EAI Endorsed Transactions
on Internet of Things

Research Article EAI.EU

A Comparative Study on Machine Learning Classifiers for Cervical Cancer Prediction: A Predictive Analytic Approach

Khandaker Mohammad Mohi Uddin[1*], Iftikhar Ahammad Sikder[2] and Md. Nahid Hasan[2]

[1]Department of Computer Science and Engineering, Southeast University, Dhaka 1208, Bangladesh
[2] Department of Computer Science and Engineering, Dhaka International University, Dhaka1205, Bangladesh.

Abstract

INTRODUCTION: Cervical cancer is a significant global health concern, particularly in underdeveloped nations where preventive healthcare measures are limited. Early identification of the risks associated with cervical cancer is essential for both prevention and treatment.
OBJECTIVES: In recent years, machine-learning algorithms have gained popularity as potential techniques for determining a person's risk of developing cancer based on demographic and medical information. This study uses a dataset that contains patient demographics, clinical history, and results from diagnostic tests to examine how machine learning-based algorithms can be used to predict the risks of cervical cancer.
METHODS: Various machine learning approaches are used to create predictive systems, including Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), K-Nearest Neighbors (KNN), Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), Nearest Centroid (NC), Multilayer Perceptron(MP), and AdaBoost (AB).
RESULTS: The prediction capability of these models is assessed using performance metrics such as accuracy, sensitivity, specificity, f-measure, precision, and area under the receiver operating characteristic curve (AUC-ROC). Our results show that the decision tree has the highest accuracy, precision, and f1-score (98.91%, 97.81%, and 0.9889). Additionally, model performance was optimized by the use of hyperparameter tuning. After hyperparameter adjustment, the Support Vector Machine (SVM) showed superior accuracy of 99.64%, precision of 99.26%, and an F1-score of 0.9963, thereby indicating its potential in cervical cancer probability prediction. We also created a web application that uses a machine-learning model to estimate the risk of cervical cancer.
CONCLUSION: The findings of this study highlight the significance of SVM and demonstrate the potential and capabilities of machine learning techniques to enhance accurate prediction and patient outcomes for cervical cancer screening.

Keywords: Cervical cancer, Machine learning algorithms, Early diagnosis, Hyperparameter tunning, Support Vector Machine (SVM).

Received on 01 06 2024, accepted on 18 11 2024, published on 19 11 2024

doi: 10.4108/eetiot.6223

[*] Corresponding author. Email: jilanicsejnu@gmail.com

1. Introduction

Cervical cancer poses a significant risk to public health in situations where women have limited access to comprehensive medical treatment [1]. If the tissue cells around the cervix expand and multiply uncontrollably without adhering to the normal cell division route, cervical cancer, another term for the carcinogenic tumor, may form [2]. Cervical cancer continues to pose a serious threat to international healthcare systems in spite of notable advancements in diagnosis and treatment techniques. According to the World Health Organization's (WHO) most recent data, over 90% of cervical cancer-related deaths happened in low- or middle-income nations in 2018, and there were around 570,000 new instances of cervical cancer detected worldwide [3]. Smoking has been found to be a substantial risk factor for cervical cancer, increasing the likelihood of developing the illness [4]. Additionally, a disproportionately high percentage of cervical cancer occurrences globally are caused by HIV, underscoring the need for improved healthcare systems and preventative measures [5]. The ability of machine learning models to enhance diagnosis, prognosis evaluation, and treatment planning across a number of medical specialties has drawn a lot of interest in recent years [6]. The goal of this endeavor is to assist in the advanced prediction of cervical cancer by utilizing machine learning models. We want to identify the most effective models and characteristics for improving patient outcomes and diagnosis accuracy using Kaggle cervical cancer datasets [7].

Researchers employed a range of attributes and data to forecast cervical cancer. Ratul et al. [8] applied eleven machine learning approaches to identify the early risk of cervical cancer with a dataset of the UCI library of machine learning. Multi-Layer Perceptron (MLP) method was used to anticipate early threats and reach the best accuracy of accuracy of 93.33%. Bhavani et al. [9] used the dataset from UCI to predict cervical cancer. They tried numerous classification and ensemble strategies, including SMOTE and RFERF, and they came to the conclusion that the ensemble method was working better. From their analysis, they found that the Bagging Decision Tree had the greatest result in terms of accuracy, achieving 91.20%, with a sensitivity of 95%. In another study, Pramanik, Rishav, et al. [10] used a fuzzy distance-based ensemble of deep-learning algorithms to identify cervical cancer. They employed three transfer learning algorithms: Inception ResNet V2, Inception V3, and MobileNet V2, with added layers for learning data-specific features and achieving accuracy. After adopting an ensemble technique, the accuracy is 96.96%, which is greater than above each of the models.

Ali, Md Shahin, et al [11], applied an ensemble machine learning classifier to diagnose cervical cancer on two separate datasets in medical records. The findings of the study showed that the ensemble technique beat other various models in accuracy, recall, precision, and f-measure. Current approaches were surpassed by the ensemble algorithm, with accuracies of 98.06% and 95.45% for datasets 1 and 2, accordingly. Pacal, Ishak, et al. [12] integrate vision transformer (ViT) and convolutional neural network (CNN) techniques to produce an upgraded diagnosis system. Max-voting is an ensemble learning technique for visual transformer models, with the greatest classification performance rate, reaching 92.95% accuracy, and 93.30% f1-score. Ilyas et al. [13] proposed an ensemble strategy to predict cervical cancer, which integrated multiple machine-learning classifiers. Additionally, it claims that this ensemble method attained a maximum accuracy of 94%, surpassing the performance of individual classifiers tested on the same datasets.

The important contributions of our study are as follows,

a) To enhance the standard and usefulness of raw data for analysis, preprocessing uses a variety of techniques, such as feature selection, null value handling, and imbalance data handling.
b) Null values were handled by filling them with the mean, assuring the missing data points were substituted with a representative value to protect the integrity of the dataset through analysis.
c) To improve the prediction performance of our cervical cancer diagnostic models, we used SelectKBest in combination with XGBoost for feature selection.
d) By boosting minority class representation, Random Over-sampling improves machine learning algorithm performance and equity while addressing class imbalance.
e) To find the best-performing model to predict cervical cancer, a variety of machine learning models are utilized, including Support Vector Machine, K-Nearest Neighbors, Random Forest, Naïve Bayes, Decision Tree, Nearest Centroid, Logistic Regression, Gradient Boosting, Multilayer Perceptron, and Adaboost.
f) By enhancing accuracy, hyperparameter tuning is used to improve performance.
g) The attainment of greater accuracy in comparison to the mentioned research work and ML-based web application to predict cervical cancer represents a significant contribution.

