Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 13(2), 747-759, 2023
Journal of the Institute of Science and Technology, 13(2), 747-759, 2023
ISSN: 2146-0574, eISSN: 2536-4618
Computer Engineering **DOI: 10.21597/jist.1222764**
Research Article
Received: 22.12.2022 **Accepted: 16.01.2023**
To Cite: Kılıçarslan, S., Göğebakan, M., & Közkurt, C. (2023). Cervical Cancer Prediction Using SMOTE Algorithm and Machine Learning Approaches. *Journal of the Institute of Science and Technology*, 13(2), 747-759.

Cervical Cancer Prediction Using SMOTE Algorithm and Machine Learning Approaches

Serhat KILIÇARSLAN[1*], Maruf Göğebakan[2], Cemil Közkurt[3]

Highlights:

- Majority voting
- SMOTE
- Classification

Keywords:

- Data mining
- Majority voting
- SMOTE algorithm
- Cervical cancer
- Classification

ABSTRACT:

Cervical cancer is one of the most successful types of treatment when diagnosed early. In this study, it is aimed to find and classify the disease with data mining methods on the digitized data set obtained as a result of the pap-smear test. Two-stage architecture has been proposed for the diagnosis of cervical cancer. In the first stage of the study, missing data were extracted from the used dataset, and in the second stage, a new dataset was obtained by using the Synthetic Minority Oversampling Technique (SMOTE) algorithm to balance the target classes in the dataset. By applying the majority voting (MV) method to the dataset used in the study, the structure with 4 target variables was reduced to a single target variable. On two data sets, Artificial Neural Network (ANN), Support Vector Machines (SVM), Decision Trees (DT), Random Forest (RF), and K-Nearest Neighbors (KNN) algorithms from data mining methods were used for the diagnosis of cervical cancer. The results obtained from the original dataset and the dataset produced with Smote were compared. ANN is the best method evaluated according to classification success and F-score, and the major voted target variable in the balanced data group produced with the Smote algorithm gave the most successful result. The experimental results showed that the use of MV and SMOTE algorithms together increased the classification success from 93% to 99%.

[1] Serhat KILIÇARSLAN (Orcid ID: 0000-0001-9483-4425), Bandirma Onyedi Eylul University, Department of Software Engineering, Türkiye

[2] Maruf GÖGEBAKAN (Orcid ID: 0000-0003-0447-8311), Bandirma Onyedi Eylul University, Department of Maritime Business Administiration, Maritime Faculty, Türkiye

[3]Cemil KÖZKURT (Orcid ID: 0000-0003-1407-9867), Bandirma Onyedi Eylul University, Department of Transportation Engineering, Türkiye

*** Corresponding Author:** Serhat KILIÇARSLAN, e-mail: skilicarslan@bandirma.edu.tr

INTRODUCTION

Cancer is expressed as malignant tumors that multiply uncontrollably in various parts of our body. Cancer is one of the most dangerous diseases that can cause serious illness and even death if left untreated. Cervical cancer is the fourth most common type of cancer worldwide, resulting in approximately 604,000 new cases and 342,000 deaths, according to the 2020 WHO report (Adem et al., 2019). Again, according to the WHO's 2020 report, about 90% of new cases and deaths occur in low- and middle-income countries. When cervical cancer is detected early, it has been observed that the treatment success rate is quite high. Although cervical cancer can be easily eradicated, it is still a serious threat to women's health (GÜRE et al., n.d.). Therefore, early diagnosis is very important for the treatment of cervical cancer. To reduce the number of deaths and diseases related to cervical cancer, many researchers around the world have started to conduct research. Decision support systems are being developed with the use of data mining methods for the early diagnosis of the disease (Tanimu et al., 2022). The diagnostic process developed with computer-aided decision systems helps doctors to examine and diagnose the pap-smear images of many patients in a short time.

Many data mining algorithms are based on the assumption that the distribution of intraclass labels in data sets is balanced. Imbalanced class distribution is one of the situations encountered in many real-life problems such as disease search and spam filtering (Liu et al., 2008). When trying to analyze (classify) imbalanced data sets with these algorithms, the classification success of the algorithms remains very low. For this reason, unbalanced data distributions cause the analyzes to obtain biased and erroneous accuracy values (He & Garcia, 2009). In imbalanced datasets, many different algorithms have been developed to multiply and stabilize the minority dataset. Among these methods, the SMOTE algorithm, which was developed based on the nearest neighbor algorithm, is among the most used (Kartal & Özen, 2017). In the literature, methods such as SAE, DT, MLP, AdaBoost, and SVM have been used for the diagnosis of cervical cancer based on the clinical dataset obtained from Pap smear images (Adem et al., 2019; CH et al., 2022; Khanam, 2021; Ratul et al., 2022; Tanimu et al., 2022; Zhang et al., 2021). The cervical cancer image obtained by the Pap Smear method is shown in Figure 1. Studies that diagnosed cervical cancer on the clinical data set obtained from their images are presented below.

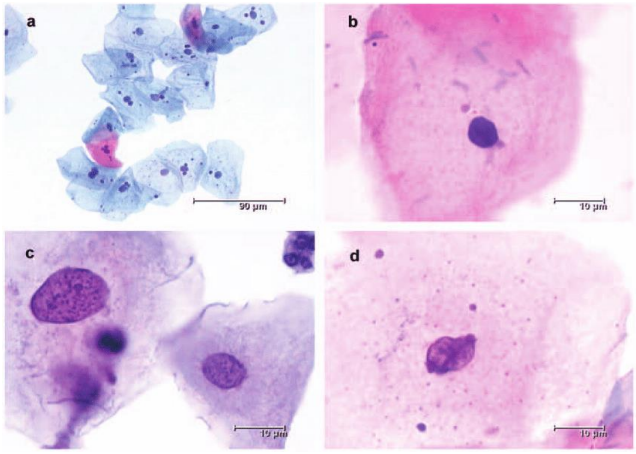

Figure 1. Pap-smear images

In the study of Ali et al. (2021), the cervical cancer clinical dataset was trained on 858 samples, 36 features, and 4 classes using Random Tree (RT), Random Forest (RF), and Instance-Based K-nearest neighbor (IBk) methods. According to the results obtained from the study, the RT method classified the biopsy and cytology classes well, while the RF method correctly classified the hinselmann and schiller classes (Ali et al., 2021). In the study of Islam et al. (2019), training was carried out using the Decision

Tree (DT), RF, Logistic Model Tree (LMT), and Artificial Neural Network (ANN) methods over the cervical cancer clinical dataset. As a result of the experimental evaluation in this study, the RF method correctly classified the biopsy and hinselmann classes, the LMT method Schiller and DT, and the cytology classes (Islam et al., 2019).

