**Improving Reliability of Cervical Cancer Early Detection by**

**Utilizing Multiple Machine Learning Approaches**

**1. Abstract:**

Cervical cancer is the fourth most common cancer in women compared to other types of cancer, with an estimated 604 000 new cases and 342 000 deaths in 2020. About 90% of the new cases and deaths worldwide in 2020 occurred in low- and middle-income countries. Therefore, early detection of cervical cancer is crucial to reduce this disease's deadliness. Several predictive models are built based on an 858-sample dataset from UCI Machine Learning Repository with 32 features and 4 targets, which are also the 4 most common tests for cervical Cancer: Hinselmann, Schiller, Cytology, and Biopsy. This dataset suffers from imbalance with only less than 9% positive patients and approximately 20% missing values. Besides, 32 attributes appear redundant to feed a predictive model, which may lead to potential overfitting. Therefore, several machine learning approaches have been deployed to deal with the aforementioned problems, such as feature engineering, resampling, and feature selection. This paper concludes that Random Forest Classification, with the support of Border-SMOTE and Meta-transformer for selecting features based on importance weights for the Hinselmann, shows the most outstanding performance, with 9 chosen features, generating an accuracy of 98.18%. The last section of the work proposes several implications of the finding.

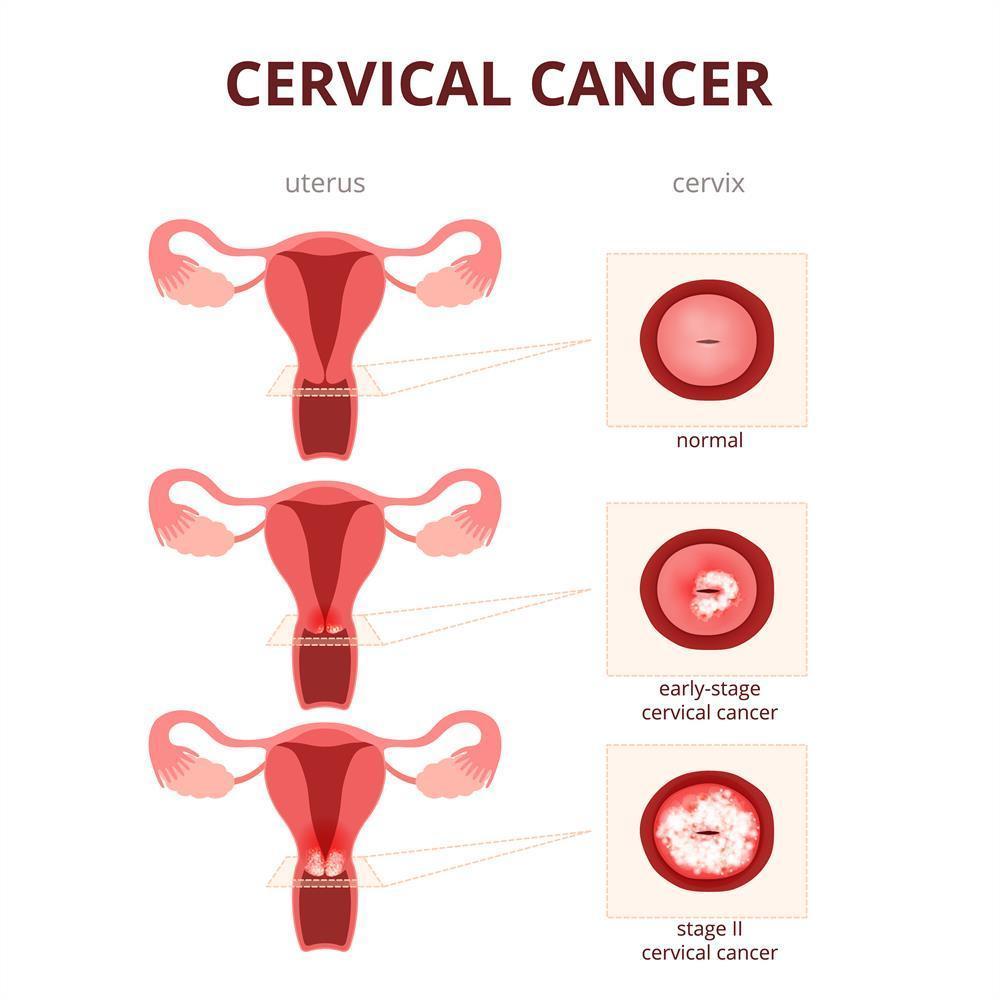
Keywords: cervical cancer, feature selection, imbalanced dataset, resampling, border-SMOTE, random forest, grid-search cross-validation.

**2. Introduction:**

Cancer is one of the most deadly diseases in the world. According to the World Health Organization [1], cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020. The lack of diagnosable symptoms at an early cancer stage creates difficulty in early detection and cure for late stages. Medical sciences have focused on finding solutions to the deadly disease, and a number of progress have been made in the early detection and treatment of cancer. In the United States, Cancer has been the second leading cause of the disease. This paper works on one type of cancer prevalent in women, called cervical cancer. Cervical cancer is the fourth most common cancer in women compared to other types of cancer, with an estimated 604 000 new cases and 342 000 deaths in 2020. About 90% of the new cases and deaths worldwide in 2020 occurred in low- and middle-income countries.

According to the Center for disease control and prevention [[2]](https://www.cdc.gov/cancer/cervical/basic_info/), cancer is defined as a disease in which abnormal cells divide out of control and can invade other tissue. The blood and lymph system in the human body is the medium through which cancer can spread to other body parts. Cervical cancer is a type of cancer in which abnormal cell growth occurs in the cervix, the lower part of the uterus. The cervix connects the uterus to the vagina. The figure depicts a cervix with abnormal cell growth developing into a tumor (see fig. 1).

For the detection of cervical cancer, we have four crucial tests that are used. These four tests are Hinselmann, Schiller, Cytology, and Biopsy.



[Fig.1](https://www.quellerfisher.com/blog/wp-content/uploads/sites/464/2018/02/Cervical-cancer-simple-illustration.jpg) Cervical Cancer illustration; Queller, Fisher et al. “Erin Andrews’ Story Affirms Importance of Proper Physicals and Early Diagnosis of Cervical Cancer.” *New York Personal Injury Law Blog*, 29 July 2020, quellerfisher.com/blog/erin-andrews-story-affirms-importance-of-proper-physicals-and-early-diagnosis-of-cervical-cancer.

In Section 3, several previous findings, so-called related works, are mentioned and discussed. Sections 4 covers all methods used to develop research results, which are carefully presented in Section 5. In the last section of this paper, Section 6, a discussion about the research conclusion and future work are proposed.