2. Methodology

The datasets used in this investigation were gathered from the Kaggle repository [7]. The dataset is first subjected to a number of data preparation techniques, such as feature selection, random oversampling, and null value management. Then, in order to train and test models appropriately, the dataset was split into 80% and 20%. Several supervised machine-learning models were used to determine best best-performing model. Additionally, hyperparameter tuning is used to optimize accuracy and improve model performance.

To determine accuracy, recall, precision, and F-measure, the models are evaluated on testing data after being trained on patterns from the training data. Additionally, machine learning techniques enhanced the understanding of the system's organizational structure and the relationships between its various components. Figure 1 illustrates the main architecture of the proposed model.

2.1. Dataset Overview

For this study, we chose datasets from Kaggle repositories [7], a publicly available dataset of 835 patient samples in 36 characteristics, comprising 54 cancer patients and 781 non-cancerous persons [7]. The values in the "Biopsy" column are set to 1 for cervical cancer and 0 for healthy. The bulk of this dataset consists of a number of patient history characteristics, including age, number of years smoked, age of first sexual encounter, usage of hormonal contraceptives, number of pregnancies, presence of STDs, and 30 more factors. Therefore, machine learning models might be used to assess the data in order to determine the significance and strength of the correlations between patient features and the diagnosis of cervical cancer.

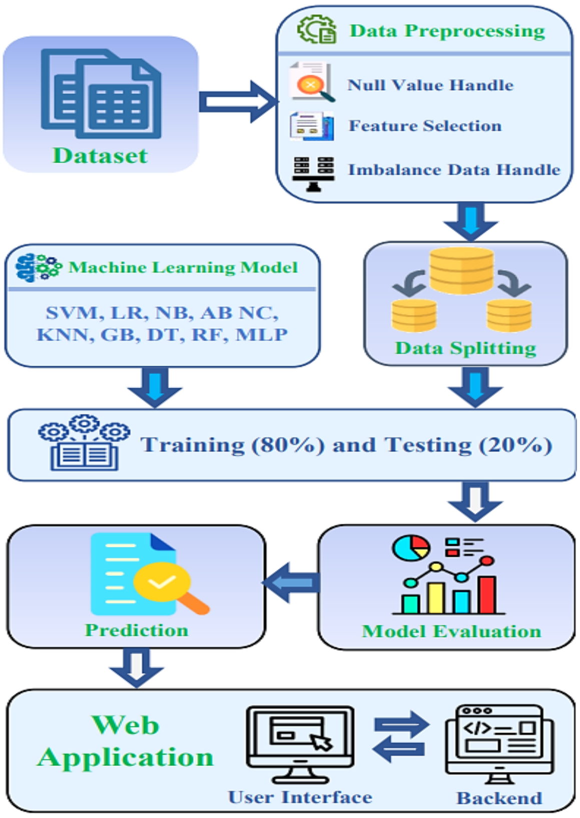


Figure 1. Proposed Methodology of our system

2.2. Data Preprocessing

Preprocessing is a crucial step in data analysis and machine learning that involves organizing, cleaning, and converting raw data into a format required for study or model training. Preprocessing aims to address imbalanced data, handle missing data, identify pertinent features, and eventually enhance the overall quality of the data.

2.2.1. Missing Value Handle

Null values must be handled effectively during the preprocessing of data in order to offer suitable analysis while preventing bias and ensuring data integrity [14]. Mean, median, and mode are common options for replacing missing data. In the current study, we employed mean to manage null values because of its simplicity and effectiveness in maintaining data integrity.

2.2.2. Feature Selection

For the present cervical cancer dataset, a comprehensive array of 36 characteristics has been provided [7]. For easier feature selection, we used a methodology that combined the SelectKBest and XGBoost algorithms. Below are 14 of the most important features achieved through this method. Additionally, table 1 provides a detailed summary of the selected features and targeted column.

2.2.3 Imbalanced Data Handle

In binary classification, where one class greatly outnumbers the other, imbalanced data handling is a common issue. The unbalanced data before Random Over-sampling (ROS) is shown in Figure 2(a). We have used Random Over-sampling on the dataset to balance the data since unbalanced data leads to biased models that adversely impact the minority class [22]. The success of Random Over-sampling in resolving class imbalance concerns and producing a more balanced dataset for further analysis and modeling is seen in Figure 2(b), which visualizes the class similarities in terms of their distribution with ROS.

2.3. Exploratory Data Analysis

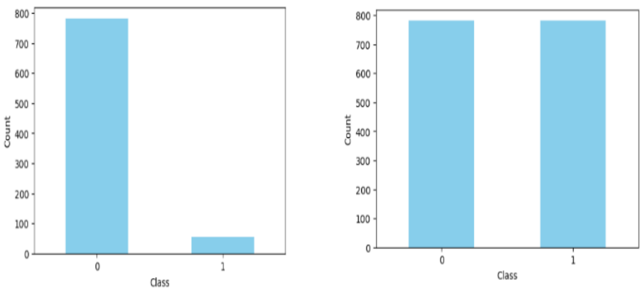
To learn more about the features, we used Boxplot to analyze the raw data. The boxplot helps identify outliers and distribution features by providing a visual representation of the dataset's central tendency, dispersion, and skewness. Histograms were created for every feature separately in order to understand the range.

The boxplots of a few key features extracted from this dataset [7] are displayed in Figure 3, providing information about their distribution and emphasizing their importance in the prediction process. The histograms for three distinct parameters are shown here. Because abnormal cells appear white and are not stained, Schiller's assay is an essential cervical cancer screening technique.

The boxplot indicates dependable qualities by demonstrating a discernible difference within classes. "Dx" refers to a diagnostic history, which may include previous cervical issues. Similar to the number of years smoked, the number of

Table 1. Feature importance score in cervical cancer detection

No	Feature’s Name	Features Scores	**Feature Description**
1	Schiller	0.6826245	Appearance of the cervix[15]
2	Dx	0.0756582	Cancer diagnosis status
3	STDs: Number of diagnosis	0.0408548	Count of STDs diagnosed [16]
4	Smokes (packs/year)	0.0307381	Packs of cigarettes smoked yearly [17]
5	Age	0.0290981	Age of the individual
6	Hormonal Contraceptives	0.0245152	Use of hormonal birth control [18]
7	STDs	0.0213901	Presence of STDs [19]
8	Citology	0.0213313	Cervical cytology results [20]
9	Num of pregnancies	0.0173793	Total number of pregnancies
10	Smokes (years)	0.0164721	Duration of smoking in years
11	First sexual intercourse	0.0146702	Age at first intercourse
12	Hinselmann	0.0131811	Presence of Hinselmann sign
13	IUD	0.0065912	Usage of an intrauterine device [21]
14	IUD (years)	0.0054959	Duration of intrauterine device use



(a) **(b)**

Figure 2. Data Balancing Using ROS

packs consumed annually is a risk indication. It is noteworthy and suggests that smoking for an extended period of time is a risk factor.