The comparative results of the proposed data mining methods with the studies on the same subject in the literature are shown in Table 1.

Table 1. A comparison of the literature on cervical cancer

Authors	Year	Method	Remarks
Alam et al. **(Alam et al., 2019)**	2019	Boosted Decision tree (BDT), Random Forests (RF)	Best performance achieved in BTD.
Ilango and Nithya **(Nithya & Ilango, 2019)**	2019	DT, RF, KNN, SVM	DT and RF gave the best results
Yang et al. **(Yang et al., 2019)**	2019	MLP	The most associated risk factors for cervical cancer achieved with MLP.
Mudawi and Alazeb **(Al Mudawi & Alazeb, 2022)**	2022	DT, SVM, KNN, RF	Best performance achieved with SVM.
Abdullah et al. **(Abdullah et al., 2019)**	2019	SVM, RF	Best performance achieved with RF.
Suman and Hooda **(Suman & Hooda, 2019)**	2019	SVM, Bayes Net, Naive Bayes, RF, MLP, J48	Best performance achieved with Bayes Net.
Karani et al. **(Karani et al., 2022)**	2022	KNN, RF, SVM, Logistic Regresion	Best performance achieved with SVM.
Gan et al. **(Gan et al., 2020)**	2020	Cost-sensitive classification algorithm	Better success by editing unbalanced data.
Ilyas and Ahmad **(Ilyas & Ahmad, 2021)**	2021	DT, SVM, RF, KNN, NB	Best performance achieved with SVM.
Proposed Model	2022	ANN, SVM, DT, RF, KNN	Best performance achieved with ANN-MVS

Cervical cancer is the most common type of cancer that women are affected by the increasing number of cancer cases today. A large number of women worldwide suffer from cervical cancer (EYÜPOĞLU, 2020). Therefore, the main motivation of this study is to achieve successful results in early diagnosis by minimizing the classification error caused by imbalanced data in data mining methods for the diagnosis of cervical cancer. For this purpose, a two-stage architecture is proposed to perform the diagnosis and diagnosis of the disease on the cervical cancer dataset. In the first phase of the study, data with missing observations on the cervical cancer dataset were extracted. In the second stage, the SMOTE algorithm was used to balance the classes of the extracted data set. As a result of both stages, new datasets were obtained. On the two new datasets obtained, artificial neural network (ANN), support vector machine (SVM), decision tree (DT), random forest (RF), and nearest neighbor (KNN) algorithms from data mining methods were used for the diagnosis of cervical cancer. Performance comparisons were made by applying each model used in the study to the dataset. Therefore, data mining methods are used to aid early detection in the diagnosis of cervical cancer. Balancing the cancer data and reducing the four target variables used for diagnosis to a single class with a statistical method based on frequency distributions increased the classification success. With the data mining methods applied on the target variable of the class distributions balanced with the SMOTE algorithm, combined with majority voting, the accuracy in early diagnosis increases the success rate to approximately 99%.

In the second part of the study, the cervical cancer dataset and methods used are given. In the third chapter, experimental evaluation results and discussion are given. In the last section, conclusions and future work are presented.

MATERIALS AND METHODS

In the study, artificial neural networks (ANN), support vector machine (SVM), decision tree (DT), random forest (RF), and nearest neighbor (KNN) algorithms from data mining methods were used in the classification of cervical cancer data.

ANN: They are mathematical models developed to enable nervous systems such as the human brain to make learning predictions on computers like humans, inspired by information processing methods (Mitchell et al., 1990).

SVM: It is a supervised learning algorithm based on statistical learning/Vapnik-Chervonenkis (VC) theory. It aims to separate the data belonging to two classes in the most appropriate way based on the risk minimization principle (Cortes & Vapnik, 1995).

RF: It is an algorithm that uses more than one decision tree and works based on the same logic as the Decision Trees algorithm. It is aimed to increase classification success by creating multiple decision trees.

DT: It is an algorithm that determines the tree structure from the top to the bottom in classifying the data. The tree structure is named roots, branches, and leaves, starting from the top. The nodes in the tree structure correspond to the leaves in the tree to show the rules (Kotsiantis, 2013).

k-NN: It is one of the most frequently used sample-based unsupervised learning algorithms in solving classification problems. The learning process is performed with the training set taken from the data set, and the data is trained based on the similarity of the distances of the k-nearest data. In this algorithm, it is determined which class the incoming sample belongs to by looking at the k nearest neighbors (Dudani, 1976). With $X = x_1, \dots, x_n$ and $Y = y_1, \dots, y_n$ being the set of points belonging to two classes, the distance between them is obtained as in (1).

$$\|XY\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

The majority voting method was applied to reduce the target variables of cervical cancer data to a single class based on frequency distributions. In addition, to eliminate the imbalanced cluster distribution problem among the features in the data set, synthetic data was produced with the SMOTE algorithm to produce minority class members.

Majority Voting

Cervical cancer data consisting of 668 observations, 30 features, and 4 target variables were classified by data mining methods based on statistical learning. A new target variable was obtained from the Hinselmann (C_H), Schiller (C_S), Citology (C_C) and Biopsy (C_B) target variables (label vector) in the data set by the majority voting (MV) method based on frequency distribution. The majority voting classifier algorithm specifically compares the results of each class label and decides on the class with the most votes (Lam & Suen, 1997). The majority voting flow diagram is presented in Figure 2.

