International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Evaluation of Machine Learning Algorithms and Feature Selection Methods for Cervical Cancer Risk Prediction using Clinical Features

Sokaina El Khamlıchı[1], Ikram Ben Abdel Ouahab[2], Mohammed Bouhorma[3], Fatiha Elouaaı[4], Abdelfettah Sedquı[5], Amal Maurady*[6]

Submitted: 10/09/2022 **Accepted:** 20/12/2022

Abstract: Cervical cancer is one of the most frequent gynecological cancers worldwide. It is associated to several risk factors like sexually transmitted diseases, human papillomavirus and smoking. The early diagnosis of this disease is crucial to lower fatality rates. Furthermore, its early prediction can support clinicians and patients to have an effective treatment. This study intends to compare machine learning classifiers to determine the best model to predict cervical cancer and identify its most significant risk factors. This work compares five machine learning algorithms: K-Nearest Neighbor, Gaussian Naïve Bayes, Logistic Regression, Random Forest and Decision Tree (DT). Afterwards, the study continues to enhance the outcome of DT algorithm through balancing the data with Synthetic Minority Oversampling Technique (SMOTE), selecting the most important features with Recursive Feature Elimination (RFE) and tuning hyperparameters with Grid Search technique. Overall, the combination of Decision Tree classification technique with SMOTE and tuning hyperparameters with Grid Search method presents the most performing model.

Keywords: *Cervical Cancer, Decision Tree, Feature Selection, RFE, SMOTE, Supervised Machine Learning.*

1. Introduction

Cervical cancer was the fourth most common cancer in terms of incidence and mortality in2018, with 570,000 new diagnoses and 311,000 deaths around the world [1], [2]. It is the third most prevalent cancer among women all over the world, with 85% of cases occurring in low- and middle-income countries. It represents almost 12% of all female-associated cancers. It is the second most frequent cancer among Moroccan women, with an age-standardized prevalence rate of 17.2 per 100,000 ladies and an age-standardized death rate of 12.6 in 2018 [3].

This cancer, originating from the cervix, is generated from the mutations of genes controlling the cell's growth and division functions. It is capable to expand from the cervix to different organs of the body. In the initial stages, no symptoms can be noticed. It can only be detected early by regular check-ups. In the late phase, signs like pelvic pain and vaginal bleeding become visible. Moreover, cervical cancer can spill over other parts of the body such as lungs and abdomen. Additionally, the advanced stage of this disease reveals some symptoms such as leg pain, back pain, bone fractures, tiredness and weight loss. This illness can be detected to some degree using diffusion-weighted imaging (DWI) and magnetic resonance imaging (MRI) [4], [5]. On the other hand, in developing countries, people have a poor understanding of regular screening importance. Moreover, this malady has become a significant reason of death in low-income countries due to a shortage of physician expertise and restricted medical equipment [6].

Human papillomavirus (HPV) is the main cause of cervical cancer [7]. Besides, there are numerous other causes of this ailment including cigarette smoking, the use of contraceptives, several pregnancies, and a number of other causes. For example, if an HPV-positive patient smokes, the risk of cervical cancer will rise by two to three times [8]. In the meantime, it is discovered that women who use contraceptives have a three-fold higher incidence of this illness than women who do not. Additionally, if contraceptives are utilized for more than ten years, the occurrence will increase to four times. When it comes to multiple pregnancies, female HPV-positive patients without pregnancies have a lower risk of cervical cancer than those having multiple full-term pregnancies [9].

[1] *Laboratory of Innovative Technologies, National School of Applied Sciences of Tangier, Abdelmalek Essaâdi University, Tangier, Morocco,ORCID ID: 0000-0001-8848-9661*
[2] *Computer science, systems and telecommunication laboratory (LIST), Faculty of sciences and techniques, University Abdelmalek Essaadi, Tangier, Morocco, ORCID ID: 0000-0003-0955-6382*
[3] *Computer science, systems and telecommunication laboratory (LIST), Faculty of sciences and techniques, University Abdelmalek Essaadi, Tangier, Morocco, ORCID ID: 0000-0002-5687-5231*
[4] *Computer science, systems and telecommunication laboratory (LIST), Faculty of sciences and techniques, University Abdelmalek Essaadi, Tangier, Morocco, ORCID ID: 0000-0002-7139-5682*
[5] *Laboratory of Innovative Technologies, National School of Applied Sciences of Tangier, Abdelmalek Essaâdi University, Tangier, Morocco, ORCID ID:0000-0002-7446-0400*
[6] *Laboratory of Innovative Technologies, National School of Applied Sciences of Tangier, Abdelmalek Essaâdi University, Tangier, Morocco Faculty of sciences and techniques, Abdelmalek Essaâdi University, Tangier, Morocco*
ORCID ID:0000-0001-9298-717X
** Corresponding Author Email: amal.maurady.ma@gmail.com*

Currently, applications of Artificial Intelligence (AI) have become very widespread and crucial in medical domain. Machine learning algorithms have demonstrated their prominence in various medical data. For instance, fundus images were used to detect diabetic retinopathy [10] and to screen glaucoma [11]. Region-Based Convolutional Neural Network was used to detect skin cancer [12]. AI presents also an immense promise in the field of mental healthcare [13]. Furthermore, deep learning algorithms were used to analyze electronic health records [14] and much more challenging tasks such as classifying the structure of trabecular bone in osteoporotic individuals [15] as well as medial image segmentation for brain and spine [16].

The aim of our study is to predict cervical cancer risk using an efficient classifier based on clinical features. For this reason, we started with a comparison between five machine learning algorithms: K-Nearest Neighbor, Logistic Regression, Random Forest, GNB and Decision Tree. Since Decision Tree algorithm yielded the best result in this comparison, the study went on with Decision Tree through balancing the data, tuning hyperparameters and applying Recursive Feature Elimination (RFE) in order to select the most relevant features. We obtained encouraging results.

