EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

Data Driven Prognosis of Cervical Cancer Using Class Balancing and Machine Learning Techniques

Mamta Arora[1,2,*], Sanjeev Dhawan[3] and Kulvinder Singh[3]

[1]Ph.D. Research Scholar, Department of Computer Science & Engineering, University Institute of Engineering & Technology (U.I.E.T), Kurukshetra University, Kurukshetra.

[2]Assistant Professor, Department of Computer Science and Technology, Manav Rachna University, Faridabad

[3]Faculty of Computer Science & Engineering, Department of Computer Science & Engineering, University Institute of Engineering & Technology (U.I.E.T), Kurukshetra University, Kurukshetra

Abstract

INTRODUCTION: With the progression of innovation and its joint effort with health care services, the world has achieved a lot of benefits. AI procedures and machine learning techniques are constantly improving existing statistical methods for better results in the medical field. These improved methods will assist health care providers in providing intelligent medical services.
OBJECTIVES: This Cervical cancer is the fourth most common cancer among the other female cancers. This cancer is preventable with early diagnosis. This reason becomes the motivation of the research work. For efficiently and timely prognosis of cervical cancer require a computer-assisted algorithm
METHODS: The work demonstrated in this paper contributes to the techniques of machine learning for diagnosing cervical cancer. The machine learning algorithms used in this research are K Nearest Neighbour, Support Vector Machine and Random Forest Tree. These classification algorithms are used with class balancing techniques including under-sampling, Oversampling and SMOTE.
RESULTS: The evaluation metrics used for comparative analysis includes accuracy, sensitivity, specificity, negative predicted accuracy, and positive predictive accuracy. The results show the Random Forest algorithm with SMOTE technique delivered more promising results over SVM and KNN for four target variables Schiller, Biopsy, Hinselmann , and Cytology respectively.
CONCLUSION: It is concluded that with the limited amount of data which also suffers from the unbalancing problem the promising results drawn using the proposed model.

Keywords: Cervical Cancer, Random forest, Support vector machine, K-Nearest Neighbour, Random over-sampling, random under-sampling, SMOTE.

Received on 31 August 2019, accepted on 01 May 2020, published on 07 May 2020

doi: 10.4108/eai.13-7-2018.164264

*Corresponding author. Email:imamta.arora@gmail.com

1. Introduction

Cancer is the second leading driving factor for death all around the world. The facts show that approximately 9.6 million deaths in 2018 are due to cancer [1]. Globally, one out of six deaths is due to this malignant disease. Cervical cancer is ranked as the fourth most common cancer among other female cancers. Cervical cancer is the cancer of the cervix which is caused due to human

papillomavirus (HPV). This virus causes the abnormal development of cells in the cervix that can lead to the malignant stage. It takes one or two decades to reach from pre-cancerous to the cancerous stage. Thus it can be preventable with a timely diagnosis and prognosis. The machine learning techniques showed the state-of-art in various medical applications. This gives the motivation in developing a computer-based algorithm that can assist the health care providers in providing a timely diagnosis so that it can decrease the rate of mortality. The experimental work presented in this paper is done on the cervical cancer dataset publically available on the UCI repository [2]. Several machine learning classification algorithms are applied to achieve a higher rate of sensitivity so that no patient left untreated with cancer. The machine learning techniques used for this work are KNN, SVM-Linear, SVM-POLY and random forest. These classification algorithms are applied with the various class balancing techniques like random under sampling, random oversampling and SMOTE. The ensemble algorithm random forest with SMOTE gives promising results over the other classification methods. The uniqueness of this study exists in the fact that with the limited amount of data which also suffers from the unbalancing problem the promising results drawn using the proposed model. The work presented in this paper has implications for healthcare providers that can use this model for serving the mass population by giving a timely diagnosis to the patients. The presented work of the study is structured into four sections. The survey of the recently published papers is covered in section 2. Section 3 covers the machine learning and its approaches used in the study for conducting the experiments. Section 4 describes dataset, class balancing techniques and the experimental results. At last, the conclusion and future work are discussed in Section 5.

2. Related Work

In this section, we explained the publications that work on numerical clinical values. Table 1 summarizes a few recent publications in this domain. The attributes of the table include data source, ML technique, type of data, No. of patient records and Results. In 2019, Author W. Yang, Xin Gou et al. published a paper [3] in which they described the cervical cancer prediction model using a Multilayer perceptron and random forest approach. The experiment is conducted on the cervical cancer dataset publically available on the UCI repository. The dataset is consisting of 858 patients and 4 target variables. Random forest classifier reported the highest prediction accuracy as 97.6% for Hinselmann target variable.