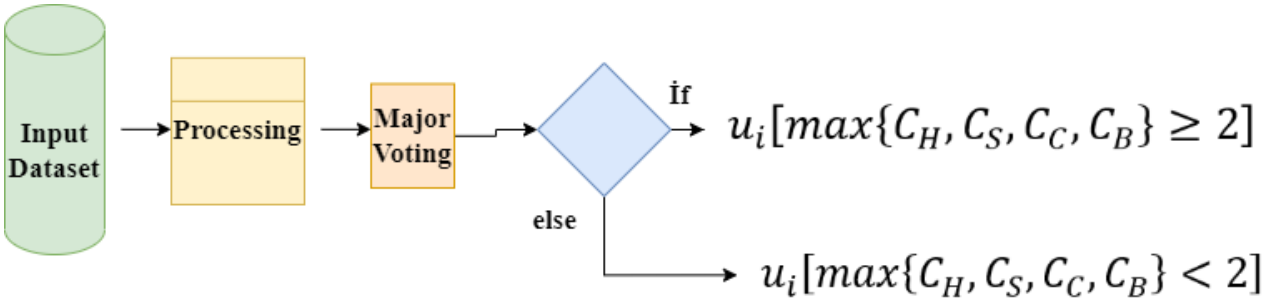

Figure 2. Majority voting flow diagram

In Figure 2, If the weighted class labels take values as $u_i[max\{C_H, C_S, C_C, C_B\} \geq 2]$, they are assigned to the C_1 label vector, while if values are obtained as $u_i[max\{C_H, C_S, C_C, C_B\} < 2]$, they are assigned to the C_0 label vector. Here, the tag vectors C_H, C_S, C_C and C_B represent the classes Hinselmann, Schiller, Citology and Biopsy, respectively. Classes C_0 and C_1 represent the label vectors for majority voting.

Synthetic Minority Over-sampling Technique (SMOTE):

The SMOTE algorithm method is among the best data generation methods to minimize the error in imbalanced cluster distributions in classification. Unlike algorithms that generate random samples, it works with the logic of generating members based on the closest minority members by using the k-NN method from minority members in class memberships. The most important difference from other synthetic data generation methods is that instead of producing minority class tags by copying method, synthetic members are obtained according to their distance from their nearest neighbors. (Chawla et al., 2002)**.** The minority class member generation diagram based on SMOTE and k-NN algorithms is shown in Figure 3 as follows.

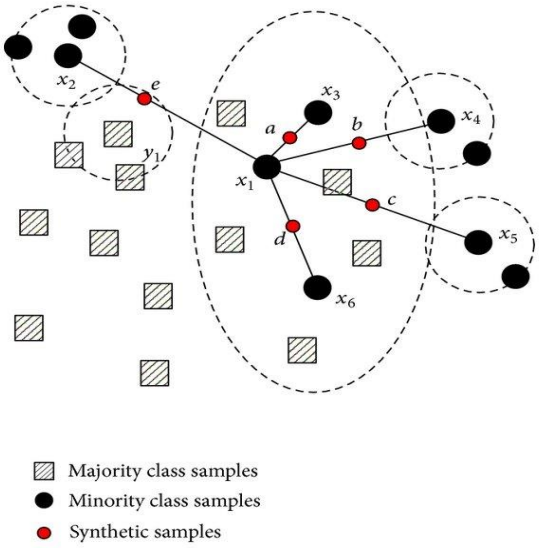

Figure 3. Minority class member generation diagram based on SMOTE algorithm (Hu & Li, 2013)

As demonstrated in Figure 2, while learning from unbalanced data, there is a higher likelihood of being close to a negative example and even being close to the mode of the positive distribution for a brand-new query x because there aren't many positive instances in the training set. The suggested method entails altering the spacing between the instances by the class structure.

Algorithm 1: Generating minority class synthetic data with SMOTE algorithm

Steps	Descriptions of Algorithm steps
Step-1	The k-NN members of each member in the minority class are appointed.
Step-2	The difference of distances in $\mathbb{R}^n$ is calculated by the equation$(\sum_{i=1}^{n}\|x_i - y_i\|^p)^{1/p}$ of the minority class member and the k-NN members.
Step-3	The class label α is multiplied by the value (distance) obtained from Step-2, which will be $\forall\alpha \in \{0,1\}$.
Step-4	A new class member is obtained with the $x_{new} = x_i + (x_j - x_i)\alpha$ equation
Step-5	Step 1-4 continues until the desired number of class members is produced.
Step-6	The algorithm is finished.

Cervical Cancer Data

In the study, a dataset with 858 observations, 32 features, and 4 target variables was obtained from the UC Irvine Machine Learning Repository database for the classification of cervical cancer disease by

classical data mining methods (Fernandes et al., 2017). “Time since first diagnosis” and “Time since last diagnosis” features from 32 features in the study were removed from the feature list to protect the confidentiality of patients' personal information. In addition, 190 lines of missing observations were excluded from the data obtained from the patients. Thus, experimental studies were carried out with the data set consisting of the remaining 668 observations and 30 features. The four target variables in the dataset are named Hinselmann, Schiller, Cytology, and Biopsy. Table 2 describes the numerical definition of the attributes of the dataset.