**3. Related Works:**

In this section, a few recent publications on the domain are discussed. [Mehmood M.](https://www.frontiersin.org/articles/10.3389/fpubh.2021.788376/full#h6) et al. [3] published a paper in which they built a model called CervDetect to predict the chance of a woman having cervical cancer. They claimed that their model accurately predicts cervical cancer, outperforms the state-of-the-art studies, and achieved an accuracy of 93.6%, mean squared error (MSE) error of 0.07111, false-positive rate (FPR) of 6.4%, and false-negative rate (FNR) of 100%. However, a flawed result was presented since they did not pre-process the problem of the imbalanced dataset, leading to the fact that their ML models predict all patients are cancer-free. Regardless of high accuracy, their precision and sensitivity scores are 0%, meaning they failed to predict whether a person has cancer or not. The same issue was observed in the findings of [Parikh D. et al](https://www.mecs-press.org/ijmsc/ijmsc-v5-n1/IJMSC-V5-N1-5.pdf) [4]. The authors deployed the K-nearest neighbor algorithm, yielding low precision and sensitivity scores. It will be wrong if a person with cancer is told they do not have cancer. It's possible that it will be too late when symptoms appear.

Improving a classification model without considering the actual cost of each error case could result in unreliable results. Fortunately, an improvement was found in the research of [Fatlawi H.](http://www.ijctjournal.org/Volume4/Issue4/IJCT-V4I4P19.pdf) [5]. The researcher achieved higher sensitivity and precision of 42.9% by proposing a model that depends on a decision tree classifier with a cost matrix with different cost values. It has a higher error cost in cases where infected patients have positive medical tests but are classified as not infected. The proposed model produces more accurate results in both binary and multiclass classification.

Another finding by [W. Wu and H. Zhou](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8070120) [6] produce the highest accuracy of 94.13% for the Biopsy test with perfect sensitivity of 100% and specificity of 90.21%. Furthermore, by applying SVM-PCA, there was a significant reduction in features to 11 selected attributes, yielding a slightly lower accuracy of 94.03%. In the meantime, they could preserve the sensitivity but decrease the specificity slightly. They found that the standard Support Vector Machine (SVM) method properly classifies malignant and benign cancers. SVM-RFE and SVM-PCA, in particular, can reduce feature numbers from 30 to 8 to complete the classification. Meanwhile, the classification speed can be significantly improved. Furthermore, while the SVM method can accurately classify cervical cancer data, its high computation cost is a drawback. The problem can be effectively solved using SVM-RFE and SVM-PCA.

[Arora. M. et al](https://eudl.eu/pdf/10.4108/eai.13-7-2018.164264) [7] found that the problem of the dataset imbalance can be treated by using several resampling approaches, including Random Over Sampling (ROS), Random Under Sampling (RUS), and Synthetic Minority Over-sampling Technique (SMOTE), making the data balanced before feeding the data to the proposed model. The results show that the Random Forest (RF) with SMOTE algorithm outperforms four target variables: Schiller, Biopsy, Hinselmann, and Cytology, whereas the KNN outperforms very poorly. They achieved an accuracy of 93.29% and a sensitivity of 100%.

Feature selection was performed in a research conducted by [Chaudhuri A. et al.](https://www.mecs-press.org/ijisa/ijisa-v13-n5/IJISA-V13-N5-5.pdf) [8] to reduce unimportant factors before feeding the ML model. In this paper, only five features were selected for a Logistic Regression (LR) model, yielding an accuracy and sensitivity of 96% and 95% specificity. Since LR is a low bias, high variance algorithm, the apparent error rates in LR results may underestimate the true value because the model concentrates near the observed points that fit the best. As a result, these points may give a false impression of the model's true precision. This is a unique approach with only five features but achieving such high performance. The finding of [Sagala M.](https://iopscience.iop.org/article/10.1088/1742-6596/1255/1/012022/pdf) [9] also takes advantage of feature reduction, a Correlation-based Filter (CFS), to increase the accuracy and reduce the noise of the dataset, with only four features being selected. The most outstanding results achieved, 96.24% accuracy, were for Biopsy and Cytology targets as combinations of CFS and K-Nearest Neighbors (KNN) and Naive Bayes (NB) algorithms, respectively. The researcher concluded that the performance of Naive Bayes with CFS is superior to that of other classifiers with CFS or Naive Bayes with/without feature selection methods in terms of accuracy, specificity, and sensitivity. This is to be expected because Naive Bayes may perform better on a smaller number of relevant attributes. On the contrary, SVM performance improved when using a larger number of attributes. Based on the results of the experiments, the authors discovered that KNN performed best at k = 7. In terms of computational time, Naive Bayes takes less time than other classifiers to build a model. The research conducted by [Sharma M.](https://link.springer.com/article/10.1007/s12553-019-00375-8#article-info) [10] also boosted accuracy by various ML approaches, including feature selections. With a genetic algorithm, the number of attributes is reduced to 7, and an adaptive boosting technique is proposed for further performance improvement. For Support Vector Machine Radial Bias Function (SVM RBF), SVM Linear, and Decision Tree, the improved accuracy lies between 94.17% and 94.69%, sensitivity 97.36% - 98.90%, specificity 93.37% - 94.72%, and precision 93% - 95.17%.

Six features were selected in the finding of [Al-Wesabi, Y. M. S., et al.](https://arxiv.org/ftp/arxiv/papers/1812/1812.10383.pdf) [11]. ROS, RUS, and a combination of the two methods were used to address the unbalanced data in this research. Using feature selection methods, more analysis is carried out. The authors get the best overall result with accuracy exceeding 97 percent by selecting the standard features among the DT and KNN.

Table 1

Publications relevant to Machine Learning methods used in Cervical Cancer detection

| Reference | No. of Features | ML Techniques | Accuracy | Sensitivity | Specificity |
| --- | --- | --- | --- | --- | --- |
| [[3]](https://www.frontiersin.org/articles/10.3389/fpubh.2021.788376/full#h6) | 12 | RF + ANN | 93.6% | 0% | 93.2% |
| [[4]](https://www.mecs-press.org/ijmsc/ijmsc-v5-n1/IJMSC-V5-N1-5.pdf) | 32 | KNN | 82.2% | N/A | N/A |
| [[5]](http://www.ijctjournal.org/Volume4/Issue4/IJCT-V4I4P19.pdf) | N/A | DT | 42.9% | 42.9% | N/A |
| [[6]](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8070120) | 11 | SVM + PCA | 94.0% | 100% | 90.1% |
| [[7]](https://eudl.eu/pdf/10.4108/eai.13-7-2018.164264) | 32 | RF | 93.29% | 100% | 0% |
| [[8]](https://www.mecs-press.org/ijisa/ijisa-v13-n5/IJISA-V13-N5-5.pdf) | 5 | LR | 96% | 96% | 95% |
| [[9]](https://iopscience.iop.org/article/10.1088/1742-6596/1255/1/012022/pdf) | 4 | NB | 96.24% | 100% | 92.5% |
| [[10]](https://link.springer.com/article/10.1007/s12553-019-00375-8#article-info) | 7 | DT + IGAAB | 95.97% | 98.9% | 93.5% |
| [[11]](https://arxiv.org/ftp/arxiv/papers/1812/1812.10383.pdf) | 6 | DT | 97.52% | 100% | 95.03% |

**4. Methods**

4.1. Exploratory Data Analysis

The Cervical Cancer dataset used in this study came from the University of California at Irvine's repository. The dataset comprises 858 patients' clinical histories, which are described using 32 features and four targets (Schiller, Hinselmann, Biopsy, and Cytology).