We also examine the correlation matrix to understand the relationships among the variables in a dataset. Understanding the links between the features is made easier by the correlation matrix, which includes zero, negative, and positive correlations. While negative correlation suggests opposite value changes, an increase in one is correlated with a drop in the other, and vice versa. A positive correlation indicates that parameters change together. Complete independence between variables is implied by a zero correlation. Furthermore, this thorough study advances our understanding of the intricate relationships between several factors and the occurrence of cervical cancer.

2.4. Machine Learning Classifiers

For our research, we employed a range of machine learning models [23], including Random Forest and Decision tree, due to their inherent advantages when handling complex datasets.

Because decision trees are so interpretable, we can readily comprehend and illustrate the decision-making process. Their effectiveness lies in detecting nonlinear patterns in datasets. By merging the results of each decision tree, Random Forest reduces overfitting and makes each decision tree more robust.

Furthermore, hyperparameter adjustment significantly improved the Support Vector Machine (SVM) classifier's accuracy. SVM is an effective model that works when there are more dimensions than occurrences. By adjusting parameters such as the kernel type, regularization term, and gamma, hyperparameter tuning makes it possible for the SVM algorithm to perform more predictively and interpret data more precisely [24].

2.4.1. Decision Tree

Decision trees are versatile and widely used machine learning algorithms that can do both regression and classification tasks [25]. They function by segmenting the feature region into smaller parts and evaluating each segment based on its feature values. This process creates a structure like a tree, where each leaf node represents the anticipated result and classification, and each inner node presents attributes and a judgment based on its value. Decision trees are represented mathematically as a collection of if-else conditions based on the input's characteristics.

This classifier is used as it can naturally handle both categorical and continuous data. Decision Trees can also have splits with underlying conditions for each column, for example, a threshold (i.e., numeric) with respect to continuous features such as "Age," "Num of pregnancies" and category membership such as "Hormonal Contraceptives," "IUD." Mixed data offers advantage such as class isolation and category-based splits, allowing for targeted class selection and classification of the target variable "Biopsy."

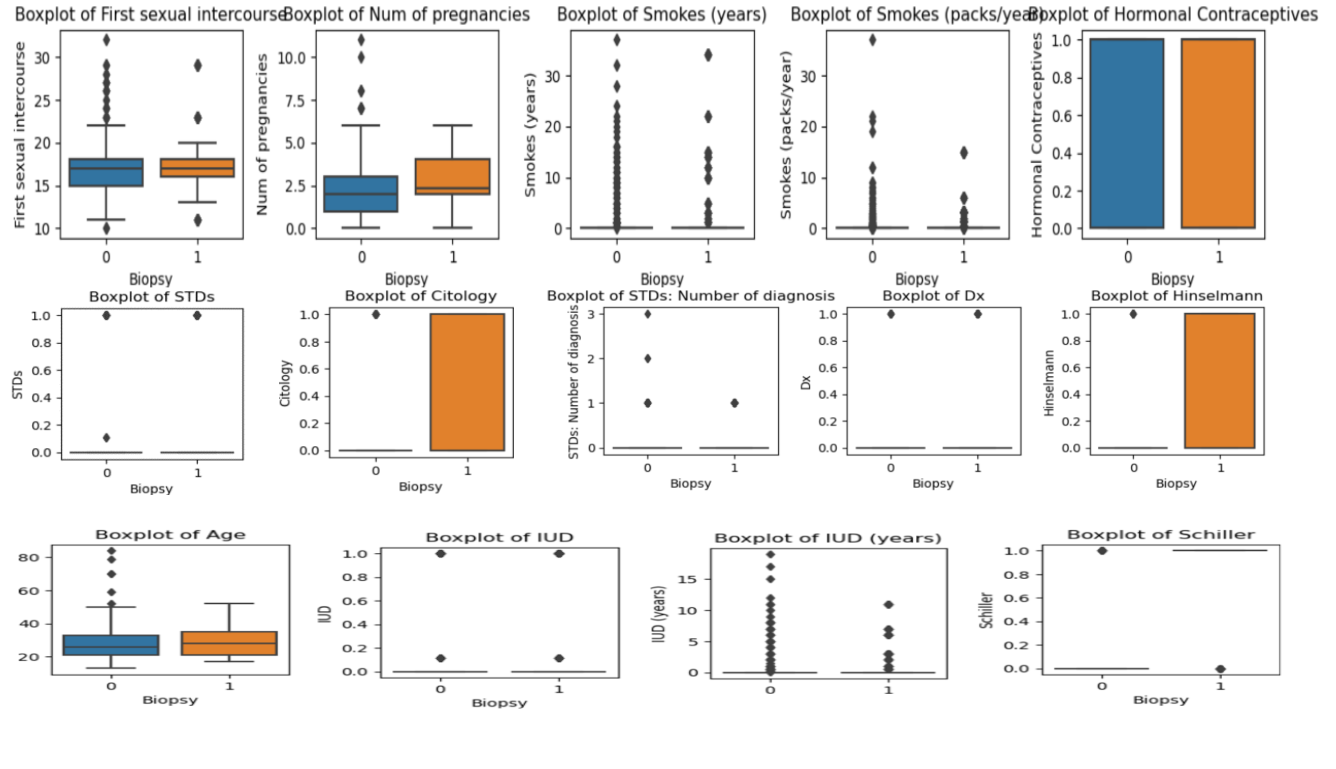


Figure 3. Sample Dataset Boxplot

Decision trees provide a natural representation of the decision process due to its tree-like structure. Each node illustrates a certain decision-based technology and the users can trace back their way from root to leaf illustrating how each feature affects the final prediction. Such accessibility helps people recognize the influential variables in determining risk behavior associated with cervical cancer and establishes confidence in the predictions made by this model. In addition, the visualization of decision trees improves communication of complicated decisions to both technical and non-technical audiences which provides an overall boost in interpretability. Decision trees are not only able to capture the relationship between feature and target variable more effectively compared to splitting boundary but by dividing the whole feature space. The tree splits the data in such a way that each leaf node corresponds to one subset of the data, meaning that it can make predictions rooted in localized information as opposed generalized assumptions This means the model is now more effective at identifying patterns and as a result makes a better prediction.

2.4.2. Random Forest

Multiple decision trees are used in the Random Forest classification model, an ensemble learning technique, to increase durability and expected performance [26]. Averaging the results of several trees training on different subsets of the given data and attributes, reduces over-fitting. Classification equations for Random Forest classifiers are defined by,

$$\hat{y}=mode(f_1(x),f_2(x),\ldots,f_B(x)) \qquad (1)$$

Here, $\hat{y}$ indicates the class label of the input instance x, and the mode is the estimated value of the b-the Random Forest decision tree given the input sample x, which is represented by the statistical mode function, $fb(x)$, where B is the total number of Random Forest decision trees.