Table 2. Description of the features in the data set after preprocessing

Attributes	Mean	S. Dev.	Attributes	Mean	S. Dev.
Age	26.8205	8.497948	STDs: vaginalcondylomatosis	0.004662	0.068159
# of partners	2.51165	1.644759	STDs: vulva-perinealcondylomatosis	0.050117	0.218313
Age of 1st intercourse	16.9790	2.797653	STDs: syphilis	0.020979	0.143398
# of pregnancies	2.19230	1.434395	STDs: pelvic inflammatory disease	0.001166	0.034139
Smokes	0.14335	0.350641	STDs: genital herpes	0.001166	0.034139
Smokes years	1.20124	4.060623	STDs: molluscumcontagiosum	0.001166	0.034139
Smokes packs/year	0.44627	2.210351	STDs: AIDS	0	0
Hormonal contraceptives	0.68648	0.464194	STDs: HIV	0.020979	0.143398
Hormonal contraceptives years	1.97239	3.597888	STDs: Hepatitis B	0.001166	0.034139
IUD	0.09673	0.295771	STDs: HPV	0.002331	0.048252
IUD years	0.44460	1.814218	STDs: Number of diagnosis	0.087413	0.302545
STDs	0.09207	0.2893	Dx: Cancer	0.020979	0.143398
STDs number	0.15501	0.529617	Dx: CIN	0.01049	0.101939
STDs: condylomatosis	0.05128	0.220701	Dx: HPV	0.020979	0.143398
STDs: cervicalcondylo-mitosis	0	0	DX	0.027972	0.164989

RESULTS AND DISCUSSION

This work is carried out in Spyder 3.10.8 Python development environment on a laptop Intel (R) Core (TM) i5-10400 CPU@2.90 GHz and 8-GB RAM running on Windows 10. In the study, the processing steps of the decision support system developed for the detection of cervical cancer are shown in Figure 4.

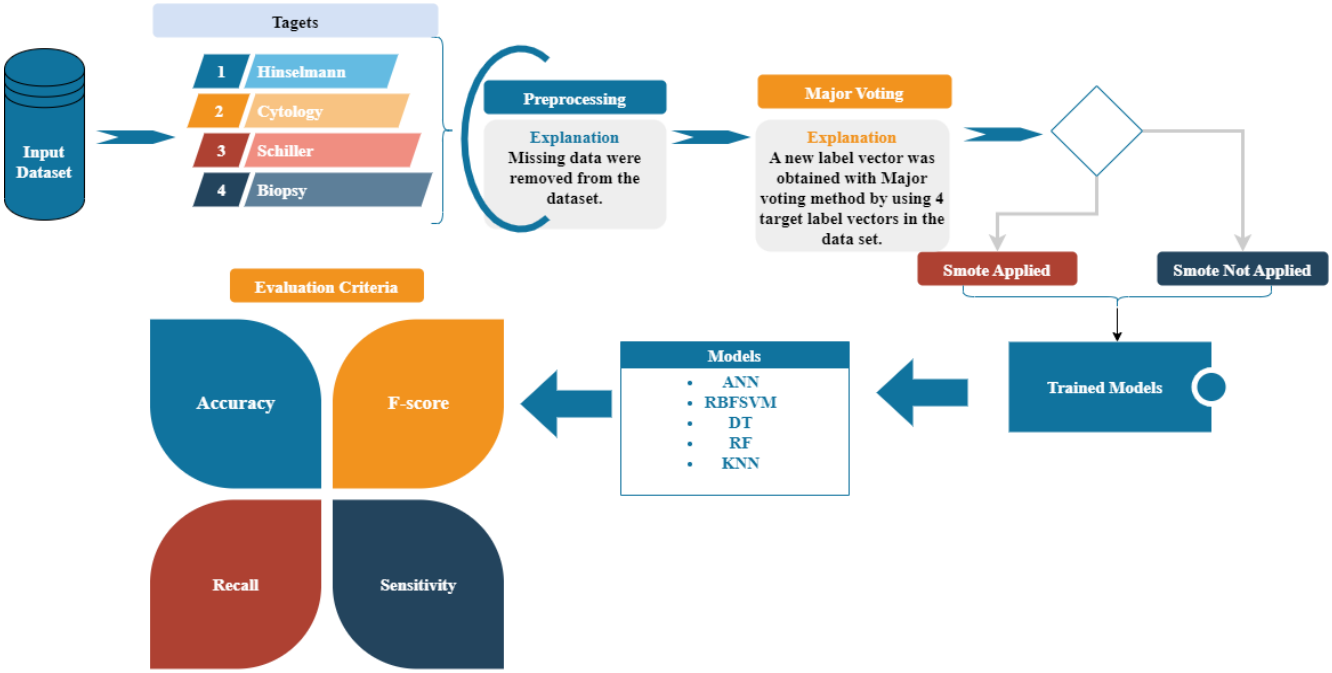

Figure 4. Processing steps performed for the detection of cervical cancer data

Experimental evaluations were carried out using the k-fold=10 cross-validation method of default values in the Scikit-learn [34] library as the parameters of the data mining methods used in the study.

In the study, the performances of various data mining algorithms are compared and the classification successes of these algorithms are given. The performances of the classification algorithms were compared according to accuracy, precision, recall, and f-score criteria in Equations 2, 3, 4, and 5, respectively. Accuracy is the difference between the actual value in the measurement of the physical property and the value indicated by the device. Precision, on the other hand, shows how many of the values we estimated as Positive are Positive. Recall, on the other hand, is a metric that shows how much of the operations we need to estimate as Positive, we estimate as Positive. The F1 Score value shows us the harmonic mean of Precision and Recall values.

$$\text{accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

$$\text{precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$F - score = \frac{2 * precision * recall}{precision + recall} \tag{5}$$

To test the performances of machine learning algorithms, the k-fold crossover method is used as shown in Figure 5.

