International Journal of Innovative Research in Engineering
Volume 3, Issue 3 (May-June 2022), PP: 186-190.
www.theijire.com
ISSN No: 2582-8746

Cervical Cancer Prediction using Outlier deduction and Over sampling methods

K.Gowri[1], M.Saranya[2]
[1]*Department of Computer Science, Sri Kaliswari College (Autonomous), Tamilnadu, India.*
[2]*Assistant professor, Department of Computer Science, Sri Kaliswari College (Autonomous), Tamilnadu, India.*

How to cite this paper:
K.Gowri[1], M.Saranya[2], "Cervical Cancer Prediction using Outlier deduction and Over sampling methods", IJIRE-V3I03-186-190.

Abstract: *Cervical cancer is one of the disease considered to be fourth among the most common types of cancer in women around the world. The early deduction of cervical cancer helps to raise number of recovery patients and reduce death rates. This research work aims to use machine learning algorithms to predicting cervical cancer with high accuracy using an outlier deduction and over sampling method. Analyse the cervical cancer data available the dataset from UCI repository. In this research, first step removes outliers by using outlier detection method such as density-based spatial clustering of applications with noise (DBSCAN) and by increasing the number of cases in the dataset in a balanced way through the synthetic minority over-sampling technique (SMOTE) and SMOTE with Tomek link (SMOTETOmek). Finally, it employs random forest (RF) as a classifier to check the accuracy. Thus, the prediction model Have a two scenarios: (1) DBSCAN + SMOTETomek + RF, (2) DBSCAN + SMOTE+ RF. I found that combination of DBSCAN with SMOTE provided better performance than DBSCAN with SMOTETomek. Also observed that RF performed the best among several popular machine learning classifiers. Furthermore, the proposed research work showed better accuracy than previously proposed methods for forecasting cervical cancer.*

Key Word: *Machine learning, cervical cancer, Outlier deduction, Oversampling, SMOTE, SMOTETomek, DBSCAN.*

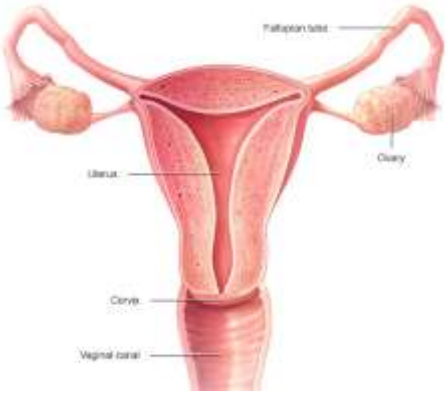
I. INTRODUCTION

Cervical cancer is one type of a cancer that occurs in the Cells of cervix — the lower part of the uterus that connects to the vagina. The Various strains of the human papilloma virus (HPV), a sexually transmitted infection, play a role in causing most cervical cancer. When exposed to HPV, the body's immune system typically prevents the virus doing any harm. In a small percentage of people, however, the virus survives for years, contributing to the process that causes some cervical cells to become cancer cells. Can reduce your risk of developing cervical cancer by having screening tests and receiving a vaccine that protects against HPV infection.

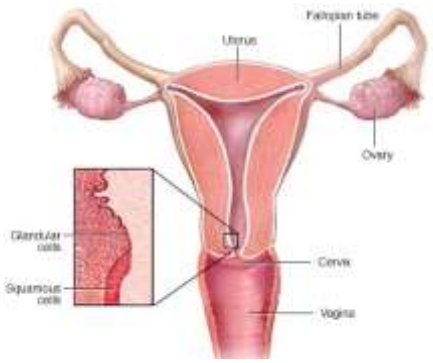
1.1 Symptoms

Early-stage cervical cancer generally produces no signs or symptoms. Signs and symptoms of more-advanced cervical cancer include:

- Vaginal bleeding after intercourse, between periods or after menopause
- Watery, bloody vaginal discharge that may be heavy and have a foul odour
- Pelvic pain or pain during sexually intercourse



Female reproductive system



where the cervical cancer begins

1.2 Causes

Cervical cancer begins when healthy cells of the cervix develop and changes (mutations) in their DNA. A cell's DNA contains the instructions that tell a cell what to do. Healthy cells grow and multiply at a set rate, eventually dying at a set time. The mutations tell the cells to grow and multiply out of control, and they don't die. The accumulating abnormal cells form a mass (tumor). Cancer cells invade nearby tissues and can break off from a tumor to spread (metastasize) elsewhere in the body. It isn't clear what causes cervical cancer, but it's certain that HPV plays a role. HPV is very common, and most people with the virus never develop cancer. This means other factors such as your environment or your lifestyle choices also determine whether you'll develop cervical cancer.