The rest of this paper is organized as follows: section 2 presents related works of cervical cancer classification. The methodology of our study is explained in section 3. Then, section 4 focuses on the proposed solution. The results and discussion are illustrated in section 5. Furthermore, Section 6 deals with analysis and comparison. Finally, section 7 covers the conclusion.

2. Related works

Recently, number of researchers has studied data on cervical cancer. [17] provided the publicly available dataset used in the present paper. Transfer learning techniques were used to predict the patient's risk through transmitting information between linear classifiers on analogous tasks. Hence, the main goal of the research was illustrating the effect of transform learning techniques on improving accuracy.

In another research, [18]analyzed the same data using Decision Tree classifier to perform cost-sensitive classification. Besides, [6] applied three SVM-based methods where the SVM approach yielded the best results comparing to the SVM-PCA and the SVM-RFE approaches. Moreover, [19] studied the above-mentioned dataset using sampling methods to balance the data and feature selection to enhance the accuracy of the model. The findings reveal that age, number of pregnancies, first sexual intercourse, hormonal contraceptives, STDs:genital herpes and smokes are the principal predictive attributes. More attention was paid to feature selection methods. [20]analyzed the same dataset using Random Forest classifier with two feature selection methods Recursive Feature Elimination and Principal Component Analysis. Synthetic Minority Oversampling Technique was used to balance the data. Hence, SMOTE-RF model outperforms SMOTE-RF-RFE and SMOTE-RF-PCA models.

On the other hand, there are other studies dealing with cervical cancer classification in other datasets. These datasets have different features. For instance, [21] analyzed a dataset containing 145 patients from Iran and 23 attributes. The study predicted cervical cancer using QUEST tree, C&R tree, RBF-ANNs, SVM and MLP-ANNs algorithms. Decision Trees performed better than the other algorithms. Therefore, the main features predicting this ailment comprises marital status, individual wellbeing level, social status, schooling level, contraceptive dosage taken and the number of caesarean deliveries[21].

Moreover, [22] examined a dataset of 799 cells taken from patients in Malaysia's Hospital UniversitiSains. The inputs are cytoplasm size, nucleus size, nucleus grey level as well as cytoplasm grey level, and the target variable has three outputs including High grade Squamous Intraepithelial Lesion, Low-grade Squamous Intraepithelial Lesion as well as normal cases. This research used a hybrid Multi-layered Perceptron trained with Genetic Algorithm. This hybrid algorithm outpaced the Hybrid Radial Basis Function trained with Adaptive Fuzzy K-means and Moving K-means Clustering algorithm.

Ref	***Title***	***Database***	***Method***	***Result***
[18]	"Enhanced Classification Model for Cervical Cancer Dataset based on Cost Sensitive Classifier"	A dataset collected from patients at Hospital Universitario de Caracas in Caracas [23]	-Decision Tree -Cost sensitive Decision Tree	Cost sensitive Decision Tree: • True Positive rate=0.429
[6]	"Data-Driven Diagnosis of Cervical Cancer With Support Vector Machine-Based Approaches"		-SVM -SVM-PCA -SVM-RFE	SVM: • Accuracy= 94.13% • Sensitivity = 100% • Specificity = 90.21% • PPA(Precision) = 86.07% • NPA = 100%
[19]	"Classification of Cervical Cancer Dataset"		-Gaussian Naive Bayes (GNB) -KNN -DT -LR -SVM	Decision Tree using the over sampling method and picking out selective attributes : • Accuracy = 97.5%
[20]	"Cervical Cancer Diagnosis		-SMOTE-RF	SMOTE-RF:

	Using Random Forest Classifier With SMOTE and Feature Reduction Techniques"		-SMOTE-RF-RFE -SMOTE-RF-PCA	• Accuracy= 96.06% • Sensitivity= 94.55% • Specificity = 97.51% • PPA(Precision) = 97.33% • NPA = 94.91
[21]	"Supervised Algorithms of Machine Learning for the Prediction of Cervical Cancer"	Dataset referred to Shohada Hospital of Tehran, Iran. In 2017–2018.	-SVM, -QUEST tree, -C&R tree, -MLP-ANNs -RBF-ANNs	Decision Trees : • Accuracy = 95.55% • Sensitivity = 90.48% • Specificity = 100% • AUC = 95.20%
[22]	"Classification of Cervical Cancer Using Hybrid Multi-layered Perceptron Network Trained by Genetic Algorithm"	The data was obtained from Hospital UniversitiSains in Malaysia (HUSM).	- Hybrid Multi-layered Perceptron (HMLP) trained with Genetic Algorithm - Hybrid Radial Basis Function (HRBF) trained with Adaptive Fuzzy K-means and Moving K-means Clustering (AFKM)	HMLP trained with Genetic Algorithm: • Accuracy = 74.82% • Sensitivity = 72.50% • Specificity = 86.76%

A summary of all related works is presented in (Table 1), highlighting the used databases, methods as well as results

3. Methodology

3.1. Dataset

The dataset used in the present study is taken from UCI Repository [23]. It is collected from patients at Hospital Universitario de Caracas in Caracas, Venezuela. It encompasses 858 patients, 35 features and the result of a Biopsy examination as a target variable. The features include demographic details, behaviors and previous medical reports as presented in (Table 2).