Table 2

Feature Description

| Number | Features | Non-null Data | Missing Data | Data Type |
| --- | --- | --- | --- | --- |
| F1 | Age | 858 | 0% | Integer |
| F2 | Number of sexual partners | 832 | 3% | Integer |
| F3 | First sexual intercourse(age) | 851 | 1% | Integer |
| F4 | Number of pregnancies | 802 | 7% | Integer |
| F5 | Smokes | 845 | 2% | Boolean |
| F6 | Smokes(years) | 845 | 2% | Boolean |
| F7 | Smokes(packs/years) | 845 | 2% | Boolean |
| F8 | Hormonal Contraceptives | 750 | 13% | Boolean |
| F9 | Hormonal Contraceptives(years) | 750 | 13% | Integer |
| F10 | Intrauterine Device(IUD) | 741 | 14% | Boolean |
| F11 | IUD(years) | 741 | 14% | Integer |
| F12 | Sexually Transmitted disease (STD) | 753 | 12% | Boolean |
| F13 | STDs (number) | 753 | 12% | Integer |
| F14 | STDs: condylomatosis | 753 | 12% | Boolean |
| F15 | STDS: cervical condylomatosis | 753 | 12% | Boolean |
| F16 | STDS: vaginal condylomatosis | 753 | 12% | Boolean |
| F17 | STDs: vulvo-perineal condylomatosis | 753 | 12% | Boolean |
| F18 | STDs: syphilis | 753 | 12% | Boolean |
| F19 | STDs: pelvic inflammatory disease | 753 | 12% | Boolean |
| F20 | STDs: genital herpes | 753 | 12% | Boolean |
| F21 | STDs: molluscum contagiosum | 753 | 12% | Boolean |
| F22 | STDs: AIDS | 753 | 12% | Boolean |
| F23 | STDs: HIV | 753 | 12% | Boolean |
| F24 | STDs: Hepatitis B | 753 | 12% | Boolean |
| F25 | STDs: HPV | 753 | 12% | Boolean |
| F26 | STDs: Number of diagnosis | 858 | 0% | Integer |
| F27 | STDs: Time since first diagnosis | 71 | 92% | Integer |
| F28 | STDs: Time since last diagnosis | 71 | 92% | Integer |
| F29 | Dx: Cancer | 858 | 0% | Boolean |
| F30 | Dx: Cervical Intraepithelial Neoplasia (CIN) | 858 | 0% | Boolean |
| F31 | Dx: Human Papillomavirus(HPV) | 858 | 0% | Boolean |
| F32 | Dx (Diagnosis) | 858 | 0% | Boolean |

The description of these four target variables is given below.

* [Hinselmann test](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3279084/): Mr. Hinselmann developed the Hinselmann test [12]. This test is developed for the early discovery of cervical carcinoma, a colposcopic method. In this test, cervixes are visually inspected on a magnified scale.
* [Schiller](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2448516/pdf/ulstermedj00068-0038.pdf): Walter Schiller investigated the origins of diseases in the United States and Austria, and in 1928, he developed the Schiller test, which allows women to detect early cervical cancer [13]. The Schiller's test, also known as the Schiller's Iodine test, is a medical procedure in which an iodine solution is applied to the skin. The Schiller test" can be used as part of a colposcopic vaginal and cervix examination.
* [Cytology](https://www.mskcc.org/cancer-care/types/cervical/risk-prevention-screening#:~:text=Your%20doctor%20can%20use%20a%20Pap%20smear%20or,mucus%20and%20cells%20by%20gently%20scraping%20your%20cervix.): Liquid-based cytology is a cervical cancer screening diagnostic in which a clinician draws fluid to study the cells [14]. Changes in the cervix can be detected using liquid-based cytology, as well as HPV testing.
* [Biopsy](http://www.jsirjournal.com/Vol4_Issue1_02.pdf): In a biopsy, samples of tissue are taken from the abnormal part of the cervix to detect the cancerous cell [15]. The samples are then studied by a senior consultant pathologist.

Table 3

Target Description

| Target | Patient | Patient Percent | Data Type |
| --- | --- | --- | --- |
| Hinselmann | 35 | 4% | Boolean |
| Schiller | 74 | 9% | Boolean |
| Citology | 44 | 5% | Boolean |
| Biopsy | 55 | 6% | Boolean |

*4.2. Proposed Architecture*

Two central problems of the dataset were noticed before any feature engineering had been performed: missing values and imbalanced dataset. Instead of dropping all null values, multiple techniques have been utilized to fill missing them out case to case, depending on the number of missing values in each feature. In terms of dealing with the imbalanced dataset, three main methods will be used: Border - Synthetic Minority Over-sampling (Border-SMOTE), Border - Synthetic Minority Over-sampling Support Vector Machine (Border-SMOTE SVM), and Adaptive Synthetic ([ADASYN](http://localhost:8888/notebooks/Desktop/Data%20Sciene/Project/cervical-cancer-project/Model%20Building/Processing/Citology/ADASYN%20-%20After%20Using%20Feature%20Selection%20-%20SelectFromModel.ipynb)) algorithm. In lieu of feeding the machine learning model with all 32 features, which is proposed to be redundant, two feature selection methods will be performed, Meta-transformer for selecting features based on importance weights (Select from Model) and Univariate Feature Selection. After being processed with all mentioned steps, the clean dataset will be split into a training set and a testing set with a ratio of 80-20. Following this step, the training dataset is fed to four predictive models. Each will take advantage of the following algorithms: Decision Tree, Random Forest, Grid Search Cross-Validation, and eXtreme Gradient Boosting. Each model's performance is evaluated using test data. The predicted results are then compared to the actual results to assess the model's performance.

*4.3. Feature Engineering*

26 out of 32 features contain missing values. Since the dataset has 858 data points, excluding two features: 'STDs: Time since last diagnosis' and 'STDs: Time since first diagnosis' (more than 90% null values), there are still 190 instances that contain null values. Each column containing null values will be processed differently based on the feature characteristics.