In 2019, Dhwaani Parikh and Vineet Menon [4] demonstrated the use of various machine learning algorithms on the similar dataset of cervical cancer available on UCI repository. The authors reported that the K-NN model gives better accuracy, F1- Score, recall and precision. In the different studies, author Wu and Zhou [5] examined principal components analysis (PCA) for dimensionality reduction with SVM classifier for prediction of cervical cancer. They achieved Acc=90%, Sen=100%, Spec= 88% but their study lack the explanation of using PCA as feature selection.

In the study [6] the authors have presented a comparative analysis of 15 machine learning algorithms to diagnose cervical cancer. They have used the Pap smear benchmark database prepared by Herlev's university hospital. They applied 15 algorithms on two datasets namely old and a new dataset which consist of 500 and 917 single cell Pap smear images respectively. Among the 15 machine learning algorithms, Ensemble of Nested Dichotomies (END) outperformed for both dataset with the accuracy of 77.38% for the first dataset and 78.28% for the second dataset. On the other side, this study also shows that Naive Bayes is the worst performer with accuracy of about 50% and 60% on the first and second dataset respectively.

In [7] K. Hemlatha et al. demonstrated the experimental work on the modified cervical Pap smear dataset (MCPS). Authors applied four most commonly used neural networks namely Multi-Layer Perceptron (MLP), Radial Basis Function (RBF), Probabilistic Neural Network (PNN), and Linear Vector Quantization (LVQ) for classification. They have used the MCPS dataset with only 4 features. The MCPS dataset is obtained from their previous work [8] which deals with the segmentation of cervical images. This segmentation is obtained by applying the fuzzy Edge Detection method that segments the cytoplasm and nucleus part. They applied the Fuzzy edge detection method on the Pap smear old dataset (Herlev's University Hospital) which originally consisted of 917 images that are further described using 20 features. After this segmentation neural networks are applied on the modified dataset for classifying the entire data into two classes' i.e. normal and abnormal class. Each network is trained, tested and validated with 70%, 15%, and 15% sample size respectively. They have evaluated their performance using mean squared error (MSE). The algorithm with the less MSE is considered to be best amongst others. The study shows that BF has the highest classification accuracy with 100% but with a higher MSE value of 1. So RBF can't be used for classification as its classification will be more error-prone. But MLP gives the classification accuracy of 92.03% with the small MSE value of 0.0616. The work reveals that the MLP network outperformed among other three networks.

The published work in [9] is different from the above-stated work as they have used the biopsy test data instead of Pap smear data for the prediction of cervical cancer. For the sake of classifying the data into normal or cancer cervix, the authors applied a powerful data mining algorithm on biopsy numerical data. The data collected from NCBI (National Centre for Biotechnology Information) which consists of 500 records and 61 biopsy features including a gene identifier. They have selected a sample of 100 records for training and testing purposes. On the selected sample Classification and regression Tree

algorithm (CART), the Random Forest tree algorithm (RFT) and RFT with the K-Means learning algorithm were applied for classifying the data into the normal and cancerous cervix. The study reveals that the proposed hybrid algorithm RFT with k- means outperformed among CART and RFT with the accuracy of 96.77% on NCBI biopsy data for prediction of cervical cancer.

In the study [10] authors Prableen Kaur and Manik Sharma presented a comprehensive review of supervised machine learning and nature-inspired computing techniques used in the analysis of human psychological disorders. Their analysis revealed that the application of supervised machine learning techniques in identifying psychological disorders achieves an accuracy of more than 90%.

In a published study [11] author Manogaran, Gunasekaran, et al. demonstrated the use of big data analytics and machine learning algorithms for identifying the changes in DNA sequencing. Big data analytics is commonly used in applications related to DNA study. The amalgamation of big data analytics with machine learning techniques is justified in the work by achieving 86.55% accuracy.

In [12] authors review the available literature on diagnosis of cancer and diabetic using five different insect-based optimization techniques viz. Ant Colony Optimization (ACO), Artificial Bee Colony (ABC), Glow-Worm Swarm Optimization (GSO), Firefly Algorithm (FA) and Ant Lion Optimization (ALO). The study highlights two things: first, most of the disease diagnostic work has been carried out using ACO, whereas GSO found to be least explored and second, high predictive accuracy achieved using the hybridization of ACO and neural network.

In [13] the author presented a detailed review of machine learning algorithms used in the prognosis of breast cancer. The commonly used machine algorithms are the Support vector machine, Decision tree, K-nearest neighbor and artificial neural network. The data used for the experiment drawn from Wisconsin Breast Cancer Database (WBCD) which is a benchmark dataset for breast cancer.