Table 2. Features from the database

1 Age	10 IUD : Intrauterine device	19 STDs: pelvic inflammatory disease	28 STDs: Time since last diagnosis
2 Number of sexual partners	11 IUD (years)	20 STDs: genital herpes	29 Dx: Cancer
3 First sexual intercourse (age)	12 STDs : Sexually Transmitted Diseases	21 STDs: molluscum contagiosum	30 Dx: CIN
4 Number of pregnancies	13 STDs (number)	22 STDs: AIDS	31 Dx: HPV
5 Smokes	14 STDs: condylomatosis	23 STDs: HIV	32 Dx
6 Smokes (years)	15 STDs: cervical condylomatosis	24 STDs: Hepatitis B	33 Hinselmann
7 Smokes (packs/year)	16 STDs: vaginal condylomatosis	25 STDs: HPV	34 Schiller
8 Hormonal Contraceptives	17 STDs: vulvo-perinealcondylomatosis	26 STDs: Number of diagnosis	35 Cytology
9 Hormonal Contraceptives (years)	18 STDs: syphilis	27 STDs: Time since first diagnosis	36 Biopsy: target variable

3.2. Preprocessing data

In the original database, two features comprising more than 80% of missing values, STDs: Time since first diagnosis and STDs: Time since last diagnosis, were dropped off. On the other hand, the lines that have the attributes Smokes and First sexual intercourse missing were omitted, since these columns contain the fewest missing records. Therefore, the dataset is made of 838 and 34 columns.

After that, the database still contains missing observations. So, machine learning models have been used to fill the missing values. The following 7 steps have been sued (Fig. 1):

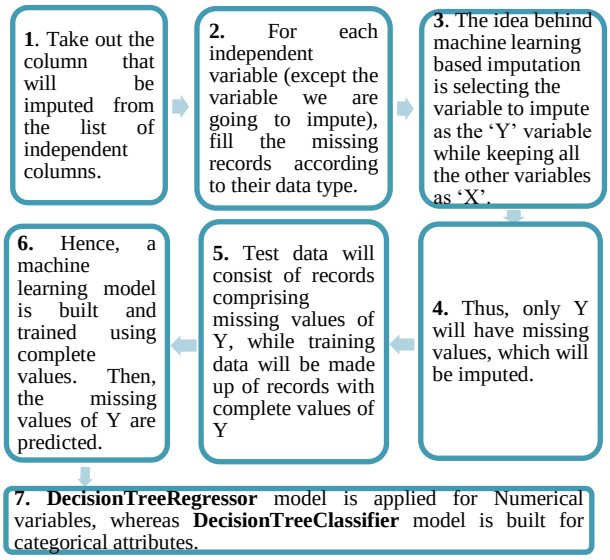

Fig. 1. Cleaning data and Filling missing values using Machine learning

3.3. Machine learning algorithms

In order to predict cervical cancer risk based on medical features, we compare 5 machine learning algorithms. Our goal is to find the most appropriate model in our case. These algorithms are very popular and the most used for prediction and classification problematics.

Logistic Regression is a statistical algorithm that belongs to the Generalized Linear Models. In spite of its name, it is applied for classification not for regression. In the literature, the log-linear classifier, logit regression and maximum-entropy classification (MaxEnt) are all terms used to describe logistic regression. In this algorithm, a logistic function is used to model a dichotomous variable dependent on any type of explanatory variables.

Logistic regression uses an equation similar to linear regression, below is an example of it:

$$y = \frac{e^{(b_0+b_1 \times x)}}{1+e^{(b_0+b_1 \times x)}} \quad (1)$$

Where,

- *b_0 is the bias*
- *b_1 is the coefficient for the single input value (x)*
- *y is output value*
- *x is input value*

Decision Tree is a simple and performant non-parametric technique of data analysis, which is used in supervised learning process. It solves classification and regression problems. Its aim is to build a model predicting the target variable's value through learning basic decision instructions deduced from the data attributes.

Random Forest is a supervised machine learning algorithm used for classification and regression tasks. It is based on ensemble learning method, which consists of combining various kinds of algorithms or same algorithm several times, in order to build a more efficient prediction model. Random Forest generates Decision Trees from randomly chosen data samples. After taking the prediction obtained by each tree, it determines the best result through voting.

Naive Bayes is a probabilistic statistical classification method that applies Bayes theorem. It is one of the most basic supervised learning algorithms. It assumes that every pair of features are independent between each other. This assumption makes the calculations simpler. Therefore, it is called naïve. It is also known as class conditional independence. When dealing with continuous data and Gaussian normal distribution, the Gaussian Naïve Bayes algorithm is used.

K-Nearest Neighbor or KNN is a non-parametric algorithm used in a supervised learning for classification problems. It may also be used for regression tasks. The theory behind this algorithm is very simple. It basically computes the distance between a new data point and every data point in the training set. Any kind of distance may be used, such as Euclidean, Hamming or Manhattan distances. The K-nearest data points are then chosen, where K is an integer. Finally, the data point is assigned to the class containing the majority of the K data points. In our implementation, we use Euclideandistance defined in the formula below, where p and q are two points.

$$d(p,q) = d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (2)$$

3.4. Balancing data and dimensionality reduction methods

To deal with imbalanced data issue, SMOTE technique was used. Then, RFE and PCA were used to select relevant features.

Synthetic Minority Oversampling Technique (SMOTE) is an oversampling approach that was introduced to enhance random oversampling. In this method, synthetic samples for the minority class are created. It addresses the overfitting problem caused by random oversampling. SMOTE concentrates on identifying instances in the feature space that are close together and create new ones by interpolating between the instances (Fig. 2)

Algorithm *SMOTE*(T, N, k)
Input: Number of minority class samples T; Amount of SMOTE N%; Number of nearest neighbors k
Output: $(N/100) * T$ synthetic minority class samples

1. (∗ *If N is less than 100%, randomize the minority class samples as only a random percent of them will be SMOTEd.* ∗)
2. **if** $N < 100$
3. **then** Randomize the T minority class samples
4. $T = (N/100) * T$
5. $N = 100$
6. **endif**
7. $N = (int)(N/100)$ (∗ *The amount of SMOTE is assumed to be in integral multiples of 100.* ∗)
8. k = Number of nearest neighbors
9. *numattrs* = Number of attributes
10. *Sample*[ ][ ]: array for original minority class samples
11. *newindex*: keeps a count of number of synthetic samples generated, initialized to 0
12. *Synthetic*[ ][ ]: array for synthetic samples
(∗ *Compute k nearest neighbors for each minority class sample only.* ∗)
13. **for** $i \leftarrow 1$ **to** T
14. Compute k nearest neighbors for i, and save the indices in the *nnarray*
15. Populate(N, i, *nnarray*)
16. **endfor**