Table 4

Missing values and outliers’ solution description

| Feature | Data Type | No. of Null Values | Missing values and outliers’ solutions |
| --- | --- | --- | --- |
| Number of sexual partners | Integer | 26 | Replace missing values with the mean. Remove all outliers greater than 11 |
| First sexual intercourse | Integer | 7 | Replace missing values by the mean |
| Num of pregnancies | Integer | 56 | Replace missing values by the mean |
| Smokes | Boolean | 13 | Replace missing values by the mode, which is 0 |
| Smokes (years) | Boolean | 13 | Replace missing values by 0 because if a person does not smoke, they will have 0 years of smokes |
| Smokes (packs/year) | Integer | 13 | Replace missing values by 0 because if a person does not smoke, they will consume 0 packs/year |
| Hormonal Contraceptives | Boolean | 108 | Fill missing values using KNN Imputer |
| Hormonal Contraceptives (years) | Integer | 108 | If the instance does not use Hormonal Contraceptives, fill Hormonal Contraceptives years with 0. Otherwise, replace the left missing values using the Univariate Approach |
| IUD | Boolean | 117 | Replace missing values by building a Decision Tree model to predict missing values |
| IUD (years) | Integer | 117 | If the instance does not use IUD, fill IUD years with 0. Otherwise, replace the left missing values using the mean |
| STDs | Boolean | 105 | Replace missing values by building a Decision Tree model to predict missing values |
| STDs (number) | Integer | 105 | Replace the missing values using a Univariate Approach |
| STDs:condylomatosis | Boolean | 105 | Replace missing values by building a Decision Tree model to predict missing values |
| STDs:cervical condylomatosis | Boolean | 105 | Drop because there is only one unique value |
| STDs:vaginal condylomatosis | Boolean | 105 | Drop because only one instance has output 1; thus, it will have very little predictive return |
| STDs:vulvo-perineal condylomatosis | Boolean | 105 | Drop because it is 99% correlated with ‘STDs:vaginal condylomatosis’ column |
| STDs:syphilis | Boolean | 105 | Replace missing values by building a Decision Tree model to predict missing values |
| STDs:pelvic inflammatory disease | Boolean | 105 | Drop because only one instance has output 1; thus, it will have very little predictive return |
| STDs:genital herpes | Boolean | 105 | Drop because only one instance has output 1; thus, it will have very little predictive return |
| STDs:molluscum contagiosum | Boolean | 105 | Drop because only one instance has output 1; thus, it will have very little predictive return |
| STDs:AIDS | Boolean | 105 | Drop because there is only one unique value |
| STDs:HIV | Boolean | 105 | Replace missing values by building a Decision Tree model to predict missing values |
| STDs:Hepatitis B | Boolean | 105 | Drop because only one instance has output 1; thus, it will have very little predictive return |
| STDs:HPV | Boolean | 105 | Drop because only two instances have output 1; thus, it will have very little predictive return |
| STDs: Time since first diagnosis | Integer | 787 | Drop feature |
| STDs: Time since last diagnosis | Integer | 787 | Drop feature |

After completing feature engineering, 32 features were reduced to 21 features for further mining stages: 'Age', 'Number of sexual partners', 'First sexual intercourse', 'Num of pregnancies', 'Smokes', 'Smokes (years)', 'Smokes (packs/year)', 'Hormonal Contraceptives', 'Hormonal Contraceptives (years)', 'IUD', 'IUD (years)', 'STDs', 'STDs (number)', 'STDs:condylomatosis', 'STDs:syphilis', 'STDs:HIV', 'STDs: Number of diagnosis', 'Dx:Cancer', 'Dx:CIN', 'Dx:HPV', 'Dx'. Only 2 instances were dropped, making the number of instances decrease from 858 to 856.

*4.3. Choice of Data Mining Methods*

According to [Issam El Naqa](https://link.springer.com/chapter/10.1007/978-3-319-18305-3_1#auth-Issam-El_Naqa) et al. [16], Machine learning is a growing field of computational algorithms that aim to mimic human intelligence by learning from their surroundings. In the new era of so-called big data, they are regarded as the workhorse. Pattern recognition, computer vision, spacecraft engineering, finance, entertainment, computational biology, and biomedical and medical applications have all benefited from machine learning techniques. Machine learning algorithms' ability to learn from the current context and generalize to new tasks would improve radiotherapy practice's safety and efficacy, leading to better outcomes.

To compare the efficiency of each data mining method, Linear Regression (LR), K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Grid Search Cross-Validation (CV), Naive Bayes (NB), eXtreme Gradient Boosting (XGB), and Artificial Neural Network (ANN) will be deployed before resampling and after resampling to observe the differences in terms of performance metrics. Therefore, several most precise methods will be selected for the subsequent model improvements, including Feature Selection.

*4.4. Performance Metrics:*

The following measures are used to evaluate the proposed work's performance. The Confusion Matrix is used to evaluate a learning model's performance. The performance matrices are created using four terms related to the confusion matrix. True Positive is the number of cervical cancer patients classified as cancer patients (TP). The number of non-cancer patients classified as cervical cancer patients is called False Positive (FP). The number of non-cancer patients without cervical cancer is known as True Negative (TN). The number of patients classified as cancer patients without cervical cancer is known as false-negative (FN).

Table 5

Confusion Matrix

|  | | Predicted | |
| --- | --- | --- | --- |
| 0  Negative | 1  Positive |
| Actual | 0  Negative | True Negative  (TN) | False Positive  (FP)  Type I Error |
| 1  Positive | False Negative  (FN)  Type II Error | True Positive  (TP) |

Accuracy: the proportion of correctly predicted instances to total instances.

Precision (positive predicted rate): the proportion of patients with cervical cancer and patients predicted by an algorithm to be cancer patients.

Sensitivity (recall or True positive rate): the proportion of individuals predicted to be cancer patients and the actual total number of cancer patients.

Specificity: the proportion of individuals without cervical cancer and the total number of non-cancer individuals.

F1 score: the harmonic mean between precision and sensitivity.

From the metrics table{6}, it is clear that among implemented algorithms, DT, RF, CV, and XGB are so far the most effective models with good performance. These algorithms will be used for further data mining methods, including several combinations of resampling data techniques (Border-SMOTE, Border-SMOTE SVM, ADASYN) and feature selections (Select from Model and Univariate Feature Selection)

As expected, the precision and sensitivity of most algorithms before utilizing any resampling methods tend to be 0% or undefined. The dataset is exceptionally imbalanced, leading to the underfitting problem. In this case, the ML models only learn that in most cases in the training sets, a patient is cancer-free; thus, they predict that almost all patients in the testing sets are also cancer-free. As a result, the accuracy was very high, up to 96%, but it is unreliable.