We can see that the Random Forest model uses feature randomness, each decision tree has been trained on a random subset of features. By doing this, the trees become more diverse and this will allow the model to learn different patterns in the data. This randomness lowers the overfitting risk and improves the overall computational effectiveness of the model accordingly. Random forest is, therefore, a strong classifier that reduces variance and increases accuracy by averaging the predictions of multiple trees.

The Random Forest significantly reduced the overfitting risk for multiple reasons. One, it reduces the noise and variance that can come from big statistics by averaging together the predictions of many different decision trees trained on various parts/subsets of your upstream data. This ensemble method reduces the overall surface and therefore more generalized models. Also, the construction of each tree in the forest using a random subset of features at every split makes trees different from each other to an even greater extent. That randomness stops any one tree from becoming overly complex and memorizing noise, rather than representing the data distribution. Thus, the Random Forest method performs well in overarching bias and variance by improving predictive performance on unseen data.

2.4.3. Support Vector Machine

Support Vector Machine (SVM) is a machine learning technique that has been applied to both regression and classification tasks [27]. Finding the best hyperplane to divide feature classes is the aim. The following could be used to illustrate the linear SVM selection process:

$$f(x)=w\cdot x+b \qquad (2)$$

Here, the input feature vector is represented by *x*, the bias component is indicated by b, the decision function is represented by *f(x),* and the weights assigned to the features are represented by w. The data point belongs to one class if *f(x)* is positive, and to the other class if *f(x)* is negative.

2.5. System Specifications

The system is driven by an Intel(R) Core (TM) i5-6300U CPU clocked at 2.40GHz, which ensures consistent speed of processing. It has 16 GB of RAM, which allows for excellent multitask and application management. The Intel® HD Graphics 520 GPU manages graphical tasks and serves basic graphics requirements. The system runs on Windows 11 Pro and uses Jupyter Notebook [28] as the primary experimentation tool.

Table 2. System specification for the proposed system

Resource	**Details**
CPU	Intel(R) Core(TM) i5-6300U CPU @ 2.40GHz
RAM	16 GB
GPU	Intel® HD Graphics 520
Experimental Tool	Jupyter notebook
Operating System	Windows 11 Pro

3. Result

AB, MLP, DT, RF, LR, KNN, SVM, GB, NC, and NB were among the ML classifiers used in this study. The decision tree and random forest outperformed the feature selection methods SelectKBest and XGBoost, with respective scores of 98.91% and 98.54%. The model's performance was then optimized through hyperparameter adjustment. The 99.64% accuracy of the Support Vector Machine (SVM) model demonstrates its promise for forecasting the risk of cervical cancer.

3.1. Confusion Matrix

To further understand the performance of our model, we used a confusion matrix [29] to show our findings. We used a 2x2 confusion matrix to achieve our goal of a binary classifier. Here, TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives [30]. Accuracy, sensitivity, specificity, precision, and f-measure are all calculated with the aid of equations 3 through 7 [31]. The percentage of true positives that the mode correctly detects is measured by sensitivity. Another name for it is recall. Specificity is the proportion of true negatives that the prediction algorithm correctly detects. Precision is defined as the proportion of positive results that turned out to be truly correct. The F-measure, often called the F1 Score, is the harmonic mean of Precision and Recall. It provides a single statistic that balances precision and recall. Because they reveal details about the model's ability to accurately identify both positive and negative circumstances, sensitivity, specificity, accuracy, and F-measure are important performance indicators in the classification process.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad (3)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \qquad (4)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \qquad (5)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (6)$$

$$\text{F-Measure} = \frac{\text{TP}}{(\text{TP+0.5(FP+FN)})} \qquad (7)$$

The decision tree outperforms all other machine learning classifiers in terms of accuracy, scoring 98.91%. The high accuracy of decision tree classifiers demonstrates their ability to effectively categorize occurrences in the dataset.

Table 3 displays the model evaluations for four performance indicators: sensitivity, specificity, accuracy, and f-measure.

Table 3. Model Evaluation (before hyperparameter tunning)

Model	Sensitivity	Specificity	Precision	F-Measure
SVM	100.00%	95.71%	95.71%	97.81%
RF	100.00%	97.14%	97.10%	98.53%
KNN	100.00%	91.43%	91.78%	95.71%
DT	100.00%	97.86%	97.81%	98.89%
NB	94.78%	85.71%	86.39%	90.39%
LR	88.81%	94.29%	93.70%	91.19%
GB	100.00%	95.71%	97.71%	97.81%
NC	85.07%	92.86%	91.94%	88.37%
MLP	100.00%	96.43%	96.40%	98.17%
AB	94.03%	94.29%	94.03%	94.03%

Model assessment evaluates the effectiveness and utility of prediction models using a variety of criteria and methodologies to assure their dependability and relevance. In this cervical cancer diagnosis study, we used a number of machine learning classes. Particularly, Decision Tree (DT) and Random Forest (RF) had the greatest accuracies of

98.91% as well as 98.54%, accordingly, demonstrating their efficiency for this challenge. Although Multilayer Perceptron (MLP) and Gradient Boosting (GB) scored admirably, having accuracy levels of 98.18% and 97.81%, accordingly, other models like Support Vector Machine (SVM) or k-Nearest Neighbors (KNN) delivered comparable outcomes of 97.81% and 95.62% accuracy. These different levels of performance show the necessity of picking a proper machine-learning model for predicting cervical cancer accurately.

The accuracy of each machine learning technique is shown in Figure 4, which also provides performance metrics for various models and helpful information about how efficient they are in comparison.

The effectiveness of binary classification models is gauged by the receiver operating characteristic curve (ROC) [32]. It contrasts the False Positive and True Positive rates for each valuation variable. This curve shows how specificity and sensitivity can be balanced, enabling the best threshold selection based on the particular needs of the application. A quantitative measure derived from the ROC curve is called the Area Under the Curve (AUC) [33]. The overall discriminative power of a classification algorithm is assessed by the AUC curve (20, 40, 49). AUC values range from 0 to 1, where 1.0 denotes perfect classification and 0.5 denotes unpredictability.