Populate(N, i, *nnarray*) (∗ *Function to generate the synthetic samples.* ∗)
17. **while** $N \neq 0$
18. Choose a random number between 1 and k, call it nn. This step chooses one of the k nearest neighbors of i.
19. **for** $attr \leftarrow 1$ **to** *numattrs*
20. Compute: $dif = Sample[nnarray[nn]][attr] - Sample[i][$
21. Compute: gap = random number between 0 and 1
22. $Synthetic[newindex][attr] = Sample[i][attr] + gap * dif$
23. **endfor**
24. $newindex{+}{+}$
25. $N = N - 1$
26. **endwhile**
27. **return** (∗ *End of Populate.* ∗)
End of Pseudo-Code.

Fig. 2. SMOTE algorithm pseudo code proposed in [24]

Principal Component Analysis (PCA) is a statistical method used for dimensionality reduction, rising interpretability while keeping as much information as possible. It accomplishes this by generating new uncorrelated variables, which are linear functions

of the original variables in the dataset and which consecutively maximize variance. Consequently, the original features are transformed into a reduced number of principal components comprising the majority of variance in the original features. PCA steps are as follow:

1) Standardization: Compute the mean of each dataset's dimension, excluding the labels. Scale the data with the aim that every variable has an equal impact on the analysis. z, x, μ and σ are the scaled value, the initial value, mean and standard deviation, respectively, in the equation below:

$$z = \frac{x-\mu}{\sigma} \quad (3)$$

2) Covariance Matrix Calculation: The covariance of two variables X and Y can be calculated using the formula:

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y}) \quad (4)$$

The covariance matrix of A can be computed using the formula above. Additionally, the outcome would be a square matrix with dimensions of x*x.

3) *Calculate Eigenvectors and corresponding Eigenvalues:* In linear algebra, an eigenvector, characteristic vector of a linear transformation, is defined as a nonzero vector, which mostly modifies by a scalar factor while applying the lineartransformation to it. The factor with which the eigenvector is scaled is the corresponding eigenvalue.

Generally, the vector verifying the condition below is the eigenvector of a matrix A:

$$A\vec{v} = \lambda\vec{v} \quad (5)$$

Knowing that the scalar, λ, is the eigenvalue. This indicates that λ defines the linear transformation, and the equation may be expressed as: $\vec{v}(A - \lambda I) = 0$(6)

Where the identity matrix is represented by I

It is noteworthy that these both eigenvectors are unit eigenvectors, which means that their length is the same and equals 1. These eigenvectors allow us to extract the most useful patterns from the data.

4) Select the k eigenvectors with the highest eigenvalues as follows: Sort the eigenvectors in decreasing order of eigenvalues, then pick out of them k eigenvectors, knowing that k is the number of dimensions obtained in the new dataset.

Recursive Feature Elimination is a backward feature selection technique. The purpose of this technique is to pick attributes through investigating recursively fewer sets of attributes. Initially, the model is built on the whole set of features and the importance score is assigned to each one. Then, the least important attributes are pulled out, the model is built again and the importance scores are recalculated. This process is recursively applied until the specified number of attributes to select is attained.

3.5. Evaluation metrics

In statistics, a confusion matrix presents the summary of the prediction results on a classification. In this context, we use these terms: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). From confusion matrix, we get the model's accuracy, recall, precision, and f1-score. These measures are evaluation metrics and help to compare various trained models.

- Accuracy is the percentage of correct predictions classified by the model. It is calculated as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

- Recall is the classifier's capacity to identify all the positive observations. It is computed like so:

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

- Precision is the classifier's capacity not to identify a negative sample as positive. The following equation gives the formula defining precision:

$$Precision = \frac{TP}{TP+FP} \quad (9)$$

- F1-score is the weighted average of the model's precision and recall. It provides a better assessment of misclassified samples than the accuracy metric. It is defined as follows:

$$f1_{score} = 2 * \frac{precision*recall}{precision+recall} \quad (10)$$

- ROC_AUC score calculates the Area Under the Receiver Operating Characteristic Curve (ROC curve). It calculates the capability of a classifier to separate between classes. The ROC curve represents the True Positive Rate (TPR) against the False Positive Rate (FPR).

3.6. Tuning hyperparameters with Grid Search

Grid search is a method for tuning hyperparameters of a machine learning model. Contrary to parameters, discovering hyperparameters in training data is unachievable. Thus, to identify the optimal hyperparameters, a model for every combination of hyperparameters has been created. Since all feasible combinations are "brute-forced", Grid search is consequently characterized as a very classic hyperparameter optimization technique. Cross validation is then used to assess the models. Obviously, the model with the highest performance is maintained as the best.

4. Proposed solution

The purpose of this study is to determine an effective classifier using a small number of features. Thereby, clinicians could predict cervical cancer thoroughly and effectively at an early stage and obtain a preliminary diagnosis of cancer.

The present study focuses on comparing five machine learning algorithms: K-Nearest Neighbor, Logistic Regression, Random Forest, GNB and Decision Tree. Decision Tree classifier presented the best results in this comparison. Consequently, the analysis went on improving the results of this algorithm using feature selection techniques such as RFE and PCA to select the

most important features in the dataset, as well as balancing the data using SMOTE method.

The current work has started with data preprocessing which aims to clean the dataset and fill the missing values using machine learning algorithms. Then, a comparison of five machine learning models has been done with default parameters. After that, these classifiers have been improved using Grid Search. Consequently, the Decision Tree classifier has outperformed all the other algorithms and the study has gone on improving Decision Tree results. Therefore, the data have been balanced using SMOTE. Whereas, PCA and RFE methods have been used to select the most relevant features for the study (Fig. 3).