The survey of the recently published studies justifies the use of ML techniques not only for cancer prediction but also for other chronic diseases. After analyzing the results and architectures discussed in previously published studies the best-suited algorithms were chosen for this research so, that a reliable machine learning model can be developed to predict cervical cancer patients. Most of the recently published studies also lack an important evaluation factor of sensitivity which is important because UCI cervical cancer dataset is suffering from heavy class imbalance problem. So the work present in this paper using ML techniques for early diagnosis and prognosis of cervical cancer and evaluated by considering one of the evaluation metric as sensitivity along with accuracy and specificity.

3. Machine learning (ML) and its Approaches

Machine learning is technique to resolve the artificial intelligence problem. It consists of set of learning algorithms that are further classified into supervised and unsupervised learning. The supervised learning algorithms work with the labelled data whereas unsupervised algorithms work with unlabelled data. The most commonly used ML techniques are K- Nearest Neighbour (KNN), Support Vector Machine (SVM) and Random Forest (RF) which are briefly described in this section below.

3.1. K- Nearest Neighbour (KNN)

K Nearest Neighbour is a classification technique that classifies the data into k classes. It is a non-parametric as it doesn't give any model. The value of k depends on the training data. The instances are classified into one of the k classes based on the distance function. The common distance functions used for KNN include Euclidean distance, Manhatten distance, Minkowski distance or Hamming distance. The Euclidean, Manhatten and Minkowski distance functions are used when the values in the dataset are continuous in nature whereas if the data is categorical then the Hamming distance is used. The equations for all the mentioned functions are mentioned below.

$$Euclidean = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad \text{[16] (1)}$$

$$Manhattan = \sum_{i=1}^{k}(x_i - y_i) \quad \text{[17] (2)}$$

$$Minikowski = \sum_{i=1}^{k}\left(\left((x_i - y_i)\right)^q\right)^{1/q} \quad \text{[18] (3)}$$

Hamming standardizes the numerical variables between 0 and 1 by using normalization.

$$Hamming = \sum_{i=1}^{k}(x_i - y_i) \quad \text{[19] (4)}$$

3.2. Random Forest Tree (RFT)

Random Forest tree (RFT) is most widely used as a supervised machine learning technique in cervical cancer prediction. This technique is first introduced by Leo Breiman [20]. RFT used for solving both classification and regression problems. In this technique, multiple trees are generated and each tree gives "vote" for the target class. In the case of classification problem the forest makes the selection of trees that are having a maximum vote for the class and in case of regression average of a different tree is computed.

Table 1. Publication relevant to ML methods used in Cervical Cancer prediction

Publication	ML Technique	No. Of Patients	No. Of Features Used	Results
W. Yang, Xin Gou et al. [3]	Random Forest	858	32	Acc=97.6%
Parikh D. and Menon V. [4]	KNN	858	32	Acc=82.2%, F1-Score =94%
Wu and Zhou [5]	PCA with SVM	668	8	Acc=90%, Sen=100%, Spec=88%
A. Sarwar et al.[6]	Ensembles of dichotomous	917	20	Acc=77.8%
K. Hemlatha and K. Usha Rani [7]	Multi class Perceptron	917	4	Acc=92.0%
R. Vidya and G.M Nasira [9]	Random Forest tree with K-Means	100	61	Acc=96.77%
Hasan et al. [14]	Ensemble of decision Trees	858	32	Acc=96%
P. Bountris et al. [15]	Ensemble of weighted Random forest trees	203	21	Acc=93.03%

For predicting a continuous variable using Random Forests, the trees are grown depending on, a random vector, in such a manner that which is the tree predictor takes on numeric values. The values of response variable are numeric and it is assumed that the training sample is drawn independently from the distribution X of random vector Y. Equation (5) shows the mean square generalization error for a numeric predictor.

$$E_{x,y}(Y - h(X))^2 \quad [21] \ (5)$$

The Random Forest predictor is constructed by taking the mean over k of the trees$[h(x, \theta_k)]$. Random Forests tend to be accurate and effective in prediction due to the right kind of randomness.

3.3. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning technique that is introduced by Vapnik [22]. It works for both linear and non-linear dataset. The principle behind the SVM is to maximize predictive accuracy by minimizing the over fitting. It transforms the non-linear dataset into the linear dataset by using a higher dimension. It constructs various hyper planes for classifying the dataset.

The core of the SVM algorithm is in the minimization:

$$minC \sum_{i=1}^{m}[y^i cost_1(\theta^T x^i) + (1 - y^i) cost_0(\theta^T x^i)] + \frac{1}{2}\sum_{i=1}^{n}\theta^2 \quad [23] \ (6)$$

Where C refers to the error penalty function and are the cost functions when y=1 and y=0 respectively.

4. Methods and Materials

4.1. Dataset

The Cervical Cancer dataset used in the experiment is obtained from the UCI repository. The dataset is consisting of a clinical history of 858 patients which is described using 32 attributes and 4 labels (Schiller, Hinselmann, Biopsy, and Cytology) are described in Table 2. The description of these four target variables is as given below.