Table 6

Performance comparison before and after using the Border-SMOTE method

| Algorithm | Before Resampling | | | After Resampling - Border-SMOTE Method | | | Target |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Accuracy (%) | Precision (%) | Sensitivity (%) | Accuracy (%) | Precision (%) | Sensitivity (%) |
| LR | 95.35% | 0.00% | 0.00% | 70.82% | 63.95% | 76.39% | Hinselmann |
| 91.86% | 66.67% | 13.33% | 65.81% | 71.01% | 59.39% | Schiller |
| 97.67% | NaN | 0.00% | 78.77% | 72.22% | 87.25% | Cytology |
| 94.77% | NaN | 0.00% | 74.14% | 74.00% | 71.61% | Biopsy |
| KNN | 95.93% | NaN | 0.00% | 94.53% | 90.38% | 97.92% | Hinselmann |
| 91.28% | NaN | 0.00% | 92.65% | 90.34% | 96.36% | Schiller |
| 97.67% | NaN | 0.00% | 95.69% | 92.99% | 97.99% | Cytology |
| 94.77% | NaN | 0.00% | 94.70% | 91.57% | 98.06% | Biopsy |
| DT | 90.70% | 0.00% | 0.00% | 94.22% | 90.85% | 96.53% | Hinselmann |
| 81.98% | 16.67% | 26.67% | 91.05% | 91.52% | 91.52% | Schiller |
| 92.44% | 0.00% | 0.00% | 95.08% | 92.36% | 97.32% | Cytology |
| 90.12% | 21.43% | 33.33% | 91.59% | 90.51% | 92.26% | Biopsy |
| RF | 95.93% | NaN | 0.00% | 98.48% | 97.93% | 98.61% | Hinselmann |
| 90.12% | 25.00% | 6.67% | 94.25% | 96.82% | 92.12% | Schiller |
| 94.77% | 0.00% | 0.00% | 97.54% | 97.96% | 96.64% | Cytology |
| 94.77% | NaN | 0.00% | 96.26% | 99.31% | 92.90% | Biopsy |
| ANN | 95.93% | NaN | 0.00% | 55.02% | 49.27% | 93.75% | Hinselmann |
| 91.28% | NaN | 0.00% | 81.15% | 87.86% | 74.55% | Schiller |
| 97.67% | NaN | 0.00% | 87.69% | 81.14% | 95.30% | Cytology |
| 94.77% | NaN | 0.00% | 88.47% | 89.33% | 86.45% | Biopsy |
| SVM | 95.93% | NaN | 0.00% | 64.74% | 60.94% | 54.17% | Hinselmann |
| 91.28% | NaN | 0.00% | 61.66% | 74.19% | 41.82% | Schiller |
| 97.67% | NaN | 0.00% | 76.31% | 68.56% | 89.26% | Cytology |
| 94.77% | NaN | 0.00% | 69.16% | 65.73% | 75.48% | Biopsy |
| CV | 95.93% | NaN | 0.00% | 96.66% | 93.46% | 99.31% | Hinselmann |
| 91.28% | NaN | 0.00% | 94.25% | 94.01% | 95.15% | Schiller |
| 97.67% | NaN | 0.00% | 96.62% | 94.81% | 97.99% | Cytology |
| 94.77% | NaN | 0.00% | 96.88% | 95.60% | 98.06% | Biopsy |
| NB | 84.88% | 4.76% | 14.29% | 56.53% | 50.17% | 100.00% | Hinselmann |
| 80.81% | 23.53% | 53.33% | 61.02% | 71.72% | 43.03% | Schiller |
| 10.47% | 1.30% | 50.00% | 54.46% | 50.17% | 99.33% | Cytology |
| 85.47% | 13.64% | 33.33% | 65.42% | 58.46% | 98.06% | Biopsy |
| XGB | 95.35% | 0.00% | 0.00% | 97.26% | 97.20% | 96.53% | Hinselmann |
| 88.37% | 22.22% | 13.33% | 93.93% | 95.63% | 92.73% | Schiller |
| 96.51% | 0.00% | 0.00% | 96.92% | 97.28% | 95.97% | Cytology |
| 94.77% | 50.00% | 11.11% | 95.95% | 98.63% | 92.90% | Biopsy |

The problem of underfitting was resolved by using the Border-SMOTE Method. Even though some methods may not perform as well as others, the precision and sensitivity increase significantly. So far, the result yielded by the Random Forest model for the Hinselmann target is the most outstanding, with 98.48% accuracy, 97.93% precision, and 98.61% sensitivity. However, since all 21 features were used to feed the model, there may be a problem of overfitting, which means the model was learning all the noise and irrelevant risk factors of this particular dataset rather than the actual underlying relationships, leading to potential poor performance on other datasets. This will be addressed by several feature selection methods proposed in Section 5.

**5. Results**

The model's evaluation metrics are accuracy, precision, sensitivity, specificity, and F-1 score. As mentioned, because the given dataset has a problem with imbalance, in which the number of target 1 (positive patients) takes up less than 9% of the total sample, accuracy cannot be used as the sole criterion for evaluating the model's performance. As a result, precision and sensitivity will be paid particular attention to and will be crucial in determining whether a case is true positive. Here are the results of our tests after conducting three resampling methods: Border-SMOTE, Border-SMOTE SVM, and ADASYN with two feature selection methods: Univariate Selection and Select from Model. Firstly, Univariate Selection with Select from Model approaches will be compared with the Border-SMOTE Resampling method. This comparison aims to observe which selection method yields almost as high accuracy as without conducting feature selection, but fewer features are selected.

Table 7

Borderline-SMOTE + Univariate Selection Performance

| Borderline-SMOTE + Univariate Selection | | | | | |
| --- | --- | --- | --- | --- | --- |
| Model | Metrics | Target (Number of Features Selected) | | | |
| Hinselmann (4) | Schiller (14) | Cytology (14) | Biopsy (18) |
| DT | Accuracy | 97.57% | 92.65% | 96.62% | 95.02% |
| Precision | 97.22% | 91.28% | 96.62% | 96.03% |
| Recall | 97.22% | 95.15% | 95.97% | 93.55% |
| RF | Accuracy | 97.26% | 95.21% | 97.23% | 96.57% |
| Precision | 98.56% | 96.30% | 98.61% | 98.00% |
| Recall | 95.14% | 94.55% | 95.30% | 94.84% |
| CV | Accuracy | 89.67% | 91.69% | 94.77% | 95.33% |
| Precision | 89.29% | 92.12% | 93.42% | 92.68% |
| Recall | 86.81% | 92.12% | 95.30% | 98.06% |
| XGB | Accuracy | 97.57% | 94.57% | 96.31% | 96.57% |
| Precision | 98.57% | 96.25% | 95.97% | 98.00% |
| Recall | 95.83% | 93.33% | 95.97% | 94.84% |

Table 8

Borderline-SMOTE + Select from Model Performance

| Borderline-SMOTE + Select From Model | | | | | |
| --- | --- | --- | --- | --- | --- |
| Model | Metrics | Target (Number of Features Selected) | | | |
| Hinselmann (8) | Schiller (7) | Cytology (6) | Biopsy (8) |
| DT | Accuracy | 93.62% | 90.73% | 93.23% | 91.90% |
| Precision | 91.76% | 88.39% | 94.05% | 90.85% |
| Recall | 95.71% | 92.57% | 92.94% | 93.13% |
| RF | Accuracy | 97.57% | 96.17% | 96.62% | 96.57% |
| Precision | 99.36% | 96.58% | 98.77% | 100.00% |
| Recall | 95.71% | 95.27% | 94.71% | 93.13% |
| CV | Accuracy | 98.18% | 88.50% | 97.54% | 94.70% |
| Precision | 98.76% | 92.42% | 97.65% | 96.73% |
| Recall | 97.55% | 82.43% | 97.65% | 92.50% |
| XGB | Accuracy | 96.66% | 93.93% | 96.31% | 95.02% |
| Precision | 96.91% | 93.88% | 98.17% | 96.75% |
| Recall | 96.32% | 93.24% | 94.71% | 93.13% |