3.2. ROC and AUC

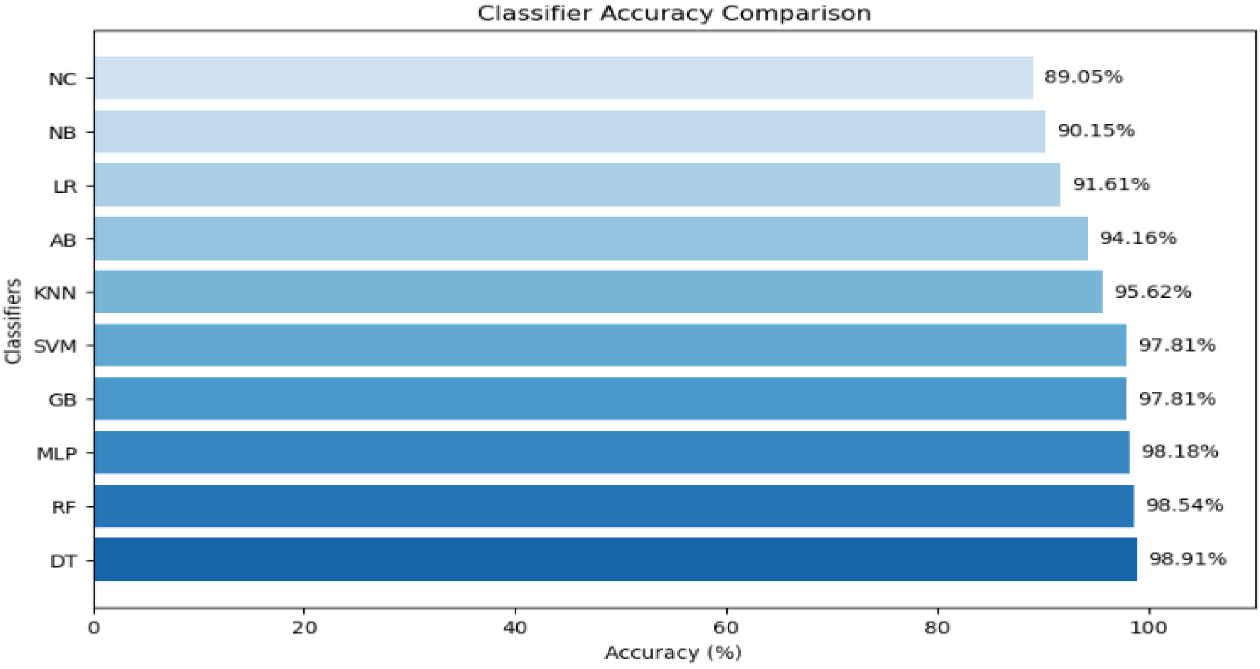


Figure 4. Comparison of different ML classifiers' accuracy before hyperparameter tuning

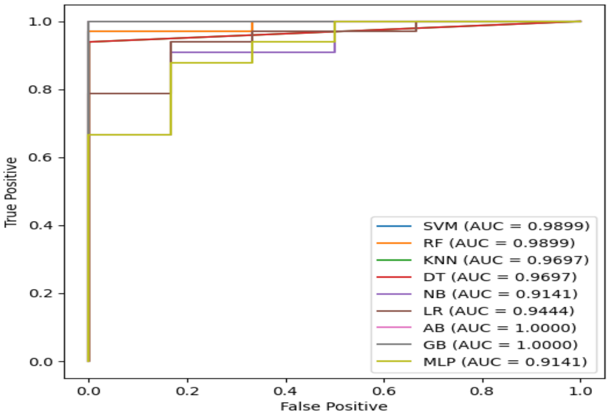


Figure 5. ML Model's evaluation based on AUC ROC curves

3.3. Hyperparameter Tuning

Hyperparameter tuning is a crucial phase in the machine-learning process that optimizes the performance of models by picking optimal hyperparameters [34]. Hyperparameters are variables that control the learning process of algorithms used for machine learning, such as regularization courage, along with the number of layers that are hidden in neural systems.

Unlike the parameters of the model, which learn from data used for training, hyperparameters are determined before the training process starts and remain fixed throughout training. Among the different methods available for hyperparameter tuning, the specific approach we employed was grid Search.

To enhance model performance through hyperparameter tuning, we employed scikit-learn's RandomizedSearchCV method. This approach systematically explores various hyperparameter combinations to identify the most effective configuration for each model. Initially, we defined our classifiers and established their respective hyperparameter grids. Each classifier was incorporated into a pipeline that standardized the data before applying the machine learning algorithm. For the Support Vector Machine (SVM), we focused on tuning three key hyperparameters: the regularization parameter C, the kernel coefficient γ, and the kernel type. The parameter C was a range between 0.1 and 100, ultimately yielding an optimal value of approximately 25.98.

For γ, which affects the degree of curvature in the decision boundary, we set options to scale, auto, and a range of values spaced logarithmically between 10−4 and 100.8 to determine an optimal value around 6.31. To simplify the process, we limit the kernel options to 'linear' and 'rbf', with the exception that the best kernel was identified as rbf.

For linearly separable data, linear kernels work which helps in mapping similarly for non-linear kernels by taking data into higher dimensions like RBF. Kernel type was chosen so that the SVM model complexity matches the dataset. This improved the interpretability of our SVM as well as reliable predictions.

Hyperparameters like C and γ in the SVM classifier significantly impact performance. By tuning these parameters, we can control model complexity and prevent overfitting. The γ parameter controls the model's sensitivity to individual data points. With each iteration of optimization, the model achieved an accuracy of 99.64%, demonstrating the significant role of hyperparameters in performance based on data nature.

Table 4. Model Evaluation (after hyperparameter tunning)

Model	Sensitivity	Specificity	Precision	F-Measure
SVM	100.00%	99.29%	99.26%	99.63%
RF	100.00%	96.43%	96.40%	98.17%
KNN	100.00%	96.43%	96.40%	98.17%
DT	100.00%	97.86%	97.81%	98.89%
LR	88.81%	94.29%	93.70%	91.19%
GB	100.00%	96.43%	96.40%	98.17%
MLP	100.00%	97.14%	97.10%	98.53%
AB	100.00%	95.00%	95.04%	97.45%
NB	94.78%	85.71%	86.39%	90.39%
NC	85.82%	95.71%	95.04%	90.20%

After hyperparameter adjustment, the Support Vector Machine's (SVM) accuracy rose dramatically to 99.64%.