- Hinselmann: This test was developed by Mr. Hinselmann. He developed a tool that is used for visual inspection of the cervix at a magnified scale [24].
- Schiller: Schiller Test was originally introduced by Walter Schiller in 1993. This test results in non-cancerous and cancerous tissues by changing the colour to brown and yellow respectively. This test is done by applying the solution of iodine and potassium iodide on the surface of the cervix [25].
- Cytology: Cytology test is a cervical cancer screening test where a doctor takes fluid to examine the cells [26].
- Biopsy: Biopsy test is an invasive test that detects the abnormal area by taking out the sample of tissue [27].

Table 2. Data Description

No.	Data Type	Attribute
1	Integer	Age
2	Integer	Number of sexual partners
3	Integer	First sexual intercourse (age)
4	Integer	Num of pregnancies
5	boolean	Smokes
6	boolean	Smokes (years)

7	boolean	Smokes (pack/year)
8	boolean	Hormonal Contraceptives
9	integer	Hormonal Contraceptives(years)
10	boolean	IUD
11	integer	IUD (years)
12	boolean	STDs
13	integer	STDs (number)
14	boolean	STDs:condylomatosis
15	boolean	STDs:cervical condylomatosis
16	boolean	STDs:vaginal condylomatosis
17	boolean	STDs:vulvo-perineal condylomatosis
18	boolean	STDs:syphilis
19	boolean	STDs:pelvic inflammatory disease
20	boolean	STDs:genital herpes
21	boolean	STDs:molluscum contagiosum
22	boolean	STDs:AIDS
23	Boolean	STDs:HIV
24	Boolean	STDs:Hepatitis B
25	Boolean	STDs:HPV
26	Integer	STDs:Number of diagnosis
27	Integer	STDs: Time since first diagnosis
28	Integer	STDs: Time since last diagnosis
29	Boolean	Dx:Cancer
30	Boolean	Dx:CIN
31	Boolean	Dx:HPV
32	Boolean	Dx
33	Boolean	Hinselmann (target variable)
34	Boolean	Schiller (target variable)
35	Boolean	Cytology (target variable)
36	Boolean	Biopsy (target variable)

4.2. Proposed Architecture

The proposed architecture is depicted in figure 1 that describes the flow of the study. The data is consisting of 858 clinical records and 32 attributes. The given data is suffering from two major issues: firstly imbalanced data and secondly Missing values. To address this issue the first step taken is of data pre-processing. Thus before feeding the data to the predictive model missing value issue is resolved by eliminating the two features namely "STDs: Time since first diagnosis" and "STDs: Time since the last diagnosis" as it doesn't contain enough data. After eliminating these two columns, 190 instances that contains missing values ('?', Null) are also dropped. So, after cleaning the data, 668 records in the raw dataset are used for experiment.

The imbalanced data issue is resolved through three class balancing technique that includes RUS, ROS, and Smote. The balanced data obtained with these methods separately are shown in Figure 2-5. The cleaned data is then divided into train and test dataset in the ratio of 70-30%. After this step, the training dataset is fed to four predictive i.e. Random forest classifier, K-nearest neighbor, Support vector machine using linear kernel and support vector machine using a polynomial kernel. The performance of each model is analyzed on test data. The predicted results are then compared with actual results for evaluating the performance of the model.

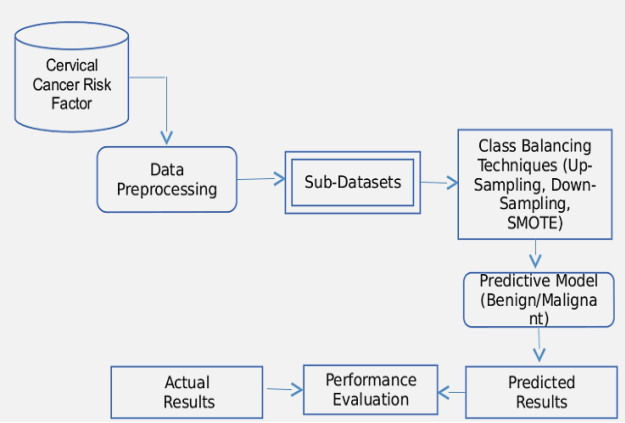

Figure 1. Proposed Architecture

4.3. Class Balancing Techniques

The given data is highly imbalanced for all the four targets. Thus before feeding the data to a predictive model for obtaining good results, there is a need to make the data balanced. Therefore three-class balancing techniques are applied namely Random under-sampling, Random Oversampling, and SMOTE. The balanced distribution of data after applying the class balancing techniques on all the four target variables is shown in Figure 2 to Figure 5. The first set of bars in figures 2-5 corresponds to original data distribution which indicates an imbalance of benign and malignant records. The second and third set of bars represents equal no. of malignant and benign instances. In the fourth set of the bar, malignant cases are synthetically upscale to benign cases.