Except for results yielded for the Hinselmann test, which requires only four attributes, all other tests using the ‘Select from Model’ method require significantly fewer attributes. Since this research aims to build a model for early cancer detection that is accessible and easy for immediate usage, significantly less required attributes will increase the chance of correct input from patients. After several tests, the authors observed that the ‘Select from Model’ method performs better than the Univariate method for Hinselmann, Schiller, and Cytology targets. Even though the method does not perform as well as the Univariate method for Biopsy targets in terms of accuracy and sensitivity, it yields better precision with significantly fewer attributes (11 attributes). This is a good trade-off to make for this research. Therefore, ‘Select from Model’ will be deployed for the final models.

Table 9

Borderline-SMOTE SVM + Select from Model Performance

| Borderline-SMOTE SVM + SelectFromModel | | | | | |
| --- | --- | --- | --- | --- | --- |
| Model | Metrics | Target (Number of Features Selected) | | | |
| Hinselmann (9) | Schiller (7) | Cytology (6) | Biopsy (8) |
| DT | Accuracy | 94.03% | 84.84% | 93.80% | 90.80% |
| Precision | 93.68% | 71.95% | 88.89% | 86.84% |
| Recall | 89.90% | 80.82% | 90.14% | 83.54% |
| RF | Accuracy | 96.64% | 91.80% | 97.93% | 95.60% |
| Precision | 97.87% | 85.33% | 97.14% | 97.22% |
| Recall | 92.93% | 87.67% | 95.77% | 88.61% |
| CV | Accuracy | 96.27% | 93.44% | 95.87% | 94.80% |
| Precision | 96.84% | 93.85% | 95.52% | 94.59% |
| Recall | 92.93% | 83.56% | 90.14% | 88.61% |
| XGB | Accuracy | 95.15% | 90.16% | 97.11% | 94.80% |
| Precision | 94.79% | 84.51% | 93.24% | 94.59% |
| Recall | 91.92% | 82.19% | 97.18% | 88.61% |

Table 10

ADASYN + Select from Model Performance

| ADASYN + SelectFromModel | | | | | |
| --- | --- | --- | --- | --- | --- |
| Model | Metrics | Target (Number of Features Selected) | | | |
| Hinselmann (7) | Schiller (7) | Cytology (6) | Biopsy (6) |
| DT | Accuracy | 96.05% | 90.28% | 93.21% | 92.02% |
| Precision | 96.91% | 88.57% | 95.63% | 92.90% |
| Recall | 95.15% | 93.37% | 91.07% | 91.81% |
| RF | Accuracy | 97.57% | 94.04% | 95.68% | 93.56% |
| Precision | 99.37% | 95.09% | 96.95% | 95.73% |
| Recall | 95.76% | 93.37% | 94.64% | 91.81% |
| CV | Accuracy | 93.92% | 86.21% | 91.98% | 88.65% |
| Precision | 95.60% | 90.67% | 89.89% | 93.51% |
| Recall | 92.12% | 81.93% | 95.24% | 84.21% |
| XGB | Accuracy | 96.35% | 93.42% | 96.60% | 95.09% |
| Precision | 96.93% | 93.41% | 98.76% | 97.55% |
| Recall | 95.76% | 93.98% | 94.64% | 92.98% |

With careful observation and consideration of the trade-off of all matrixes, the authors conclude that resampling using the Borderline-SMOTE method outstands other methods for this dataset. Hence, models with the best performance for each target are summarized and described (see table 11).

Table 11

Models with best performance for each target

| Test | Model | No of Features | Accuracy | Precision | Sensitivity | Specificity | F-1 |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Hinselmann | Grid Search CV + Borderline-SMOTE + SelectFromModel | 9 | 98.18% | 98.76% | 97.55% | 98.80% | 98.15% |
| Schiller | Random Forest + Borderline-SMOTE + SelectFromModel | 7 | 96.17% | 96.58% | 95.27% | 96.97% | 95.92% |
| Citology | Grid Search CV + Borderline-SMOTE + SelectFromModel | 6 | 97.54% | 97.65% | 97.65% | 97.42% | 96.74% |
| Biopsy | Random Forest + Borderline-SMOTE + SelectFromModel | 7 | 96.57% | 100.00% | 93.13% | 100.00% | 96.44% |