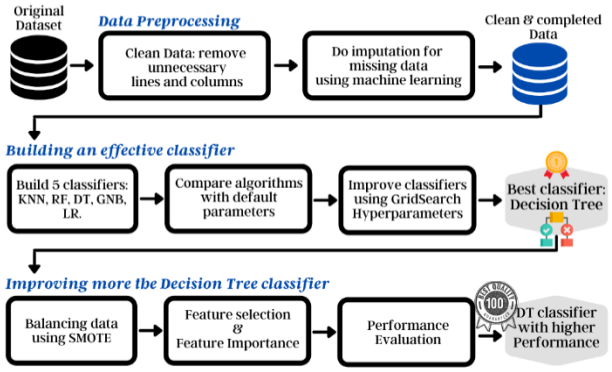

Fig. 3. Flowchart of the proposed solution

5. Results and Discussion

In this research, five machine learning algorithms have been investigated including Logistic Regression, Decision Tree, Random Forest, Gaussian Naïve Bayes and K-Nearest Neighbor. The dataset has been divided into 70% for training and 30% for testing. Random Forest has performed better than the other algorithms while using default parameters, followed byDecision Tree, then Logistic Regression, KNN and finally GNB, which has performed worse in this comparison (Table3). However, when tuning the hyper parameters, Decision Tree algorithm outperforms all the other methods with accuracy equals 96.19%, f1score equals 73.33%, recall equals 91.67%, precision equals 61.11% and ROC_AUC equals 94.07%, followed by Random Forest. Then, Logistic Regression, KNN and finally GNB with the lowest performance (Table 5). Therefore, Decision Tree will be maintained for further studies.

Table 3. Algorithms comparison with default parameters

Model	*Train_ Score (%)*	*Test_accuracy (%)*	*F1-score (%)*	*Recall (%)*	*Precision (%)*	*ROC_ AUC (%)*
GNB	14.81	9.52	11.22	100	5.94	52.02
DT	100	94.76	64.52	83.33	52.63	89.39
RF	100	*95.24*	*66.67*	*83.33*	*55.56*	*89.65*
LR	96.82	94.29	45.45	41.67	50	69.57
KNN	95.54	93.81	38.10	33.33	44.44	65.40

We have performed grid search on the previous models in order to find out the best hyperparameters and improve metrics. The best hyperparameters we found so far are presented in (Table 4). After using Grid Search, the results revealed that Decision Tree performs better than the other models. Indeed, all metrics, including accuracy, f1score, recall, precision, ROC_AUC, have increased and the overfitting problem has been eliminated (Table 5).

Table 4. Grid search results

Algorithm	*Hyper parameter*	*Value*
DT	criterion	entropy
	max_depth	3
	max_features	auto
	random_state	123
	splitter	best
LR	C	1.0
	penalty	12
GNB	var_smoothing	0.01
RF	n_estimators	90
	max_features	21
	criterion	gini
KNN	n_neighbors	27

Table 5. Algorithms comparison after grid search

Model	*Train_ Score (%)*	*Test_accuracy (%)*	*F1-score (%)*	*Recall (%)*	*Precision (%)*	*ROC_ AUC (%)*
GNB	91.56	88.57	14.29	16.67	12.50	54.80
DT	*96.82*	*96.19*	*73.33*	*91.67*	*61.11*	*94.07*
RF	100	96	70.97	91.67	57.89	93.81
LR	96.82	94.29	45.45	41.67	50	69.57
KNN	93.95	94.29	25	16.67	50	57.83

5.1. Balancing data

In order to balance the data, SMOTE technique was used. Hence, 586 samples were obtained per both classes. First, Decision Tree classifier with balanced data using default parameters has been applied. Then, tuning hyper parameters has been implemented using Grid search in order to improve the algorithm's performance. As a result, the algorithm's performance has slightly increased (Table 7).

5.2. Feature importance

Using Decision Tree algorithm, the most important features are Schiller, Number of pregnancies, Age, Smokes (packs/year), First sexual intercourse, Hormonal_Contraceptives_years, STDs number, STDs_genital_herpes, Number of sexual partners, IUD years, STDs number of diagnosis, Dx_CIN, Hinselmann, STD_HIV, Smokes_ years (Fig. 4)

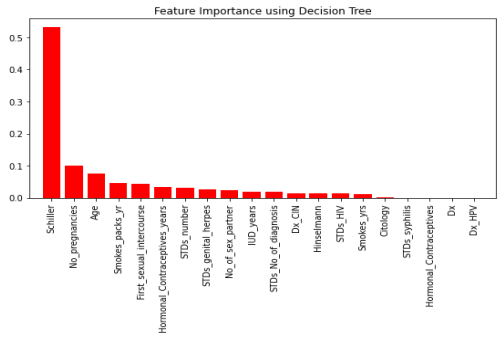

Fig. 4. Feature importance using DT

5.3. PCA before and after balancing the data

Using Principal Component Analysis, the explained variance value for first 2 components is 77.5% before balancing the data (Fig. 5). While it reached 78.5% after balancing the data (Fig. 6).