1.3 Types of cervical cancer

The type of cervical cancer that you have helps determine your prognosis and treatment. The main types of cervical cancer are:

- **Squamous cell carcinoma**: This type of cervical cancer begins in the thin, flat cells (squamous cells) lining the outer part of the cervix, which projects into the vagina. Most of the cervical cancers are squamous cell carcinomas.
- **Adenocarcinoma** : This type of cervical cancer begins in the column-shaped glandular cells that line the cervical canal. Sometimes, both types of cells are involved in cancer. Very rarely, cancer occurs in other cells in the part of cervix.

1.4 Risk factors

Risk factors for cervical cancer include:

- **Other sexually transmitted infections (STIs).** Having other STIs — such as chlamydia, gonorrhea, syphilis and HIV/AIDS — increases your risk of HPV.
- **A weakened immune system.** May be more likely to develop cervical cancer if immune system is weakened by another health condition and you have HPV.
- **Smoking.** Smoking is associated with squamous cell cervical cancer.
- **Exposure to miscarriage prevention drug.** If mother took a drug called diethylstilbestrol (DES) while pregnant in the 1950s, may have an increased risk of a certain type of cervical cancer called clear cell adenocarcinoma.

1.5 Prevention

To reduce your risk of cervical cancer:

- **Ask doctor about the HPV vaccine.** Receiving a vaccination to prevent HPV infection may reduce your risk of cervical cancer and other HPV-related cancers. Ask doctor whether an HPV vaccine is appropriate
- **Have routine Pap tests.** Pap tests can detect precancerous conditions of the cervix, so they can be monitored or treated in order to prevent cervical cancer. Most medical organizations suggest beginning routine Pap tests at age 21 and repeating them every few years.
- **Don't smoke.** If don't smoke, don't start. If do smoke, talk to doctor about strategies to help quit.

II. RELATED WORK

1. Data Driven Cervical Cancer Prediction Model with Outlier Detection and OverSampling Methods (2020) - Muhammad Fazal Ijaz , Muhammad Attique and Youngdoo Son

The work removes outliers by using outlier detection methods such as density-based spatial clustering of applications with noise (DBSCAN) and isolation forest (iForest) and by increasing the number of result in the dataset in balanced way through synthetic minority over-sampling technique (SMOTE) and SMOTE with Tomek-link (SMOTETOmek). Finally, the random forest (RF) as a classifier. Thus, model have a four scenarios: (1) DBSCAN + SMOTETomek + RF, (2) DBSCAN + SMOTE+ RF, (3) iForest + SMOTETomek + RF, and (4) iForest + SMOTE + RF. It used Chisquare as feature extraction technique. Hence, combining iForest with SMOTE and SMOTETomek can produce better results than combining DBSCAN with SMOTE and SMOTETomek

2. Prediction of Cervical Cancer using Hybrid Induction Technique: A Solution for Human Hereditary Disease Patterns (2016) - R. Vidya,G. M. Nasira

The Regression tree algorithm methodology was used for prediction. The CART binary tree gives two results, either normal cervix or cancer cervix. RFT validated the optimal precision , a new logic "combinations of two algorithms" was applied . It is also an ensampling supervised machine learning algorithm. The process of whitening used for a pre-process in k-means clustering, to get the best prediction result. The result showed 83.87% accuracy with CART TREE output. Random Forest Tree (RFT) is used for improve the prediction accuracy. With MATLAB Coding we achieved 93.54% of prediction accuracy. The K-Means algorithm is an efficient one for processing huge datasets and hence a high accuracy of 96.77% is achieved with the model of TFT – KMEAN LEARNING TREE output. By the process, accuracy of RFT with K-means for cervical cancer prediction is enhanced to 96.77% while comparing these three.