4.3.1. Random Under Sampling (RUS)

In this technique the number of instances from the majority class is reduced in order to get the balanced data. The number of instances from the majority class is selected randomly which will be equals to the instances in minority class. The number of instances in minority class will remain same. This is also known as down sampling technique.

4.3.2. Random Over Sampling (ROS)

In this technique, the numbers of instances of the minority class are increased to the number of instances in the majority. This is done by duplicating random instances of the minority class. Thus all the features of the original

data set preserved [28] as no instance dropped off. This technique is also known as up-scaling.

4.3.3. Synthetic Minority Over-sampling Technique (SMOTE)

This technique also increases the number of instances of minority classes like ROS. The difference between the SMOTE and ROS is that in SMOTE the samples are increased synthetically by using the nearest neighbor [29] approach whereas in ROS the samples are increased simply by duplicating the samples available in the minority class.

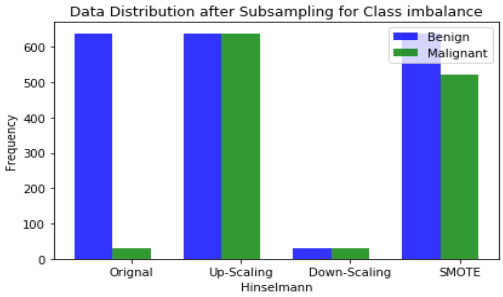

Figure 2. Balanced dataset for target variable Hinselmann

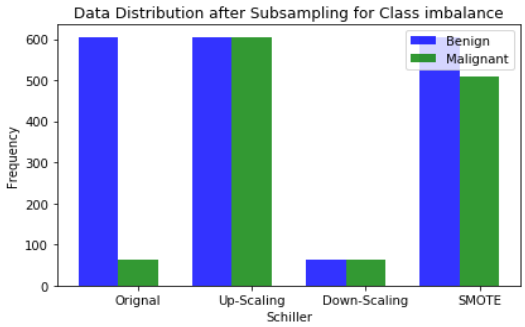

Figure 3. Balanced dataset for target variable Schiller

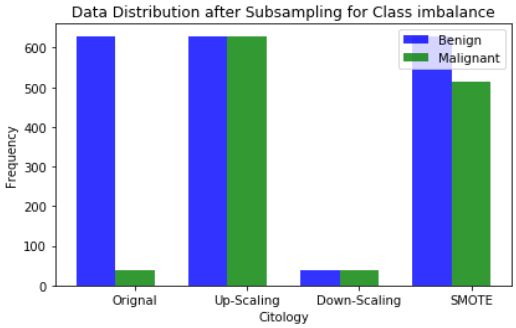

Figure 4. Balanced dataset for target variable Cytology

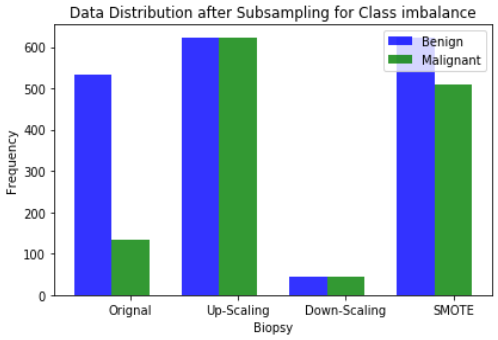

Figure 5. Balanced dataset for target variable Biopsy

4.4. Simulation Experimental Results and Analysis

The evaluation measures used for the model are accuracy, sensitivity, specificity, positive predicted accuracy and negative predicted accuracy. As the given dataset is suffering from the problem of imbalance, thus accuracy can't be taken as the only criterion for evaluation of the performance of the model. Thus the sensitivity and specificity will play a major role in diagnosis true positive and false-negative cases. The formula used for calculation accuracy, sensitivity, specificity, positive predicted accuracy, and negative predicted accuracy is given in table 3. In the listed formula TP refers to the true positive, which means the malignant samples are diagnosed as malignant whereas TN means true negative refers to the benign samples are diagnosed as benign. On the other hand, FP refers to False-negative which is equal to the number of samples diagnosed as malignant but is benign. Contrary to FP, FN is the number of samples that are malignant but stated as benign.

- TARGET VARIABLE: HINSELMANN

Under Hinselmann Test, the values are shown in table 4(a) for four different classifiers under three-class balancing techniques. The actual numbers of benign and malignant cases are 638 and 30 respectively. The data is then balanced using ROS, RUS and Smote and the classifiers are applied to obtained data. Among all the classifiers, the Random Forest algorithm shows better results for accuracy, sensitivity, specificity, PPA and NPA in table 4(a).