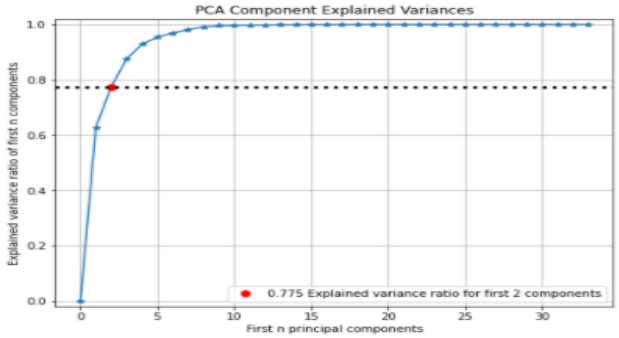

Fig. 5. PCA before balancing the data with SMOTE

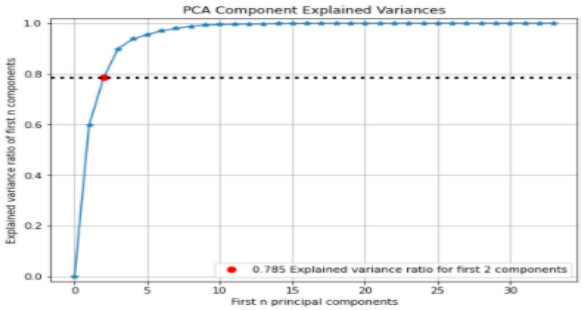

Fig. 6. PCA after balancing the data with SMOTE

5.4. Recursive Feature Elimination using balanced data (RFE+SMOTE)

Figure 7 shows that ROC_AUC is high when the number of features is equal to one feature, three features and 18 features (Fig.7). The ROC_AUC is maximal (98.23%) when the number of features is 18 and it is equal to 97.98%in case of 1 feature and 3 features (Table 6).

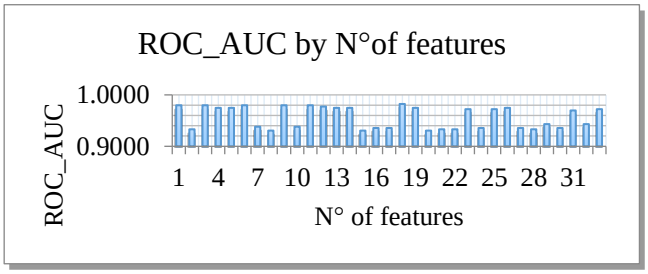

Fig.7. ROC_AUC according to the number of features for balanced data

Table 6. TOP 3 best cases when the ROC_AUC is maximal

N° of features	***ROC_AUC (%)***	***Features***
18	98.23	1. Age 2. Number of sexual partners 3. First sexual intercourse 4. Num of pregnancies 5. Smokes (packs/year) 6. Hormonal Contraceptives 7. Hormonal Contraceptives (years) 8. IUD (years) 9. STDs (number) 10. STDs:syphilis 11. STDs:pelvic inflammatory disease 12. STDs:genital herpes 13. STDs:molluscumcontagiosum 14. STDs:AIDS 15. Dx:CIN 16. Hinselmann 17. Schiller 18. Citology
1	97.98	1. Schiller
3	97.98	1. Age 2. Dx:CIN 3. Schiller

5.5. Recursive Feature Elimination using imbalanced data

In order to check whether the ROC_AUC is higher for imbalanced data than in case of balanced data, Fig.7 illustrates the ROC_AUC by the Number of features for imbalanced data. Accordingly, the maximal value of ROC_AUC is 93.81% which is very small than the ROC_AUC measured in case of using balanced data (Fig.8). Consequently, imbalanced data will not be maintained for further experiments.

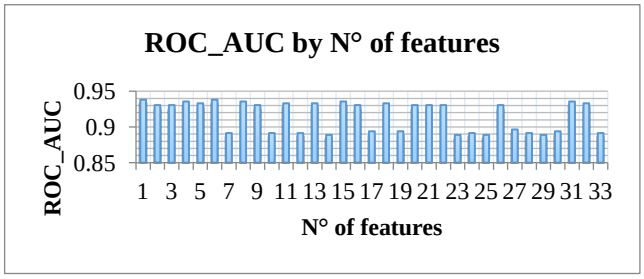

Fig. 8. ROC_AUC according to the number of features for imbalanced data

Accordingly, the best performance is given while using Recursive feature Elimination with balanced data. The optimal number of features is 18 features, namely, age, Number of sexual partners, First sexual intercourse, Number of pregnancies, Smokes, Smokes (years), Smokes (packs/year), Hormonal Contraceptives, Hormonal Contraceptives (years), IUD (years), STDs:genital herpes, STDs:molluscumcontagiosum, STDs:AIDS, STDs:HIV, STDs: Number of diagnosis, Dx:CIN, Schiller, Citology.

The comparison of Decision Tree illustrates that Decision Tree model using hyper parameters and tuned hyper parameters is the best classifier compared to Bagged Decision Tree with Hyper parameters, Decision Tree ADA Boost with Hyper parameters, and Gradient Boost (Table 7).

DT with default parameters achieved 93.81% accuracy, 60.61% f1score, 83.33% recall, 47.62% precision and 88.89% ROC_AUC. Using Grid Search, the performance increased with 96.19% accuracy, 73.33% f1score, 91.67% recall, 61.11% precision and 94.07% ROC_AUC. After balancing the data with SMOTE and tuning hyper parameters with Grid search, the performance achieved its maximum 96.19% accuracy, 75% f1score, 100% recall, 60% precision and 97.98% ROC_AUC.

On the other hand, DT-RFE and SMOTE-DT-RFE-Grid Search models reached the same performance, which is lower than the performance achieved by SMOTE-DT-Grid Search model. Moreover, Bagged Decision Tree with Hyper parameters, Decision Tree ADA Boost with Hyper parameters, and Gradient Boost has lower performance compared to SMOTE-DT- Grid Search model. Therefore, SMOTE-DT- Grid Search model with all features achieved the best accuracy, f1score, recall, precision and ROC_AUC.