3. Predicting Cervical Cancer using Machine Learning Methods (2020)-Riham Alsmariy, Graham Healy, Hoda Abdelhafez

The machine learning algorithms to find a model capable of diagnosing cervical cancer. The work through a voting method that combines three classifiers: Decision tree, logistic regression and random forest. The synthetic minority oversampling technique (SMOTE) was used to solve for the problem of imbalance dataset and, together with the principal component analysis (PCA) technique, to reduce dimensions that do not affect model precision. Then, stratified 10-fold cross-validation technique was used to

prevent the over fitting problem. In the SMOTE-voting model, accuracy, sensitivity and PPA ratios improved by 0.93 % to 5.13 %, 39.26 % to 46.97 % and 2 % to 29 %, respectively for all target variables. Moreover, using PCA technology reduced computational processing time and increasing model efficiency. It raise the accuracy, sensitivity, and ROC_AUC of predictive models to high rates as in the Schiller target variable, they reached to 98.49%, 98.60%, and 99.80%, respectively.

4. Cervical Cancer Prediction through Different Screening techniques using Data Mining (2019) – Talha Mahboob Alam, Muhammad Milhan Afzal Khan, Muhammad Atif Iqbal, Abdul Wahab, Mubbashar Mushtaq

Cervical cancer prediction through different screening Techniques using data mining techniques like Boosted decision tree, decision forest and decision jungle algorithms performed well and evaluation has done on the basis of AUROC (Area under Receiver operating characteristic) curve, accuracy, specificity and sensitivity. 10-fold cross validation method , it was utilized to authenticate the results and Boosted decision tree has given the best effects. Boosted DT provided very high prediction with 0.978. The prediction ability of the boosted decision tree measured by the AUROC curve value which outperformed decision forest and decision jungle. The low AUROC curve value used for the decision forest and decision jungle methods disqualified them as best predictive classifiers. Believe that with the growing collection of cervical cancer patient's data and the rapidly advancing methods for analyzing this data, we will begin to be able to identify best screening method for cervical cancer patients that will be informative for patient care. In future, this study can be used as a prototype to develop a healthcare system for cervical cancer patients.

5. A Contemporary Machine Learning Method for Accurate Prediction of Cervical Cancer (2021) – Jesse Jeremiah Tanimu, Mohamed Hamada, Mohammed Hassan, Saratu Yusuf Ilu

Decision tree (DT) classification algorithm and shows the advantage of feature selection approach in the prediction of cancer using recursive feature elimination technique for dimensionality reduction for improving the accuracy, sensitivity, and specificity of the work. Therefore, a combination of under and oversampling Methods called SMOTETomek that was employed. A comparative analysis of the proposed model that performed effective feature selection and class imbalance based by classifier's accuracy, sensitivity, and specificity. The Desicion Tree with selected features and SMOTETomek has better results with an Precision of 98%, sensitivity of 100%, and specificity of 97%. Decision Tree classifier is shown to a excellent performance in handling classification assign when the features are reduced, and the problem of imbalance class is addressed and solved.

III. PROPOSED METHODOLOGY

The Cervical cancer dataset is published in the repository of UCI collected at Hospital Universitario de Caracas , Venezuela. The dataset have a 858 instances with 36 features.

Table 1. Dataset features, number of entries, and missing values.