- TARGET VARIABLE: SCHILLER

Under SCHILLER Test, the actual numbers of benign and malignant cases are 605 and 63 respectively. The data is then balanced using all the three-class balancing techniques. The results for all the evaluation parameters are shown in Table 4(b).

- TARGET VARIABLE: CYTOLOGY

Under Cytology's Test, the given numbers of benign and malignant cases are 629 and 39 respectively. Among all the classifiers, the Random Forest algorithm with the SMOTE balancing technique shows better results in Table 4(c).

Table 3. Evaluation metrics summary

Summary Statistics	Equation
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$
Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{FP+TN}$
Positive predicted accuracy	$\frac{TP}{TP+FP}$
Negative predicted accuracy	$\frac{TN}{TN+FN}$

- TARGET VARIABLE: BIOPSY

Under Biopsy's Test, the values are shown in the table for four different classifiers under-three class balancing techniques. The actual numbers of benign and malignant cases are 534 and 134 respectively. The data is then balanced using ROS, RUS and Smote and the classifiers are applied to obtained data. The results are shown in Table 4(d).

4.5. Comparative Analysis

The comparative analysis of accuracy for all the four machine learning classifiers namely Random forest, KNN, SVM-poly and SVM-Linear for all the four target variables is shown in figure 6. It depicts the Random Forest Tree with the SMOTE class balancing technique is giving the highest accuracy for cytology, Schiller, Hinselmann, and biopsy. The KNN classifier is the least performing for Cytology, Schiller, and biopsy.

Similarly, figure 7 represents the analysis of sensitivity for all the four classifiers. The random forest classifier results in the highest sensitivity for all the four target variables and the KNN classifier results in low sensitivity. Thus for both accuracy and sensitivity, the results have shown a random forest algorithm with Smote balancing technique overpowered all the other classifiers.

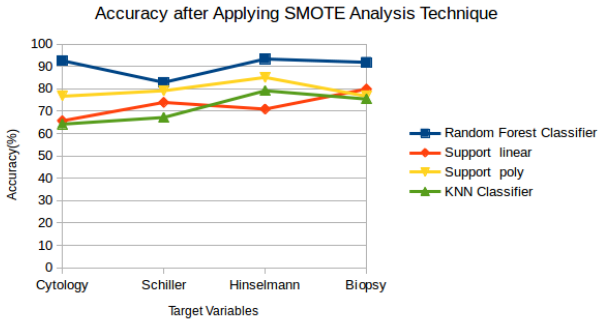

Figure 6. Comparative analysis of Accuracy

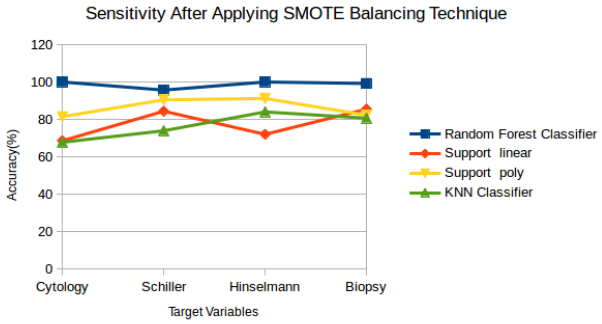

Figure 7. Comparative analysis for Sensitivity

5. Conclusion

In this paper, explanations of different ML classifiers and class balancing techniques are provided. The most commonly used ML techniques namely K-NN, SVM, and random forest tree are chosen for carrying out the experimental work. The data used in the experiment is the cervical cancer dataset which is available publicly on the UCI repository. The data obtained from the repository was imbalanced. Thus three-class balancing techniques viz. ROS, RUS, and SMOTE are used to make the data balanced before feeding the data to the proposed model. The findings of the study are predictive cervical cancer model. The results have shown the Random forest algorithm performs better with SMOTE for four target variables Schiller, Biopsy, Hinselmann, and Cytology respectively whereas the KNN is the least performer. As future work, we will use dimensionality reduction technique with the classifiers to see their influence. The second potential work is a multi-class classification with the four targets.