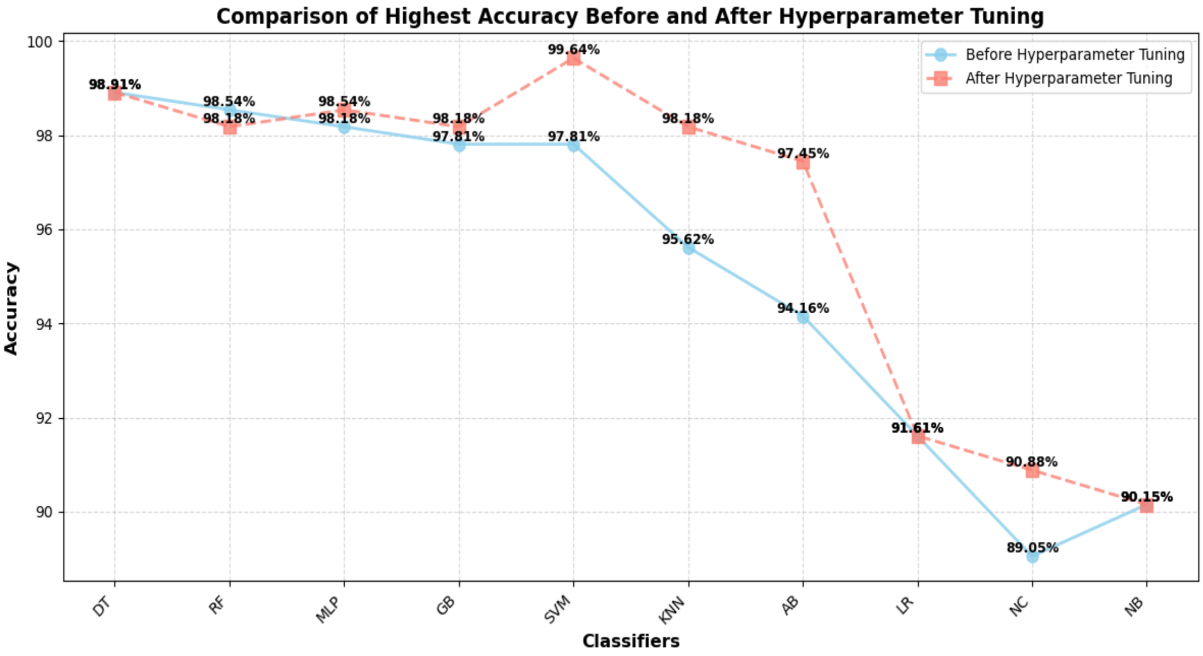


Figure 6. Effect of hyperparameter adjustment in classifier accuracy

While other classification techniques, such as K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), and Logistic Regression (LR) maintained comparable accuracy levels, the SVM's notable enhancement emphasizes how crucial hyperparameter adjustment is to optimizing classifier performance in cervical cancer diagnosis. The impact of hyperparameter modification on model accuracy is depicted in Figure 6.

Figure 7 illustrates the ROC-AUC curves of all classifiers after hyperparameter tuning. This ROC curve compares the performance of multiple classification models in distinguishing between two classes. Each curve shows the trade-off between the True Positive Rate and the False Positive Rate for a specific model. The diagonal line represents random chance (AUC = 0.5), while curves closer to the top-left corner indicate better performance. Models with higher AUC values, like Support Vector Machine and Random Forest (AUC = 1.0), demonstrate strong predictive power. Most models achieve high AUC scores, highlighting their effectiveness for this binary classification task.

4. Web Application Development

We have created a web application [35] that uses machine learning to enable users to predict their risk of developing cervical cancer. Users fill out an easy-to-use HTML-CSS form with their information. After submission, the data is routed to a Flask server, where it is processed by a machine-learning algorithm. The prediction result is then dynamically presented on the web interface as the server provides it, giving users important information about their possible health risks. The accessibility of predictive healthcare solutions is improved by the smooth integration of front-end and back-end technologies.

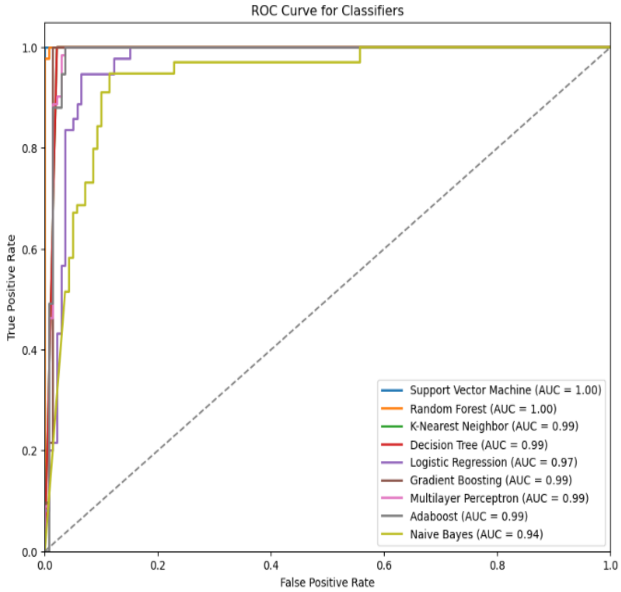


Figure 7. Classifiers evaluation based on ROC-AUC curves after hyperparameter tuning.

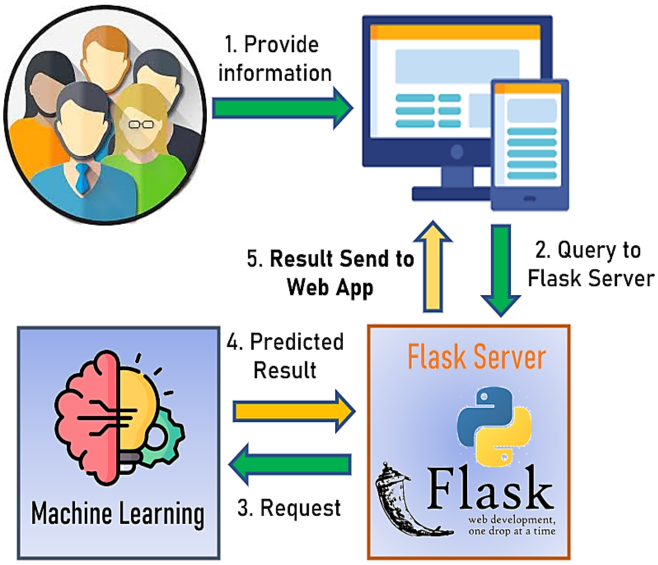


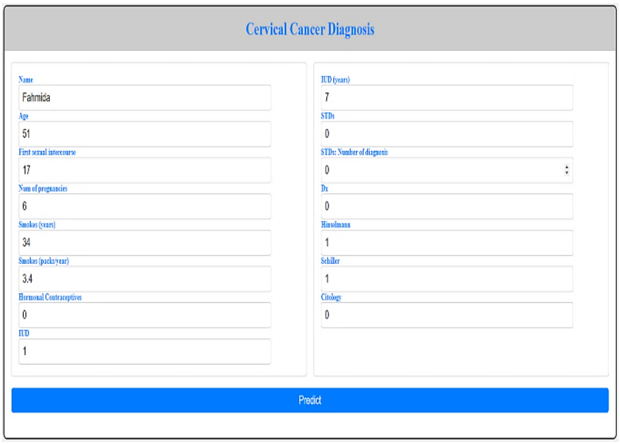
Figure 8. The workflow of the web applications.

Our web application user interface (shown in figure 9), where a woman can enter details for the prediction of cervical cancer. To enter data, one needs to know the woman's age, the base number of her first sexual contact, and the full pregnancy count range. It also asks for a woman's smoking history including how many years she has been smoking and the average number of packs smoked per year.

It also queries whether the woman has ever used hormonal contraception if she ever had an intrauterine device (IUD), and how long she has had her IUD in years. It also asks for her STD history and how many times she has been diagnosed with an STD. Diagnosis and test results the user needs to indicate if she has received a diagnosis in addition to providing the result of a variety of cervical cancer screening tests (Hinselmann, Schiller, and cytology).