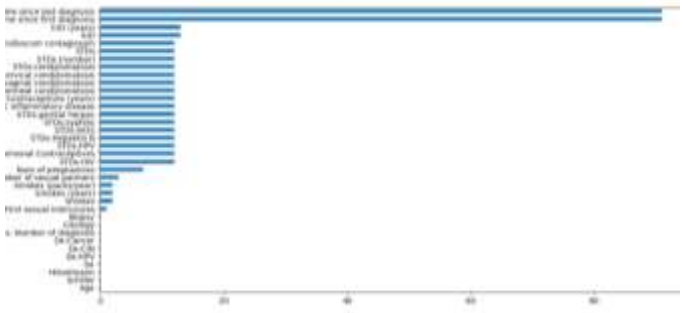
Number	Attribute Name	Type	Missing Values
1	Age	Int	0
2	Number of sexual partners	Int	26
3	First sexual intercourse (age)	Int	7
4	Num of pregnancies	Int	56
5	Smokes	bool	13
6	Smokes (years)	bool	13
7	(Smokes (packs/year)	bool	13
8	Hormonal Contraceptives	bool	108
9	Hormonal Contraceptives (years)	Int	108
10	Intrauterine Device (IUD)	bool	117
11	IUD (years)	Int	117
12	Sexually Transmitted Disease (STD)	bool	105
13	STDs (number)	Int	105
14	STDs: condylomatosis	bool	105
15	STDs: cervical condylomatosis	bool	105
16	STDs: vaginal condylomatosis	bool	105
17	STDs: vulvo-perineal condylomatosis	bool	105
18	STDs: syphilis	bool	105
19	STDs: pelvic inflammatory disease	bool	105
20	STDs: genital herpes	bool	105
21	STDs: molluscum contagiosum	bool	105
22	STDs: AIDS	bool	105
23	STDs: HIV	bool	105
24	STDs: Hepatitis B	bool	105
25	STDs: HPV	bool	105
26	STDs: Number of diagnosis	Int	0
27	STDs: Time since first diagnosis	Int	787
28	STDs: Time since last diagnosis	Int	787
29	Dx: Cancer	bool	0
30	Dx: Cervical Intraepithelial Neoplasia (CIN)	bool	0
31	Dx: Human Papillomavirus (HPV)	bool	0
32	Diagnosis: Dx	bool	0
33	Hinselmann: target variable	bool	
34	Schiller: target variable	bool	
35	Cytology: target variable	bool	
36	Biopsy: target variable	bool	

In this research, used the Python programming language and scikit-learn, pandas, numpy libraries used for machine learning models. scikit-learn library used for an outlier deduction methods. For oversampling, we have used the imbalanced-learn Python

library

Step 1: Import the Data set

Step 2: Checking the percentage of missing values and remove which feature have a high ratio of missing values



Step 3: Fill median value instead of missing values, then extract the features.

Feature extraction:

PCA(Principle Component Analysis) is a famous and common feature extraction method PCA finds the edge n vectors of a covariance matrix with the high edge n vectors. Use a data into a new subspace of equal or less dimensions

Step 4: After the feature extraction plot the result and compare to raw data.

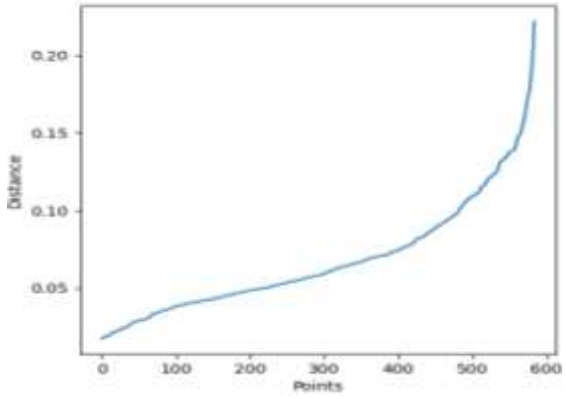
Step 5: Split the dataset into 70% train and 30%test data

Step 6: Use Scalar and normalization functions for standardized the data, next deduct the outliers

DBSCAN

To implement DBSCAN-based outlier detection, the optimum value of MinPts and eps must first to be accepted. If value of the eps is too low, it will generate more clusters and normal data may be counted as An outliers. On the other hand, if it is too large, produce the fewer clusters, and true outliers could be categorized as normal data. Specified the Minimum point value to be 5. Next, have to define the optimal number of eps(0.1)

First, we measure each point's average distance from its nearest neighbors. The value k represents MinPts and is outlined by the user. The goal is to decide the "knee" used to estimate the parameter collection of eps. A "knee" is the point at which a sharp shift occurs along the k-distance curve.



eps value finding using knee

The Figure, displays the k-dist graph sorted for the cervical cancer data set and the optimal value of eps. The "knee" shows up at the distance in the cervical cancer dataset. Lastly, the outlier data are excluded, and standard data are used for further analysis.

Step 7: now, the dataset have chance to change imbalance, it less accuracy of a model. So, we change the dataset imbalance to balance. Apply the Over sampling techniques

SMOTE and SMOTETomek.

Over-sampling methods used to increase the number of cases in a balanced way. Just applied SMOTE or SMOTETomek methods to balance the datasets. SMOTE it oversamples the minority class to randomly generate instances and increase the minority class instances, and Tomek under-samples a class to remove noise while maintaining balanced distributions.