Table 4(a) Result for target Variable Hinselmann

Class Balancing Technique	Classifier	RF	SVM-Linear	SVM-Ploy	KNN
ROS Benign=638, Malignant= 638	Accuracy	98.44	65.63	82.43	92.58
	Sensitivity	96.7	57.86	73.56	84.3
	Specificity	100	72.6	90.38	100
	PPA	100	65.43	87.26	100
	NPA	97.13	65.78	79.23	87.67
RUS Benign=30, Malignant= 30	Accuracy	41.67	66.67	58.34	91.67
	Sensitivity	42.86	57.15	57.15	100
	Specificity	40	80	60	80
	PPA	50	80	66.67	87.5
	NPA	33.34	57.15	50	100
SMOTE Benign=638, Malignant= 523	Accuracy	93.29	70.9	85.08	79.11
	Sensitivity	100	72	91.2	84
	Specificity	0	55.56	0	11.12
	PPA	93.29	95.75	92.69	92.93
	NPA	0	12.5	0	4.77

Table 4(b) Result for target Variable Schiller

Class Balancing Technique	Classifier	RF	SVM-Linear	SVM-Ploy	KNN
ROS Benign=638, Malignant= 638	Accuracy	98.77	65.29	82.65	85.54
	Sensitivity	97.46	83.06	83.06	70.34
	Specificity	100	48.39	82.26	100
	PPA	100	60.5	81.67	100
	NPA	97.64	75	83.61	77.99
RUS Benign=30, Malignant= 30	Accuracy	53.85	38.47	38.47	53.85
	Sensitivity	60	46.67	40	46.67
	Specificity	45.46	27.28	36.37	63.64
	PPA	60	46.67	46.16	63.64
	NPA	45.46	27.28	30.77	46.67
SMOTE Benign=638, Malignant= 523	Accuracy	82.84	73.89	79.11	67.17
	Sensitivity	95.66	84.35	90.44	73.92
	Specificity	5.27	10.53	10.53	26.32
	PPA	85.94	85.09	85.96	85.86
	NPA	16.67	10	15.39	14.29

Table 4(c) Result for target Variable Cytology

Class Balancing Technique	Classifier	RF	SVM-Linear	SVM-Ploy	KNN
ROS Benign=638, Malignant= 638	Accuracy	98.02	58.34	76.2	88.89
	Sensitivity	95.91	93.45	68.86	77.05
	Specificity	100	25.39	83.08	100
	PPA	100	54.03	79.25	100
	NPA	96.3	80.49	73.98	82.28
RUS Benign=30, Malignant= 30	Accuracy	31.25	43.75	37.5	18.75
	Sensitivity	50	50	0	75
	Specificity	25	41.67	50	0
	PPA	18.19	22.23	0	20
	NPA	60	71.43	60	0
SMOTE Benign=638, Malignant= 523	Accuracy	92.54	65.68	76.87	64.18
	Sensitivity	100	68.55	81.46	67.75
	Specificity	0	30	20	20
	PPA	92.54	92.4	92.67	91.31
	NPA	0	7.15	8	4.77

Table 4(d) Result for target Variable Biopsy

Class Balancing Technique	Classifier	RF	SVM-Linear	SVM-Ploy	KNN
ROS Benign=638, Malignant= 638	Accuracy	98	72	84.4	88.8
	Sensitivity	96.1	87.5	83.6	78.13
	Specificity	100	55.74	85.25	100
	PPA	100	67.47	85.6	100
	NPA	96.07	80.96	83.2	81.34
RUS Benign=30, Malignant= 30	Accuracy	61.12	66.67	66.67	50
	Sensitivity	63.64	72.73	63.64	45.46
	Specificity	57.15	57.15	71.43	57.15
	PPA	70	72.73	77.78	62.5
	NPA	50	57.15	55.56	40
SMOTE Benign=638, Malignant= 523	Accuracy	91.8	79.86	76.87	75.38
	Sensitivity	99.2	85.49	82.26	80.65
	Specificity	0	10	10	10
	PPA	92.49	92.18	91.82	91.75
	NPA	0	5.27	4.17	4

References

[1] https://www.who.int/news-room/fact-sheets/detail/cancer

[2] Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. Transfer Learning with Partial Observability Applied to Cervical Cancer Screening. In Proceedings of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA). Springer International Publishing; 2017. p. 243-250

[3] Yang, Wenying & Gou, Xin & Xu, Tongqing & Yi, Xiping & Jiang, Maohong. Cervical Cancer Risk Prediction Model and Analysis of Risk Factors based on Machine Learning.In Proceedings of 11th International Conference on Bioinformatics and Biomedical Technology;May 29 – 31; Sweden. ACM; 2019. p. 50-54.

[4] Parikh D., Menon V. (2019). Machine Learning Applied to Cervical Cancer Data. International Journal of Mathematical Sciences and Computing. 2019; 5(1): 53-64

[5] Wu, W. and Zhou, H. Data-driven diagnosis of cervical cancer with support vector machine-based approaches, IEEE Access; 2017; 5: 25189–25195.