When the woman enters these details, this information is forwarded to a Flask server and is passed through a hyper-parameter-tuned machine learning classifier. The classifier takes the input data and gives a prediction, with 1 meaning positive diagnosis of cervical cancer and 0 meaning negative. Prediction of cervical cancer (shown in figure 10) As per the output predicted a diagnostic report is generated which contains each and every detail of the woman along with diagnostic status/message. The report is provided in an easy-to-read format for the woman, so she can see her risk status and then be provided guidance based on the prediction results.

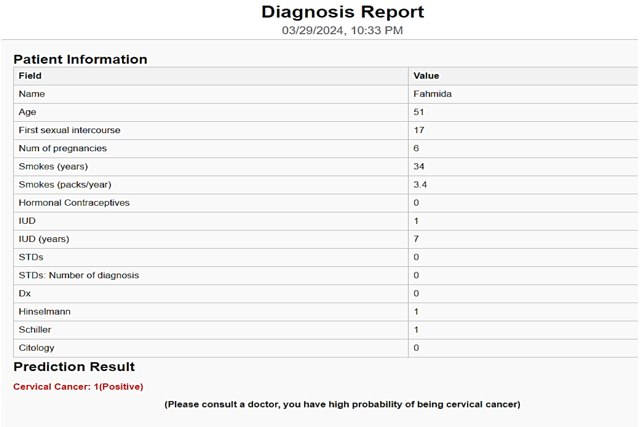
Figure 9 shows the web application's user interface allowing users to submit information, and Figure 10 shows that shows the cervical cancer prediction output.



Cervical Cancer Diagnosis

Name	Fahmida
Age	51
First sexual intercourse	17
Num of pregnancies	6
Smokes (years)	34
Smokes (packs/year)	3.4
Hormonal Contraceptives	0
IUD	1
IUD (years)	7
STDs	0
STDs: Number of diagnosis	0
Dx	0
Hinselmann	1
Schiller	1
Citology	0

Figure 9: User interface of the web application for cervical cancer diagnosis



Diagnosis Report
03/29/2024, 10:33 PM

Patient Information

Field	Value
Name	Fahmida
Age	51
First sexual intercourse	17
Num of pregnancies	6
Smokes (years)	34
Smokes (packs/year)	3.4
Hormonal Contraceptives	0
IUD	1
IUD (years)	7
STDs	0
STDs: Number of diagnosis	0
Dx	0
Hinselmann	1
Schiller	1
Citology	0

Prediction Result

Cervical Cancer: 1(Positive)

(Please consult a doctor, you have high probability of being cervical cancer)

Figure 10. Predicted result for the specific patient

5. Discussion

Our work employed various machine learning classifiers to improve the detection accuracy of cervical cancer. Random Forest, Support Vector Machines, Decision Trees, K-Nearest Neighbors, Logistic Regression, Naïve Bayes, Gradient Boosting, Nearest Centroid, Multilayer Perceptron, and AdaBoost were all used in the classification models. The decision tree has the highest accuracy of all of them, at 98.91%. Support Vector Machine (SVM) can distinguish between healthy individuals and those affected by cervical cancer, as evidenced by its 99.64% accuracy after hyperparameter tuning. These promising findings highlight the need to apply machine-learning techniques to improve healthcare outcomes and advance the field of medical diagnostics research. The development of strong and trustworthy cervical cancer detection systems may benefit from more investigation and validation of these models on various datasets.

As previously indicated, Ratul et al. [8] used machine learning (ML) algorithms to predict early threats of cervical cancer, and multi-layer perceptron (MLP) algorithms achieved the highest accuracy of 93.33%. After experimenting with a number of classification and ensemble algorithms, Bhavani et al. [9] discovered that the Bagging Decision Tree produced the best accuracy, achieving 91.20%. To detect cervical cancer, Pramanik, Rishav, et al. [10] employed a fuzzy distance-based ensemble of deep learning algorithms. The accuracy after using an ensemble technique is 96.96%. An ensemble machine learning classifier was used by Ali, Md Shahin, et al. [11] to detect cervical cancer in two datasets. The study found that the ensemble model outperformed other models in terms of f1-score, accuracy, recall, and precision. For datasets 1 and 2, the ensemble algorithm's accuracy rates were 98.06% and 95.45%, respectively. Using convolutional neural network (CNN) and vision transformer (ViT) approaches, Pacal, Ishak, et al. [12] demonstrate that max-voting produced the highest classification performance rate, with 92.95% accuracy and 93.30% f1-score. In order to predict cervical cancer, Ilyas et al. [13] developed an ensemble strategy that integrated many machine learning classifiers and had a maximum accuracy of 94%. However, our suggested model outperformed all of the previously mentioned relevant work in predicting cervical cancer, with an accuracy of 99.64% via SVM after hyperparameter adjustment. Our proposed model is contrasted with the various research performances indicated in Table 5.

6. Conclusion

A model with excellent efficiency was created to predict cervical cancer. This comparative investigation demonstrated the value of utilizing SVM and machine learning to predict patients with cervical cancer. The methods used were compared to earlier machine learning models. Future iterations of the program might include an additional dataset that analyzes spiral drawings from patients with cervical cancer in order to further improve its capabilities. The goal of this integration is to increase the model's scalability. Implementing a hybrid architecture that combines machine learning with deep learning to provide a more advanced and adaptable model is another potential development path. Furthermore, continued work might concentrate on improving feature selection to increase the general public's acceptance of this model. In the end, this model may find use in medical settings, allowing clinics or hospitals to enter test findings for the diagnosis of cervical cancer.

With a stunning accuracy percentage of 99.64%, our research demonstrates a remarkable increase in cervical cancer diagnostic accuracy utilizing a Support Vector Machine (SVM) with hyperparameter tuning. This performs better than similar recent studies. In particular, our approach outperforms a recent study that used the same dataset but reported fewer accurate results in 2024 [11]. This illustrates how effective our strategies are and how robust our framework is. To increase the quality of the data and the model's efficacy, we employed a range of preprocessing techniques. In addition to our theoretical contributions, we developed an intuitive web tool that makes reliable cervical cancer predictions. Our research makes a substantial contribution to the field of medical diagnosis, especially in the advanced detection of cervical cancer. by surpassing existing standards and providing a useful application. With more effective treatment methods, this could result in better patient outcomes.

We may consider using our prediction model to clinical practice in the future, which would enable more accurate and timely cervical cancer screening in high-risk populations. By including real-world data from clinical settings, future iterations of our work can improve the predictive model's robustness and reliability and make it more useful in real-world situations. Increase access to healthcare by facilitating personalized risk prediction with our web technology, which might be used to predict cervical cancer in distant communities in advance.