Fig. 2. Feature Importances, Hinselmann Target

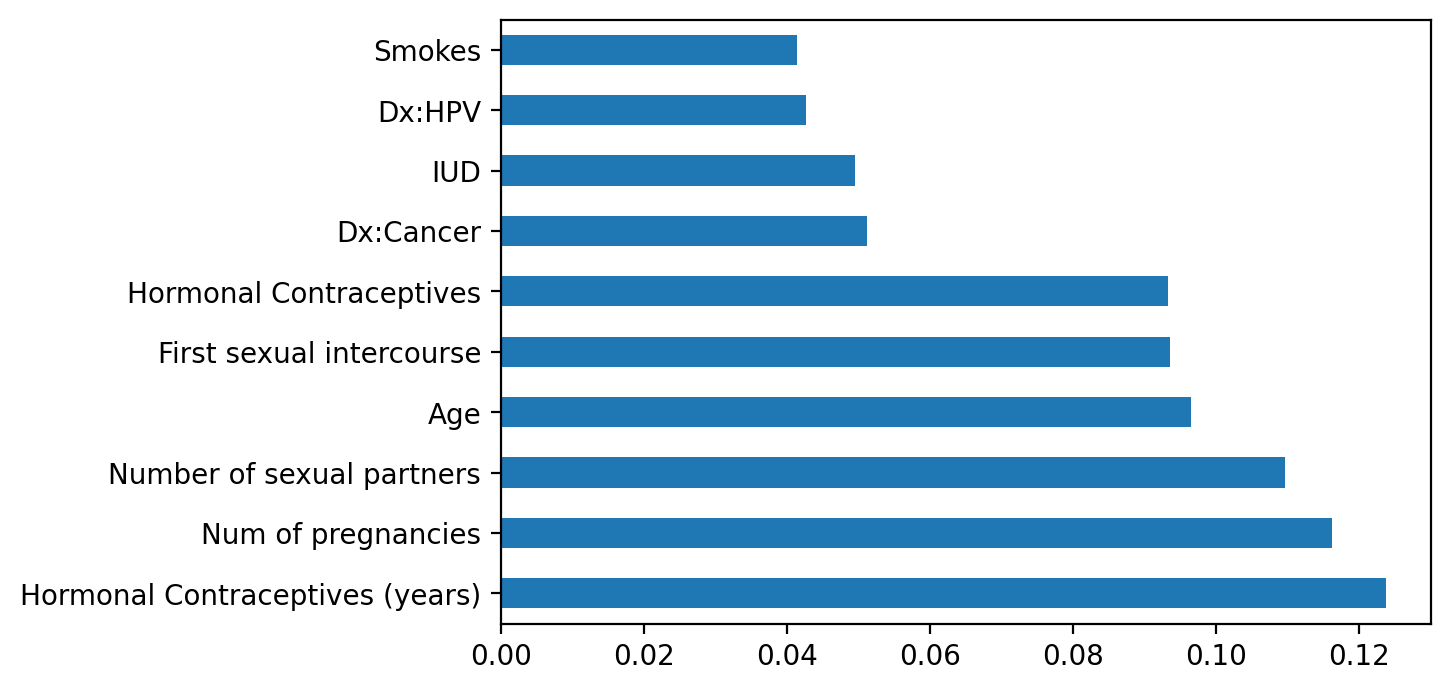


Fig. 3. Feature Importance, Schiller Target



Fig. 4. Feature Importances, Citology Target

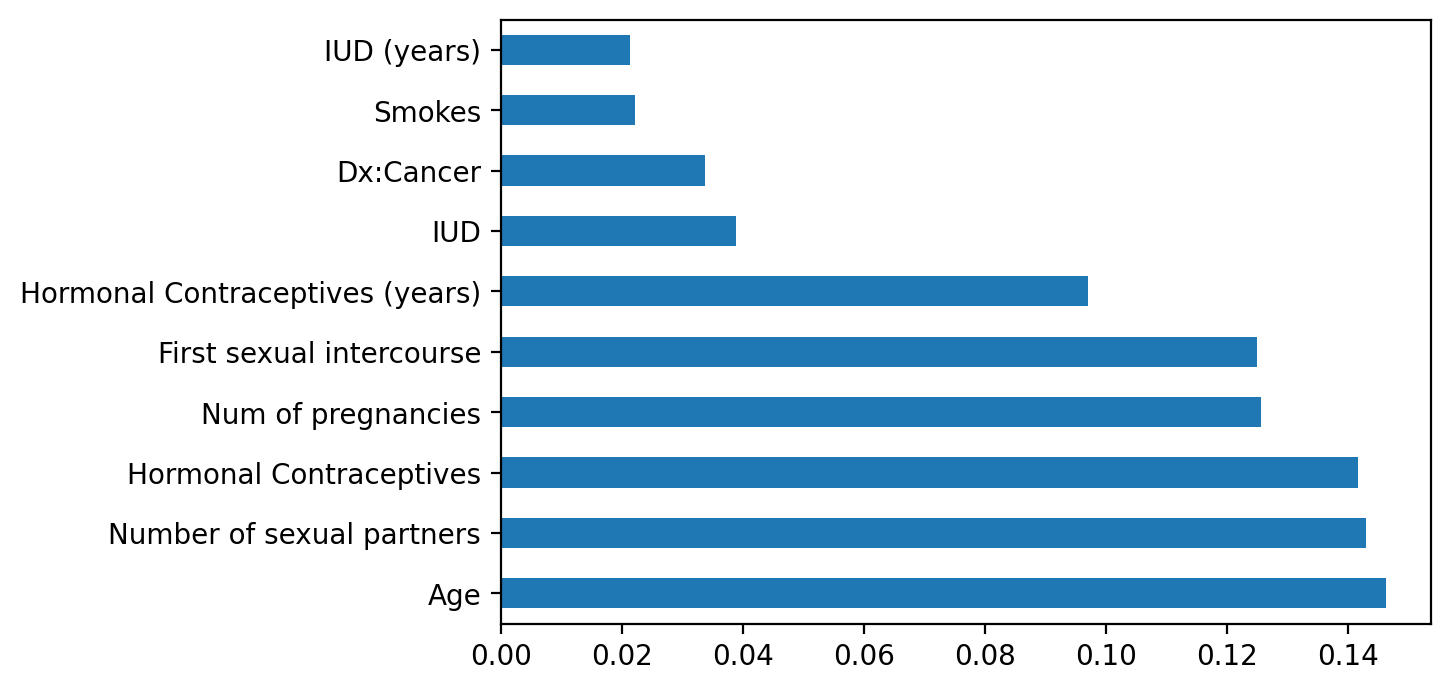
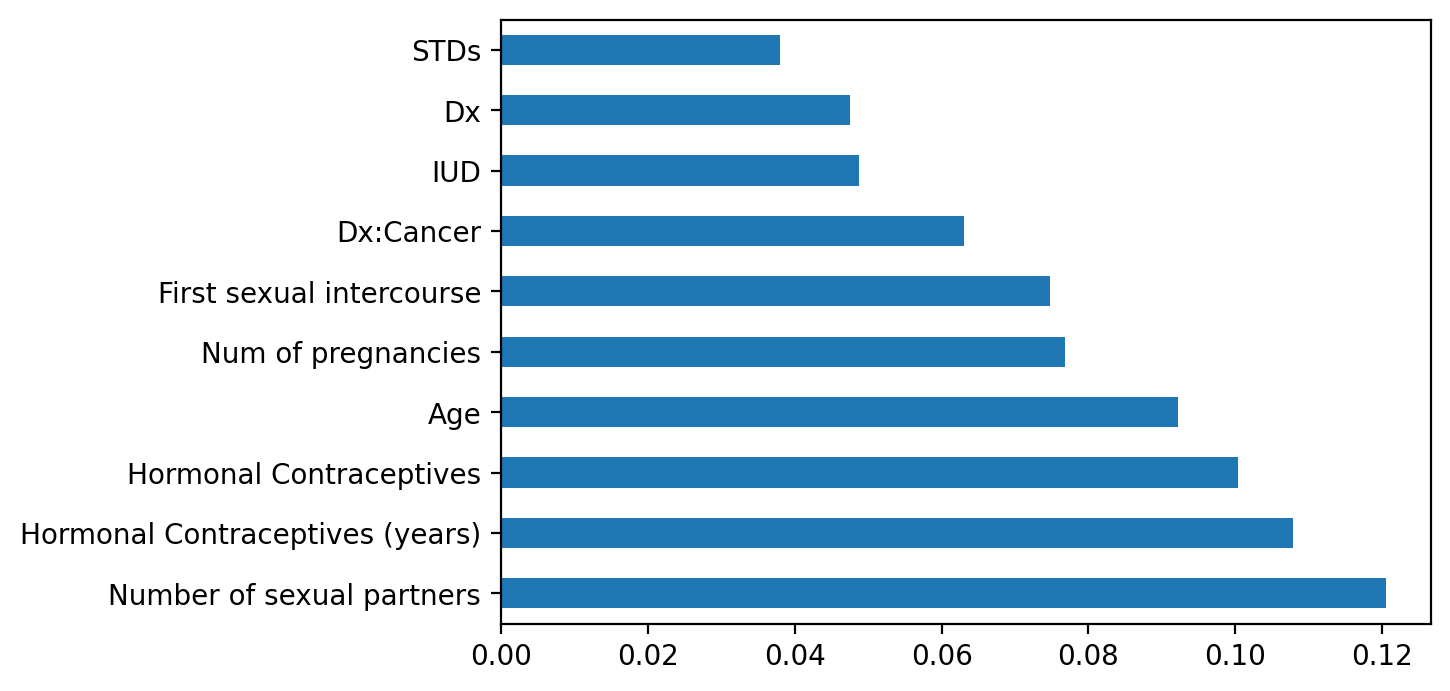