[6] Sarwar A., Ali M., and Sharma V. Performance Evaluation of Machine Learning Techniques for Screening of Cervical Cancer. In 2nd International Conference on Computing for

Sustainable Global Development (INDIACom); Mar 11-13 BVICAM New Delhi INDIA. IEEE; 2015. p. 2297-2303.
[7] Hemalatha K. and Rani U.An Optimal Neural Network Classifier for Cervical Pap smear Data. In 7th International Advance Computing Conference(IACC); Jan 5-7; Hyderabad, India. IEEE; 2017. p. 110-114.
[8] Latha D.S., Lakshmi P. and Fathima S. Staging Prediction in Cervical Cancer Patients-A Machine Learning Approach. International Journal of Innovative research and Practices. 2014; 2(2):14-23.
[9] Vidya R. and Nasira G. Prediction of Cervical Cancer using Hybrid Induction Technique: A Solution for Human Hereditary Disease Patterns. Indian Journal of Science & Technology. 2016; 9(30):1-10.
[10] Kaur, Prableen, and Manik Sharma. Diagnosis of Human Psychological Disorders using Supervised Learning and Nature-Inspired Computing Techniques: A Meta-Analysis.Journal of medical systems. 2019; 7(43): 204.
[11] Manogaran, Gunasekaran, et al. Machine learning-based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering. Wireless personal communications. 2018; 3(102): 2099-2116.
[12] Gautam, Ritu, Prableen Kaur, and Manik Sharma. A comprehensive review on nature inspired computing algorithms for the diagnosis of chronic disorders in human beings.Progress in Artificial Intelligence. 2019; 1-24.
[13] Yue, Wenbin, et al. Machine learning with applications in breast cancer diagnosis and prognosis. Designs. 2018; 2(2): 13.
[14] Hasan R., Gholamhosseini H. and Sarkar N. I. A new ensemble classifier for multivariate medical data. In 27th International Telecommunication Networks and Applications Conference (ITNAC); 22-24 Nov 2017; Melbourne, VIC, Australia, IEEE; 2018. p. 1-6.
[15] Bountris P., Haritou M. , Pouliakis A. ,Karakitsos P. and Koutsouris D. A decision support system based on an ensemble of random forests for improving the management of women with abnormal findings at cervical cancer screening. In 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 25-29 Aug 2015; Milan, Italy. IEEE; 2015. p. 8151-8156.
[16] Paul E. Black, "Euclidean distance", in Dictionary of Algorithms and Data Structures [online], Paul E. Black, ed. 17 December 2004. (accessed TODAY) Available from: https://www.nist.gov/dads/HTML/euclidndstnc.html
[17] Paul E. Black, "Manhattan distance", in Dictionary of Algorithms and Data Structures [online], Paul E. Black, ed. 11 February 2019. (accessed TODAY) Available from: https://www.nist.gov/dads/HTML/manhattanDistance.html
[18] E.F. Krause. Taxicab Geometry: An Adventure in Non-Euclidean Geometry. California: Addison-Wesley; 1987.
[19] Paul E. Black, "Hamming distance", in Dictionary of Algorithms and Data Structures [online], Paul E. Black, ed. 31 May 2006. (accessed TODAY) Available from: https://www.nist.gov/dads/HTML/HammingDistance.html
[20] E.F. Krause. Taxicab Geometry: An Adventure in Non-Euclidean Geometry. California: Addison-Wesley; 1987.
[21] Breiman, L. Random Forests. Machine Learning. 2001; 45: 5–32.
[22] Vapnik V. The Nature of Statistical Learning Theory. New York: Springer; 1995.
[23] Press, William H. , Teukolsky, Saul A. , Vetterling, William T., Flannery, Brian P. Numerical Recipes: The Art of Scientific Computing. 3rd edition. New York: Cambridge University Press. 2007 ISBN 978-0-521-88068-8.
[24] Fusco E., Padula F., Mancini E., Cavalieri A., Grubisic G. History of colposcopy: a brief biography of Hinselmann. Journal of Prenatal Medicine. 2008; 2(2): 19-23.
[25] Singer A., Monaghan J. M., Quek S.C. Lower genital tract precancer colposcopy, pathology and treatment. 2nd edition.Wiley: Blackwell Science; 2008
[26] Bentz J.S. Liquid-based cytology for cervical cancer screening. Expert Review of Molecular Diagnostics. 2005; 5(6): 857-871.
[27] Kaveri S. B., Khandelwal S. Role of Pap smear N cervical biopsy in unhealthy cervix. Journal of Scientific and Innovative Research. 2015; 4(1):4-9.
[28] R. Prati C., Gustavo E. A., BatistaP. A. and MonardM.C. Data mining with imbalanced class distributions: concepts and methods, In 4thIndian International Conference on Artificial Intelligence; 2009. p. 359-376.
[29] Chawla N.V., Bowyer K.W.,Hall L. O., Kegelmeyer W. P. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002; 16:321-357.