Table 5. Research Performance Comparative Analysis

Studies	Model	Accuracy	Precision	F1-Score
Ratul et al. [8]	MLP	93.33%	90.91%	90.52%
Bhavani et al. [9]	SMOTE + BDT	91.20%	95.00%	96.00%
Pramanik, Rishav, et al. [10]	Ensemble	96.96%	96.92%	96.91%
Ali, Md Shahin. et al [11]	Ensemble	98.06%	98.77%	98.97%
Pacal, Ishak, et al. [12]	Max-voting (Ensemble)	92.95%	93.89%	93.30%
Ilyas et al. [13]	Ensemble	93.99%	97.00%	96.00%
Present Model	ROS+RF ROS+DT **ROS+HPs-T+SVM**	98.54% 98.91% **99.64%**	97.10% 97.81% **98.26%**	98.53% 98.89% **99.63%**

References

[1] Bedell, Sarah L., et al. "Cervical cancer screening: past, present, and future." Sexual medicine reviews 8.1 (2020): 28-37.

[2] Nithya, B., and V. Ilango. "Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction." SN Applied Sciences 1 (2019): 1-16.

[3] World Health Organization (WHO). Cervical cancer. https://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/. Accessed March 27, 2024.

[4] Nagelhout, Gera, et al. "Is smoking an independent risk factor for developing cervical intra-epithelial neoplasia and cervical cancer? A systematic review and meta-analysis." Expert review of anticancer therapy 21.7 (2021): 781-794.

[5] Sung, Hyuna, et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries." CA: a cancer journal for clinicians 71.3 (2021): 209-249.

[6] Kohli, Pahulpreet Singh, and Shriya Arora. "Application of machine learning in disease prediction." 2018 4th International conference on computing communication and automation (ICCCA). IEEE, 2018.

[7] Cervical Cancer Dataset. https://www.kaggle.com/datasets/ranzeet013/cervical-cancer-dataset. Accessed March 24, 2024.

[8] Ratul, Ishrak Jahan, et al. "Early risk prediction of cervical cancer: A machine learning approach." 2022 19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). IEEE, 2022.

[9] Bhavani, C. H., and A. Govardhan. "Cervical cancer prediction using stacked ensemble algorithm with SMOTE and RFERF." Materials Today: Proceedings 80 (2023): 3451-3457.

[10] Pramanik, Rishav, et al. "A fuzzy distance-based ensemble of deep models for cervical cancer detection." Computer Methods and Programs in Biomedicine 219 (2022): 106776.

[11] Ali, Md Shahin, et al. "An ensemble classification approach for cervical cancer prediction using behavioral risk factors." Healthcare Analytics (2024): 100324.

[12] Pacal, Ishak, and Serhat Kılıcarslan. "Deep learning-based approaches for robust classification of cervical cancer." Neural Computing and Applications 35.25 (2023): 18813-18828.

[13] Ilyas, Qazi Mudassar, and Muneer Ahmad. "An enhanced ensemble diagnosis of cervical cancer: a pursuit of machine intelligence towards sustainable health." IEEE Access 9 (2021): 12374-12388.

[14] Peng, Jiaxu, Jungpil Hahn, and Ke-Wei Huang. "Handling missing values in information systems research: A review of methods and assumptions." Information Systems Research 34.1 (2023): 5-26.

[15] Ramaraju, H. E., Y. C. Nagaveni, and A. A. Khazi. "Use of Schiller's test versus Pap smear to increase detection rate of cervical dysplasias." International Journal of Reproduction, Contraception, Obstetrics and Gynecology 5.5 (2016): 1446-1451.

[16] Sinka, Katy. "The global burden of sexually transmitted infections." Clinics in Dermatology 42.2 (2024): 110-118.

[17] Malevolti, Maria Chiara, et al. "Dose-risk relationships between cigarette smoking and cervical cancer: a systematic review and meta-analysis." European Journal of Cancer Prevention 32.2 (2023): 171-183.

[18] Anastasiou, Elle, et al. "The relationship between hormonal contraception and cervical dysplasia/cancer controlling for human papillomavirus infection: A systematic review." Contraception 107 (2022): 1-9.

[19] Damayanti, Siti, Uki Retno Budihastuti, and Bhisma Murti. "Meta-Analysis: Effects of Hormonal Contraceptive Use and History of Sexually Transmitted Disease on the Risk of Cervical Cancer." Journal of Maternal and Child Health 8.6 (2023): 711-722.

[20] Barroeta, Julieta E. "The Future Role of Cytology in Cervical Cancer Screening in the Era of HPV Vaccination." Acta Cytologica 67.2 (2023): 111-118.

[21] Minalt, Nicole, et al. "Association of Intrauterine Device Use and Endometrial, Cervical, and Ovarian Cancer: an Expert Review." American Journal of Obstetrics and Gynecology (2023).

[22] Hayaty, Mardhiya, Siti Muthmainah, and Syed Muhammad Ghufran. "Random and synthetic over-sampling approach to resolve data imbalance in classification." International Journal of Artificial Intelligence Research 4.2 (2020): 86-94.

[23] Abro, Abdul Ahad, et al. "Machine learning classifiers: a brief primer." University of Sindh Journal of Information and Communication Technology 5.2 (2021): 63-68.

[24] Sun, Jiancheng, et al. "Analysis of the distance between two classes for tuning SVM hyperparameters." IEEE transactions on neural networks 21.2 (2010): 305-318.

[25] Priyanka, and Dharmender Kumar. "Decision tree classifier: a detailed survey." International Journal of Information and Decision Sciences 12.3 (2020): 246-269.

[26] Genuer, Robin, et al. Random forests. Springer International Publishing, 2020.

[27] Pisner, Derek A., and David M. Schnyer. "Support vector machine." Machine learning. Academic Press, 2020. 101-121.

[28] Jupyter notebook, https://jupyter.org/, [Last accessed: 16.04.24].

[29] Amin, Fahmy, and M. Mahmoud. "Confusion matrix in binary classification problems: a step-by-step tutorial." Journal of Engineering Research 6.5 (2022): 0-0.

[30] Schwenke, Carsten, and A. G. Schering. "True positives, true negatives, false positives, false negatives." Wiley StatsRef: Statistics Reference Online (2014).

[31] Powers, D. M. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness& Correlation. Journal of Machine Learning Technologies, 2(1), 37-63. doi: 10.5121/jmlr.2011.2103

[32] Polo, Tatiana Cristina Figueira, and Hélio Amante Miot. "Use of ROC curves in clinical and experimental studies." Jornal vascular brasileiro 19 (2020): e20200186.

[33] Turner, J. Rick. "Area under the curve (AUC)." Encyclopedia of Behavioral Medicine (2020): 146-146.

[34] Weerts, Hilde JP, Andreas C. Mueller, and Joaquin Vanschoren. "Importance of tuning hyperparameters of machine learning algorithms." arXiv preprint arXiv:2007.07588 (2020).

[35] Verma, Ankit, et al. "Web application implementation with machine learning." 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM). IEEE, 2021.