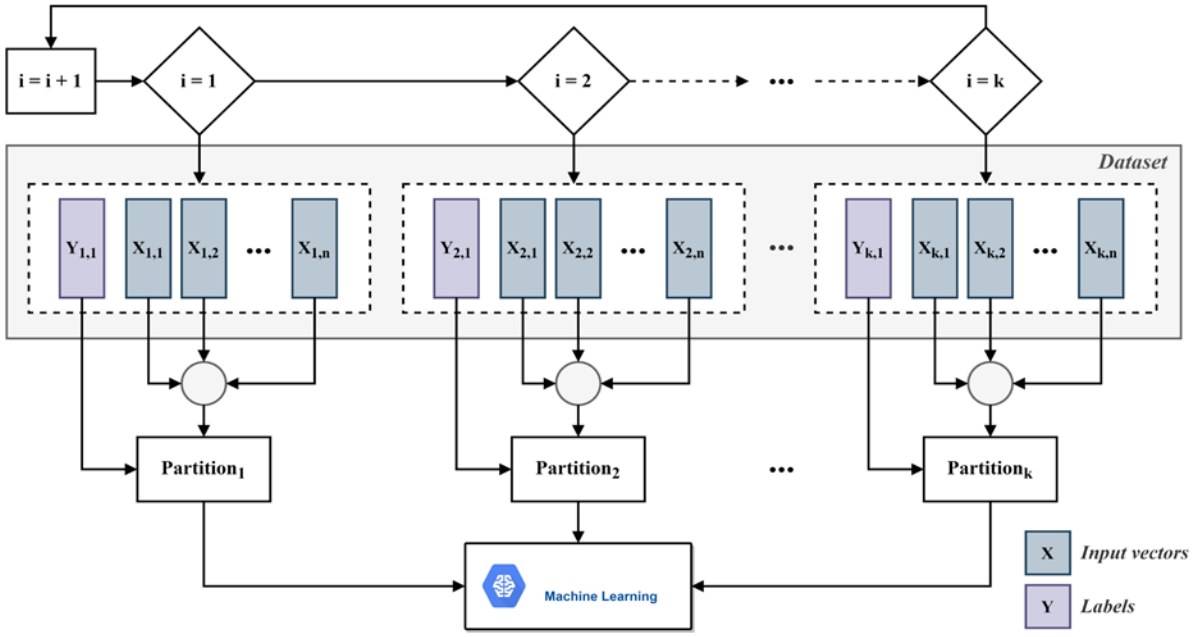

Figure 5. k-fold cross-validation of the dataset [Elen et. al., 2022]

Evaluation Indices

Accuracy is one of the most widely used evaluation criteria to measure classification success. It shows the rate of correct results achieved by data mining methods. The higher the accuracy achieved by an algorithm, the better the algorithm's performance. When there is no statistically significant difference between the accuracy values, the evaluation of the results may mislead the researchers. In this case, different evaluation criteria such as the Jaccard index, F-score, recall, and precision can be used to compare the results (Adem et al., 2019; Sharma & Seal, 2021a). In the evaluation of the results obtained from the data, Wilcoxon's signed-rank test, Wilcoxon's rank-sum test, and sign tests, which are non-

parametric hypothesis tests for two dependent samples, are evaluated at the 5% significance level (Sharma & Seal, 2021b).

In the study, experimental analyzes were carried out using data mining methods such as ANN, RBFSVM, DT, RF, and KNN. In this section, a two-stage architecture is proposed to be able to diagnose and diagnose the disease on the cervical cancer dataset. In the first stage of the study, cervical cancer data were preprocessed and experimental results were obtained. The results obtained are presented in Table 3. In the second stage, the classes in the dataset were brought into a balanced class structure with the SMOTE algorithm, experimental evaluations were made and presented in Table 4. The structure with 4 target variables was reduced to a single target variable by applying the MV method to the dataset used.

Table 3. Evaluation of Cervical Cancer Data by Data Mining Methods

			Hinselmann	
	Accuracy	**Precision**	**Recall**	**F-score**
ANN	0.9797	0.880	0.6833	0.7100
RBFSVM	0.9775	0.8500	0.7167	0.7433
DT	0.9685	0.8500	0.80	0.8800
RF	0.9706	0.9999	0.600	0.7200
KNN	0.9662	0.8000	0.8500	0.8000
			Schiller	
	Accuracy	**Precision**	**Recall**	**F-score**
ANN	0.9751	0.7417	0.8300	0.7746
RBFSVM	0.9774	0.8000	0.9400	0.8324
DT	0.9706	0.8000	0.8788	0.7997
RF	0.9752	0.8000	0.9267	0.8235
KNN	0.9345	0.8500	0.8167	0.8271
			Cytology	
	Accuracy	**Precision**	**Recall**	**F-score**
ANN	0.9120	0.8167	0.8175	8186
RBFSVM	0.9414	0.8391	0.8300	0.8692
DT	0.9188	0.8250	0.8250	0.8223
RF	0.9119	0.8333	0.8917	0.8805
KNN	0.9369	0.8667	0.8015	0.8090
			Biopsy	
	Accuracy	**Precision**	**Recall**	**F-score**
ANN	0.9593	0.9247	0.9117	0.9339
RBFSVM	0.9683	0.8000	0.8683	0.8598
DT	0.9526	0.8500	0.8767	0.8260
RF	0.9752	0.9333	0.8917	0.8360
KNN	0.9435	0.8167	0.8333	0.8871
			MV	
	Accuracy	**Precision**	**Recall**	**F-score**
ANN	**0.9910**	**0.9000**	**0.7500**	**0.9257**
RBFSVM	0.9842	1.0000	0.8650	0.9163
DT	0.9821	0.8083	0.8250	0.8150
RF	0.9842	1.0000	0.8660	0.9163
KNN	0.9503	0.8464	0.8667	0.8224

When Table 3 is examined, classification values for five target variables Hinselmann, Schiller, Cytology, Biopsy, and MV were obtained with five different data mining methods such as ANN, RBFSVM, DT, RF, and KNN. According to the results obtained, it was determined that the MV target variable showed the best classification performance with 0.9910 accuracy and 0.9257 F-score values with the ANN method. As a result, the MV variable obtained by the weighting method from the Hinselmann, Schiller, Cytology, and Biopsy target variables positively affects the classification success. The ROC curve and confusion matrix of the classification success between MV and ANN are presented in Figure 6.

