[9] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html