Fig. 5. Feature Importances, Biopsy Target



**6. Discussion and Conclusion**

In this research, various ML techniques have been deployed to increase cervical detection reliability of cervical cancer, supplementing previous findings, especially with feature engineering and imbalanced dataset solution. In terms of Machine Learning predictive algorithms, Grid Search Cross-Validation and Random Forest yielded the best performance, along with the support of proper feature selection utilized Meta-transformer for selecting features based on importance weights and Border-SMOTE resampling method. The target with the highest accuracy of 98.18% was Hinselmann with 97.55% sensitivity, 98.80% specificity, and nine features fed to the model. Several potential implications of this paper can be identified below:

* Since cervical cancer detection typically requires bloodwork and is processed with healthcare experts and practitioners, it can appear costly and not accessible to everyone. By building a model learning the underlying relationship between attributes that will affect the chance of having cervical cancer of woman, future developers can utilize the finding in building a cervical cancer tracking website in which users can put their information and quickly get an answer about whether they should pay particularly more attention to their current cervical health.
* Using feature selection, the authors could understand the most critical risk factors that may increase the chance of a woman having cervical cancer. This is also a pivotal point to help more women be aware of their current risks based on the dataset collected, helping more people increase their awareness about this disease and better care for their health.

References

| [1] | “Cancer.” *World Health Organization*, 3 Feb. 2022, www.who.int/news-room/fact-sheets/detail/cancer |
| --- | --- |
| [2] | “Basic Information About Cervical Cancer | CDC.” *Centers for Disease Control and Prevention*, www.cdc.gov/cancer/cervical/basic\_info. Accessed 13 June 2022. |
| [3] | Mehmood, Mavra, et al. “Machine Learning Assisted Cervical Cancer Detection.” *Frontiers in Public Health*, vol. 9, 2021. *Crossref*, doi: 10.3389/fpubh.2021.788376. |
| [4] | Parikh, Dhwaani, and Vineet Menon. “Machine Learning Applied to Cervical Cancer Data.” *International Journal of Mathematical Sciences and Computing*, vol. 5, no. 1, 2019, pp. 53–64. *Crossref*, https://doi.org/10.5815/ijmsc.2019.01.05. |
| [5] | Fatlawi, Hayder. “Enhanced Classification Model for Cervical Cancer Dataset Based on Cost Sensitive Classifier.” *International Journal of Computer Techniques*, vol. 4, no. 4, 2017, pp. 115–20, www.ijctjournal.org/Volume4/Issue4/IJCT-V4I4P19.pdf. |
| [6] | Wu, Wen, and Hao Zhou. “Data-Driven Diagnosis of Cervical Cancer With Support Vector Machine-Based Approaches.” *IEEE Access*, vol. 5, 2017, pp. 25189–95. *Crossref*, https://doi.org/10.1109/access.2017.2763984. |
| [7] | Arora, Mamta, et al. “Data Driven Prognosis of Cervical Cancer Using Class Balancing and Machine Learning Techniques.” *EAI Endorsed Transactions on Energy Web*, vol. 7, no. 30, 2018. *Crossref*, https://doi.org/10.4108/eai.13-7-2018.164264. |
| [8] | Chaudhuri, Avijit Kumar, et al. “A Multi-Stage Approach Combining Feature Selection with Machine Learning Techniques for Higher Prediction Reliability and Accuracy in Cervical Cancer Diagnosis.” *International Journal of Intelligent Systems and Applications*, vol. 13, no. 5, 2021, pp. 46–63. *Crossref*, https://doi.org/10.5815/ijisa.2021.05.05. |
| [9] | Sagala, Noviyanti T. M. “A Comparative Study of Data Mining Methods to Diagnose Cervical Cancer.” *Journal of Physics: Conference Series*, vol. 1255, no. 1, 2019, p. 012022. *Crossref*, https://doi.org/10.1088/1742-6596/1255/1/012022. |
| [10] | Sharma, Manoj. “Cervical Cancer Prognosis Using Genetic Algorithm and Adaptive Boosting Approach.” *Health and Technology*, vol. 9, no. 5, 2019, pp. 877–86. *Crossref*, https://doi.org/10.1007/s12553-019-00375-8. |
| [11] | Al-Wesabi, Y. M. S., et al. “Classification of Cervical Cancer Dataset.” *IISE Annual Conference*, 2018, arxiv.org/ftp/arxiv/papers/1812/1812.10383.pdf. |
| [12] | Fusco, Eugenio et al. “History of colposcopy: a brief biography of Hinselmann.” *Journal of prenatal medicine*, vol. 2, no. 2, 2008, pp. 19-23, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3279084. |
| [13] | Heatley, M. K. “A Critical Evaluation of the Use of the Schiller Test in Selecting Blocks from the Uterine Cervix in Suspected Intraepithelial Neoplasia.” *The Ulster Medical Jouraul*, vol. 64, no. 2, 1995, pp. 147–50, www.ncbi.nlm.nih.gov/pmc/articles/PMC2448516/pdf/ulstermedj00068-0038.pdf. |
| [14] | “Cervical Cancer Screening.” *Memorial Sloan Kettering Cancer Center*, 14 June 2022, www.mskcc.org/cancer-care/types/cervical/risk-prevention-screening#:%7E:text=Your%20doctor%20can%20use%20a%20Pap%20smear%20or,mucus%20and%20cells%20by%20gently%20scraping%20your%20cervix. |
| [15] | S. B, Kaveri, and Shikha Khandelwal. “Role of PAP Smear N Cervical Biopsy in Unhealthy Cervix.” *Journal of Scientific and Innovative Research*, vol. 4, no. 1, 2015, pp. 4–9. Crossref, https://doi.org/10.31254/jsir.2015.4102. |
| [16] | Naqa, Issam el, and Martin J. Murphy. “What Is Machine Learning?” *Machine Learning in Radiation Oncology*, 2015, pp. 3–11. Crossref, https://doi.org/10.1007/978-3-319-18305-3\_1. |