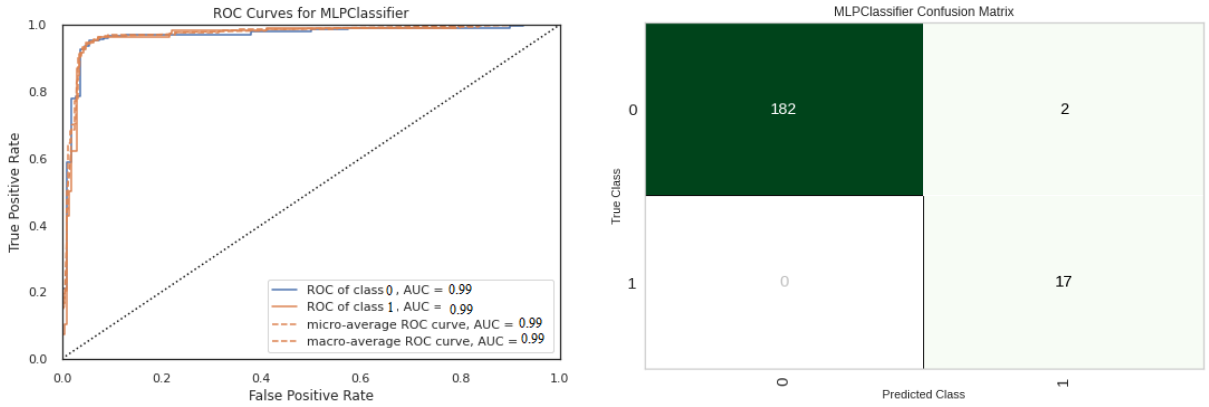

Figure 6. Classification success of ANN-MV **a)** Roc curve **b)** Confusion Matrix

After reducing the four target variables in the cervical cancer data to a single target variable, the imbalanced class structure in the data set was obtained by generating synthetic minority class members with the SMOTE algorithm. Classification success was measured by applying data mining methods to the balanced data set and the results are shown in Table 4.

Table 4. Evaluation of Cervical Cancer Data Balanced with SMOTE algorithm by Data Mining Methods

		Hinselmann		
	Accuracy	Precision	Recall	F-score
ANN	0.9420	0.8923	0.8287	0.8549
RBFSVM	0.9188	0.9154	0.7354	0.8126
DT	0.9261	0.8154	0.8008	0.8016
RF	0.9415	0.9231	0.7281	0.8061
KNN	0.9275	0.8769	0.7875	0.8234
		Schiller		
	Accuracy	Precision	Recall	F-score
ANN	0.9696	0.8724	0.9592	0.9639
RBFSVM	0.9323	0.9690	0.9449	0.9559
DT	0.9710	0.9655	0.9666	0.9656
RF	0.9638	0.9759	0.9411	0.9578
KNN	0.9145	0.9552	0.8616	0.9042
		Cytology		
	Accuracy	Precision	Recall	F-score
ANN	0.9333	0.6100	0.6062	0.5848
RBFSVM	0.9478	0.5900	0.7293	0.6323
DT	0.9116	0.5567	0.4825	0.4888
RF	0.9348	0.5567	0.6454	0.5707
KNN	0.9493	0.5533	0.7931	0.6113
		Biopsy		
	Accuracy	Precision	Recall	F-score
ANN	0.9551	0.9877	0.8981	0.9400
RBFSVM	0.9362	0.9453	0.8663	0.9162
DT	0.9420	0.9182	0.9197	0.9180
RF	0.9565	0.9917	0.8989	0.9423
KNN	0.9087	0.9552	0.8230	0.8819
		MV		
	Accuracy	Precision	Recall	F-score
ANN	0.9899	0.9964	0.9791	0.9875
RBFSVM	0.9884	1.0000	0.9723	0.9858
DT	0.9870	0.9927	0.9756	0.9839
RF	0.9870	0.9854	0.982	0.9836
KNN	0.9478	0.9709	0.9074	0.9374

Table 4 shows the experimental results of the cervical cancer data in which the SMOTE algorithm and the minority class data were balanced. The target variables of the balanced data, named Hinselmann_S, Schiller_S, Citology_S, Biopsy_S, and MV_S, were classified by data mining methods

ANN, RBFSVM, DT, RF, and KNN. According to the results obtained, the best classification performance was measured with 0.9899 accuracy and 0.9875 F-score values between MV_S and ANN. As a result, the MV_S variable obtained from Hinselmann_S, Schiller_S, Citology_S, and Biopsy_S target variables by weighting and SMOTE dataset balancing method had a positive effect on classification success. The ROC curve and confusion matrix of the classification success between MV_S and ANN are presented in Figure 7.

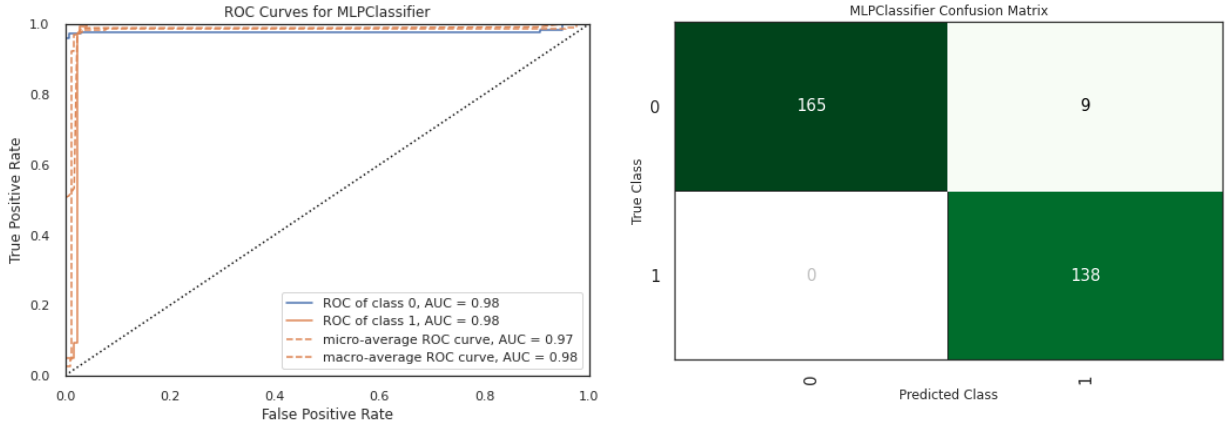

Figure 7. Classification success of ANN-MV_S **a)** ROC curve **b)** Confusion matrix

Experimental evaluation results are given in Tables 3 and 4. In the two-stage experimental evaluation, data mining methods were applied to the raw data, and the data set was balanced with SMOTE. In both stages, classification success was measured according to 5 target variables, including 4 target variables and MV target variable reduced by MV. In the first stage, it was seen that the classification between ANN and MV (99%-93%) was the most successful according to the accuracy and f-score values between the machine learning method and target variables in the raw data. Likewise, when looking at the second stage, it was seen that the best result was obtained between ANN and MV SMOTE (MV_S) (99%-99%) in terms of accuracy and F-score.