Table 7. Summary of all results

Model	*Train _Score (%)*	*Test_ accuracy (%)*	*F1-score (%)*	*Recall (%)*	*Precision (%)*	*ROC_ AUC (%)*
DT (default parameters)	100	93.81	60.61	83.33	47.62	88.89
DT-Grid Search	96.82	96.19	73.33	91.67	61.11	94.07
SMOTE-DT-Default	100	94.76	66.67	91.67	52.38	93.31
SMOTE-DT-Grid Search	98.38	96.19	75	100	60	97.98
SMOTE-DT-RFE	100	94.76	66.67	91.67	52.38	93.31
DT-RFE	100	95.24	68.75	91.67	55	93.56
SMOTE-DT-RFE-Grid Search	97.93	95.24	68.75	91.67	55	93.56
Bagged Decision Tree with Hyperparameters	97.93	95.24	61.54	66.67	57.14	81.82
Decision Tree ADA Boost with Hyperparameter	100	95.24	66.67	83.33	55.56	89.65
Gradient Boost	97.93	96.19	71.43	83.33	62.50	90.15

6. Analysis and Comparison

The outcome of this study demonstrates that Decision Tree classifier using SMOTE and Grid Search (SMOTE-DT-GridSearch) model obtained the following performance: 96.19% accuracy, 75% f1score, 100% recall, 60% precision and 97.98% ROC_AUC.

After comparing the results of our study with those of [20] and [25], we discovered that SMOTE-DT-GridSearch obtained better results in terms of accuracy and recall compared with SMOTE-RF which is the best model found in [20]. However, DT-RFE-SMOTETomek model in [25] performs better than SMOTE-DT-GridSearch in our study. Moreover, the outcome of our study illustrates that RFE for feature selection doesn't improve the performance of Decision Tree model. This result is the same as in [20] where RFE didn't enhance the performance of Random Forest model. However, balancing data with SMOTE gives better results in our study as well as in [20] (Table 8).

It is noteworthy to mention that this study has some limitations. In fact, the lack of cervical cancer datasets containing the same attributes make the comparison of our work with other studies difficult.

Table 8. Comparison of obtained results

Ref.	*Method*	*Features (%)*	*Accuracy (%)*	*Recall/ Sensitivity (%)*	***Precision/PPA (%)***
[20]	SMOTE-RF	30	96.06	94.55	97.33
	SMOTE-RF-RFE	6	95.23	94.94	95.31
		18	95.87	94.42	97.06
	SMOTE-RF-PCA	8	95.55	93.77	97.04
		11	95.74	94.16	97.58
[25]	DT	-	96	86	-
	DT-RFE	-	98	86	-
	DT-RFE-SMOTETomek	-	98	100	-
Our study	SMOTE-DT-GridSearch	-	96.19	100	60
	DT- Grid Search	-	96.19	91.67	61.11
	SMOTE-DT-RFE-Grid Search	-	95.24	91.67	55

In addition, we recommend the use of fewer features based on the features importance study we have performed. For example, only 18 features can be used with high performances namely: age, Number of sexual partners, First sexual intercourse, Number of pregnancies, Smokes, Smokes (years), Smokes (packs/year), Hormonal Contraceptives, Hormonal Contraceptives (years), IUD (years), STDs:genital herpes, STDs:molluscumcontagiosum, STDs:AIDS, STDs:HIV, STDs: Number of diagnosis, Dx:CIN, Schiller, Cytology.

7. Conclusion and future work

Cervical cancer is the fourth most prevalent kind of cancer among women worldwide and a leading cause of death. It is helpful to prevent if it is discovered in its early stages that patients would benefit from an effective treatment. The present study compares various machine learning algorithms in order to select the best classifier, which predicts accurately cervical cancer. The findings demonstrate that Decision Tree algorithm using balanced data with smote and tuned hyper parameters is the best classifier with accuracy equals 96.19% and ROC_AUC equals 97.98%. Moreover, age, Number of sexual partners, First sexual intercourse, Number of pregnancies, Smokes, Smokes (years), Smokes (packs/year), Hormonal Contraceptives, Hormonal Contraceptives (years), IUD (years), STDs:genital herpes, STDs:molluscumcontagiosum, STDs:AIDS, STDs:HIV, STDs: Number of diagnosis, Dx:CIN, Schiller, Citology are the most predictive features. Thereby, further studies in this issue would be very beneficial. It might be useful to conduct more researches using other feature selection techniques.

As perspective, we are looking forward to use the power of Deep Neural network to predict cervical cancer risk. Also, our study will focus more on prediction duration, as the task needs to be fast. In real life, we suppose that our study could be very useful in medical environments and could save lives.

8. References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, et A. Jemal, « Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries », *CA. Cancer J.* Clin., vol. 68, no 6, p. 394-424, nov. 2018, doi: 10.3322/caac.21492.

[2] Yadav, P. ., S. . Kumar, and D. K. J. . Saini. "A Novel Method of Butterfly Optimization Algorithm for Load Balancing in Cloud Computing". International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 8, Aug. 2022, pp. 110-5, doi:10.17762/ijritcc.v10i8.5683.

[3] « Cervical cancer ». https://www.who.int/westernpacific/health-topics/cervical-cancer (consulté le 11 avril 2021).

[4] W. Messoudi et al., « Cervical cancer prevention in Morocco: a model-based cost-effectiveness analysis », J. Med. Econ., vol. 22, no 11, p. 1153-1159, nov. 2019, doi: 10.1080/13696998.2019.1624556.

[5] M. Exner et al., « Value of diffusion-weighted MRI in diagnosis of uterine cervical cancer: a prospective study evaluating the benefits of DWI compared to conventional MR sequences in a 3T environment », Acta Radiol., vol. 57, no 7, p. 869-877, juill. 2016, doi: 10.1177/0284185115602146.

[6] P. Z. McVeigh, A. M. Syed, M. Milosevic, A. Fyles, et M. A. Haider, « Diffusion-weighted MRI in cervical cancer », Eur. Radiol., vol. 18, no 5, p. 1058-1064, mai 2008, doi: 10.1007/s00330-007-0843-3.

[7] W. Wu et H. Zhou, « Data-Driven Diagnosis of Cervical Cancer With Support Vector Machine-Based Approaches », IEEE Access, vol. 5, p. 25189-25195, 2017, doi: 10.1109/ACCESS.2017.2763984.