Step 8: Random Forest for predict and check the accuracy

Random Forest (RF)

Classification is a form of data analysis that extracts models describing data classes. A classification model, predicts categorical labels (classes). Numeric prediction models continuous-valued functions. Classification and numeric prediction are the algorithm used to correct the accuracy values. Random forests is a ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression). It creates many classification trees and a bootstrap sample Methods is used to train the each tree from the set of training data. Train the model using Random Forest Classification on the Training Set. Predicting the Test set results ,Confusion Matrix for checking Accuracy.

IV.RESULT

Cervical cancer if found at an early stage will help save lives of thousands of women. This research helps the real world patients and doctors to gather as much information as they can. The Two main parts implemented in this System were Outlier deduction (DBSCAN) and Over sampling (SMOTE and SMOTETomek). In this model, the first one is deduct the outliers, remove the noise points and the second one is used for balance the data. RF use to check accuracy of the model .The confusion matrix is preconceived. The model can help users find the risk of cervical cancer at an early stage. Accuracies achieved by Biopsy for DBSCAN + SMOTETomek + RF, DBSCAN + SMOTE+ RF, were 98.008%, 99.007%.

V. CONCLUSION AND FUTURE WORK

In this Research, as indicated by the World Health Organization (WHO), about 80% cases of cervical cancer are noted in developing nations. A cure ratio means the ratio of female cases that are healed from the disease. It can be boosted by classifying risk factors of the cervical cancer. This research proposed a model that used PCA as feature extraction technique. Extracted 34 features and used them to prediction. The current work proposed model by joining DBSCAN for outlier detection, with SMOTE and SMOTETomek for class balancing and RF as a classifier. The model can help users find the risk of cervical cancer at an early stage. Accuracies achieved by Biopsy for DBSCAN + SMOTETomek + RF and DBSCAN +SMOTE+ RF, were 98.008%, 99.007%.

In future work, try to employ more diverse techniques for outlier detection and over-sampling methods. And also apply each combination of a model to improve its diagnosis performance. Later the proposed method can be applied to other cervical cancer datasets. Results on these may provide additional intuitions for early diagnosis of the cervical cancer. This prediction also have a limitation, as only one dataset is employed. Since I only focus on cervical cancer in this research. In future research, the proposed model can be applied to diverse cancer datasets (such as breast, liver, lung, thyroid, and kidney) to increase the clarity and quality of result. Next limitation is that our algorithm (which is a combination of outlier technique and data balancing with RF) becomes slower and needs more memory to run, but as got a better accuracy, it serves our Cause.

References

[1] Riham Alsmariy, Graham Healy, Hoda Abdelhafez," Predicting Cervical Cancer using Machine Learning Methods", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 7, 2020

[2 Muhammed Fahri Unlersen1, Kadir Sabanci2, Muciz Özcan1, "Determining Cervical Cancer Possibility by Using Machine Learning Methods", International Journal of Latest Research in Engineering and Technology (IJLRET) ISSN: 2454-5031, Volume 03 - Issue 12 , December 2017

[3] A.Priyanga, S.Prakasam, Ph.D, "Effectiveness of Data Mining - based Cancer Prediction System (DMBCPS)", International Journal of Computer Applications (0975 – 8887) Volume 83 – No 10, December 2013

[4] Y. M. S. Al-Wesabi, Avishek Choudhury, Daehan Won, "Classification of Cervical Cancer Dataset", Proceedings of the 2018 IISE Annual Conference K. Barker, D. Berry, C. Rainwater, eds.

[5] Arif-Ul-Islam, Shamim H. Ripon, Nuruddin Qaisar Bhuiyan," Cervical cancer risk factors: classification and mining associations", APTIKOM Journal on Computer Science and Information Technologies, Vol. 4, No. 1, 2019, pp. 8~18 ISSN: 2528-2417,DOI: 10.11591/APTIKOM.J.CSIT.131

[6] R.Vidya,G.M.Nasira," A Pioneering Cervical Cancer Prediction Prototype in Medical Data Mining using Clustering Pattern", International Journal of Data Mining Techniques and Applications, Volume: 04 Issue: 02 December 2015 Page No.63-66, ISSN: 2278-2419

[7] Zeynep CEYLAN1, Ebru PEKEL, "Comparison of Multi-Label Classification Methods for Prediagnosis of Cervical Cancer", International Journal of Intelligent Systems and Applications in Engineering Advanced Technology and Science ISSN:2147-67992147.