In the literature, there are evaluations based on accuracy and f-score results. Among the results obtained, the accuracy between MV and MV_S according to the ANN method and the statistical analysis and comparison results according to the F-score are presented in Table 5.

Table 5. Descriptive statistics of MV and MV_S variables according to Accuracy and F-score values

Descriptive Statistics					
	N	**Mean**	**Std. Deviation**	**Minimum**	**Maximum**
MV_S F-score	5	0.9756	0.02143	0.94	0.99
MV_S Accuracy	5	0.9800	0.01805	0.95	0.99
MV F-score	5	0.8791	0.05537	0.82	0.93
MV Accuracy	5	0.9784	0.01604	0.95	0.99

Non-parametric tests were applied for dependent variables according to accuracy and F1-score evaluation criteria, and the results are presented in Table 6.

Table 6. Statistical analysis for Accuracy and F1-score evaluation criteria

Test Statistics[a]		
	MV_F – MV_SF	**MV_A – MV_SA**
Z	-2,023[b]	-1,214[b]
Asymp. Sig. (2-tailed)	,043	,225

When Table 6 is examined, Wilcoxon Signed Ranks Test was performed to determine whether there is a statistically significant difference between the accuracy and f-score values obtained by the MV and MV_S methods. As a result of the test in Table 6, it was determined that the f-score values obtained according to the MV_S variable were significantly higher than the f-score values obtained according to

the MV variable ($p<0.05$). In other words, considering the f-score criterion according to the ANN method, the classification success of the MV_S target variable is higher than the classification success of the MV variable. From this point of view, it can be said that applying the smote algorithm to the MV method has a positive effect on the f-score, which is an important indicator in separating the labels in classification. On the other hand, according to the results obtained with the MV variable and MV_S variable, no statistically significant difference was found in terms of the Accuracy criterion ($p>0.05$). In other words, it can be said that applying the SMOTE algorithm to the MV method has no effect on Accuracy values, which is an indicator of classification success.

According to the data mining methods applied to cervical cancer data, it has been seen that reducing the target variables to a single variable with Major Voting and balancing them with the SMOTE algorithm increases the classification success. Classification of MV_S target variable with ANN according to f-score values was statistically significant.

CONCLUSION

Cancer is expressed as malignant tumors that multiply uncontrollably in various parts of our body. Although cervical cancer is a common and lethal type of cancer worldwide, it can be treated most successfully when diagnosed early. In this study, disease identification and classification were carried out by data mining methods on the digitized cancer dataset obtained as a result of the pap-smear test. The diagnosis of cervical cancer is mostly based on the experience and knowledge of medical doctors. Therefore, the development of decision support systems for the accurate prediction of cancer is of great importance in helping doctors diagnose and treat the disease. Thanks to the developed methods and algorithms, it is important to increase the success of the prediction of cervical cancer and to prevent the wrong treatments to be applied to the patients. The classification method proposed in this study produced successful results for disease diagnosis.

The dataset used in the study was reduced in size with the majority voting (MV) method, and the 4 target variables were reduced to a single target variable. In addition, the imbalanced class structure in the variables was balanced using the SMOTE algorithm. On the new dataset obtained with the data preprocessing, artificial neural network (ANN), support vector machine (SVM), decision tree (DT), random forest (RF), and k nearest neighbor (k-NN) methods for the diagnosis of cervical cancer were used. The experimental results showed that the use of MV and SMOTE algorithms together increased the classification success from 93% to 99%. The best classification performance was obtained with ANN. In future studies, it is aimed to test MV and SMOTE algorithms on image processing methods.

Conflict of Interest

The article authors declare that there is no conflict of interest between them.

Author's Contributions

The authors declare that they have contributed equally to the article.

REFERENCES

Abdullah, A. A., Sabri, N. A., Khairunizam, W., Zunaidi, I., Razlan, Z. M., & Shahriman, A. B. (2019). Development of predictive models for cervical cancer based on gene expression profiling data. In *IOP Conference Series: Materials Science and Engineering* (Vol. 557, p. 012003). IOP Publishing.

Adem, K., Kiliçarslan, S., & Cömert, O. (2019). Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification. *Expert Systems with Applications, 115*, 557–564. https://doi.org/10.1016/j.eswa.2018.08.050

Akyol, F. B., & Altun, O. (2020). Detection of cervix cancer from pap-smear images. *Sakarya University Journal of Computer and Information Sciences*, *3*(2), 99–111.

Al Mudawi, N., & Alazeb, A. (2022). A Model for Predicting Cervical Cancer Using Machine Learning Algorithms. *Sensors*, *22*(11), 4132.

Alam, T. M., Khan, M. M. A., Iqbal, M. A., Abdul, W., & Mushtaq, M. (2019, October 23). Cervical Cancer Prediction through Different Screening Methods Using Data Mining. SSRN Scholarly Paper, Rochester, NY. Retrieved from https://papers.ssrn.com/abstract=3474371

Ali, M. M., Ahmed, K., Bui, F. M., Paul, B. K., Ibrahim, S. M., Quinn, J. M. W., & Moni, M. A. (2021). Machine learning-based statistical analysis for early stage detection of cervical cancer. *Computers in Biology and Medicine*, *139*, 104985. https://doi.org/10.1016/j.compbiomed.2021.104985

Allehaibi, K. H. S., Nugroho, L. E., Lazuardi, L., Prabuwono, A. S., & Mantoro, T. (2019). Segmentation and classification of cervical cells using deep learning. *IEEE Access*, *7*, 116925–116941.