[8] A. Gadducci, C. Barsotti, S. Cosio, L. Domenici, et A. Riccardo Genazzani, « Smoking habit, immune suppression, oral contraceptive use, and hormone replacement therapy use and cervical carcinogenesis: a review of the literature », Gynecol. Endocrinol., vol. 27, no 8, p. 597-604, août 2011, doi: 10.3109/09513590.2011.558953.

[9] P. Luhn et al., « The role of co-factors in the progression from human papillomavirus infection to cervical cancer », Gynecol. Oncol., vol. 128, no 2, p. 265-270, févr. 2013, doi: 10.1016/j.ygyno.2012.11.003.

[10] Harsh, S. ., Singh , D., & Pathak , S. (2022). Efficient and Cost-effective Drone – NDVI system for Precision Farming. International Journal of New Practices in Management and Engineering, 10(04), 14–19. https://doi.org/10.17762/ijnpme.v10i04.126

[11] « Cervical Cancer Prevention (PDQ®)–Health Professional Version - National Cancer Institute », 26 mars 2021. https://www.cancer.gov/types/cervical/hp/cervical-prevention-pdq (consulté le 11 avril 2021).

[12] S. Ganguly et al., « An Adaptive Threshold Based Algorithm for Detection of Red Lesions of Diabetic Retinopathy in a Fundus Image », p. 4, 2014.

[13] A. Agarwal, S. Gulia, S. Chaudhary, M. K. Dutta, C. M. Travieso, et J. B. Alonso-Hernandez, « A novel approach to detect glaucoma in retinal fundus images using cup-disk and rim-disk ratio », in 2015 4th International Work Conference on Bioinspired Intelligence (IWOBI), San Sebastian, Spain, juin 2015, p. 139-144. doi: 10.1109/IWOBI.2015.7160157.

[14] S. S. Han et al., « Keratinocytic Skin Cancer Detection on the Face Using Region-Based Convolutional Neural

Network », JAMA Dermatol., vol. 156, no 1, p. 29, janv. 2020, doi: 10.1001/jamadermatol.2019.3807.

[15] S. Graham et al., « Artificial Intelligence for Mental Health and Mental Illnesses: an Overview », Curr. Psychiatry Rep., vol. 21, no 11, p. 116, nov. 2019, doi: 10.1007/s11920-019-1094-0.

[16] F. Christopoulou, T. T. Tran, S. K. Sahu, M. Miwa, et S. Ananiadou, « Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods », J. Am. Med. Inform. Assoc., vol. 27, no 1, p. 39-46, janv. 2020, doi: 10.1093/jamia/ocz101.

[17] A. Singh, M. K. Dutta, R. Jennane, et E. Lespessailles, « Classification of the trabecular bone structure of osteoporotic patients using machine vision », Comput. Biol. Med., vol. 91, p. 148-158, déc. 2017, doi: 10.1016/j.compbiomed.2017.10.011.

[18] M. Kolařík, R. Burget, V. Uher, K. Říha, et M. Dutta, « Optimized High Resolution 3D Dense-U-Net Network for Brain and Spine Segmentation », Appl. Sci., vol. 9, no 3, p. 404, janv. 2019, doi: 10.3390/app9030404.

[19] K. Fernandes, J. S. Cardoso, et J. Fernandes, « Transfer Learning with Partial Observability Applied to Cervical Cancer Screening », p. 8.

[20] H. K. Fatlawi, « Enhanced Classification Model for Cervical Cancer Dataset based on Cost Sensitive Classifier », vol. 4, no 4, p. 6, 2017.

[21] Y. M. S. Al-Wesabi, A. Choudhury, et D. Won, « Classification of Cervical Cancer Dataset », p. 6.

[22] S. F. Abdoh, M. Abo Rizka, et F. A. Maghraby, « Cervical Cancer Diagnosis Using Random Forest Classifier With SMOTE and Feature Reduction Techniques », IEEE Access, vol. 6, p. 59475-59485, 2018, doi: 10.1109/ACCESS.2018.2874063.

[23] Asadi F, Salehnasab C, et Ajori L, « Supervised Algorithms of Machine Learning for the Prediction of Cervical Cancer », J. Biomed. Phys. Eng., vol. 10, no 4, août 2020, doi: 10.31661/jbpe.v0i0.1912-1027.

[24] M. F. Zorkafli, M. K. Osman, I. S. Isa, F. Ahmad, et S. N. Sulaiman, « Classification of Cervical Cancer Using Hybrid Multi-layered Perceptron Network Trained by Genetic Algorithm », Procedia Comput. Sci., vol. 163, p. 494-501, 2019, doi: 10.1016/j.procs.2019.12.132.

[25] Kumar, S., Gornale, S. S., Siddalingappa, R., & Mane, A. (2022). Gender Classification Based on Online Signature Features using Machine Learning Techniques. International Journal of Intelligent Systems and Applications in Engineering, 10(2), 260–268. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2020

[26] « UCI Machine Learning Repository: Cervical cancer (Risk Factors) Data Set », 2017. https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29 (consulté le 17 juillet 2021).

[27] N. V. Chawla, K. W. Bowyer, L. O. Hall, et W. P. Kegelmeyer, « SMOTE: Synthetic Minority Over-sampling Technique », J. Artif. Intell. Res., vol. 16, p. 321-357, juin 2002, doi: 10.1613/jair.953.

[28] EL-YAHYAOUI, A., & OMARY, F. (2022). An improved Framework for Biometric Database's privacy. International Journal of Communication Networks and Information Security (IJCNIS), 13(3). https://doi.org/10.17762/ijcnis.v13i3.5143

[29] J. Jeremiah Tanimu, M. Hamada, M. Hassan, et S. Yusuf Ilu, « A Contemporary Machine Learning Method for Accurate Prediction of Cervical Cancer », SHS Web Conf., vol. 102, p. 04004, 2021, doi: 10.1051/shsconf/202110204004.