CH, N., Sai, P. P., Madhuri, G., Reddy, K. S., & BharathSimha Reddy, D. V. (2022). Artificial Intelligence based Cervical Cancer Risk Prediction Using M1 Algorithms. In *2022 International Conference on Emerging Smart Computing and Informatics (ESCI)* (pp. 1–6). Presented at the 2022 International Conference on Emerging Smart Computing and Informatics (ESCI). https://doi.org/10.1109/ESCI53509.2022.9758241

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Chen, W., Shen, W., Gao, L., & Li, X. (2022). Hybrid Loss-Constrained Lightweight Convolutional Neural Networks for Cervical Cell Classification. *Sensors*, *22*(9), 3272. https://doi.org/10.3390/s22093272

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, (4), 325–327.

Elen, A., Baş, S., & Közkurt, C. (2022). An Adaptive Gaussian Kernel for Support Vector Machine. *Arabian Journal for Science and Engineering*. https://doi.org/10.1007/s13369-022-06654-3

Eyüpoğlu, C. (2020). Korelasyon Temelli Özellik Seçimi, Genetik Arama ve Rastgele Ormanlar Tekniklerine Dayanan Yeni Bir Rahim Ağzı Kanseri Teşhis Yöntemi. *Avrupa Bilim ve Teknoloji Dergisi*, (19), 263–271.

Fernandes, K., Cardoso, J. S., & Fernandes, J. (2017). Transfer learning with partial observability applied to cervical cancer screening. In *Iberian conference on pattern recognition and image analysis* (pp. 243–250). Springer.

Gan, D., Shen, J., An, B., Xu, M., & Liu, N. (2020). Integrating TANBN with cost sensitive classification algorithm for imbalanced data in medical diagnosis. *Computers & Industrial Engineering*, *140*, 106266.

Güre, M. D. P., Karataş, M., & Başcıllar, M. (2022). "HPV Aşısı Haktır": Halk Sağlığı Sosyal Hizmeti Perspektifinden HPV İle İlgili Tweetlerin Analizi. *Toplum ve Sosyal Hizmet*, *33*(3), 955–973.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263–1284.

Hu, F., & Li, H. (2013). A novel boundary oversampling algorithm based on neighborhood rough set model: NRSBoundary-SMOTE. *Mathematical Problems in Engineering*, *2013*.

Ilyas, Q. M., & Ahmad, M. (2021). An enhanced ensemble diagnosis of cervical cancer: a pursuit of machine intelligence towards sustainable health. *IEEE Access*, *9*, 12374–12388.

Islam, A.-U.-, Ripon, S. H., & Bhuiyan, N. Q. (2019). Cervical Cancer Risk Factors: Classification and Mining Associations. *APTIKOM Journal on Computer Science and Information Technologies*, *4*(1), 8–18.

Karani, H., Gangurde, A., Dhumal, G., Gautam, W., Hiran, S., & Marathe, A. (2022). Comparison of Performance of Machine Learning Algorithms for Cervical Cancer Classification. In *2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)* (pp. 1–7). IEEE.

Kartal, E., & Özen, Z. (2017). Dengesiz veri setlerinde sınıflandırma. *Mühendislikte Yapay Zekâ ve Uygulamaları, 1st ed., O. Torkul, S. Gülseçen, Y. Uyaroğlu, G. Çağıl, and MK Uçar, Eds. Sakarya: Sakarya Üniversitesi Kütüphanesi Yayınevi*, *109*, 131.

Khanam, F. (2021). Prediction of cervical cancer in Bangladesh using hybrid machine learning algorithms. Retrieved from http://lib.buet.ac.bd:8080/xmlui/handle/123456789/6030

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, *39*(4), 261–283.

Lam, L., & Suen, S. Y. (1997). Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *27*(5), 553–568.

Liu, X.-Y., Wu, J., & Zhou, Z.-H. (2008). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*(2), 539–550.

Mitchell, T., Buchanan, B., DeJong, G., Dietterich, T., Rosenbloom, P., & Waibel, A. (1990). Machine learning. *Annual review of computer science*, *4*(1), 417–433.

Nithya, B., & Ilango, V. (2019). Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction. *SN Applied Sciences*, *1*(6), 1–16.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Dubourg, V. (2011). " Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, p.

Ratul, I. J., Al-Monsur, A., Tabassum, B., Ar-Rafi, A. M., Nishat, M. M., & Faisal, F. (2022). Early risk prediction of cervical cancer: A machine learning approach. In *2022 19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)* (pp. 1–4). Presented at the 2022 19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). https://doi.org/10.1109/ECTI-CON54298.2022.9795429

Sharma, K. K., & Seal, A. (2021). Multi-view spectral clustering for uncertain objects. *Information Sciences*, *547*, 723–745.

Sharma, K. K., & Seal, A. (2021). Outlier-robust multi-view clustering for uncertain data. *Knowledge-Based Systems*, *211*, 106567

Suman, S. K., & Hooda, N. (2019). Predicting risk of Cervical Cancer: A case study of machine learning. *Journal of Statistics and Management Systems*, *22*(4), 689–696.

Tanimu, J. J., Hamada, M., Hassan, M., Kakudi, H., & Abiodun, J. O. (2022). A Machine Learning Method for Classification of Cervical Cancer. *Electronics*, *11*(3), 463. https://doi.org/10.3390/electronics11030463

Yang, W., Gou, X., Xu, T., Yi, X., & Jiang, M. (2019). Cervical cancer risk prediction model and analysis of risk factors based on machine learning. In *Proceedings of the 2019 11th International Conference on Bioinformatics and Biomedical Technology* (pp. 50–54).

Zhang, L., Zhu, Y., Song, Y., Han, Y., Sun, D., Qin, S., & Gao, Y. (2021). Intelligent Diagnosis of Cervical Cancer Based on Data Mining Algorithm. *Computational and Mathematical Methods in Medicine*, *2021*.