2021 International Conference on Electronics, Communications and Information Technology (ICECIT), 14–16 September 2021, Khulna, Bangladesh.

ML Classifier Comparative Performance Analysis of Prediction on Cervical Cancer

Md. Ashfaqul Haque
Dept. of Computer Science and Engineering
Brac University
Mohakhali, Dhaka – 1212, Bangladesh
md.ashfaqul.haque@g.bracu.ac.bd

Israt Jahan Dristy
Dept. of Computer Science and Engineering
Brac University
Mohakhali, Dhaka – 1212, Bangladesh
israt.jahan.dristy@g.bracu.ac.bd

Shihab Sharar
Dept. of Computer Science and Engineering
Brac University
Mohakhali, Dhaka – 1212, Bangladesh
shihab.sharar@g.bracu.ac.bd

Annajiat Alim Rasel
Dept. of Computer Science and Engineering
Brac University
Mohakhali, Dhaka – 1212, Bangladesh
annajiat@gmail.com

***Abstract*— Identification of cancer at an early stage in treating patients has been worthwhile. Cervical cancer occurs by alteration of cells connecting the uterus and vulva to anomalous changed state which can culminate in infecting deeper tissue. By making an early prediction on the result of the biopsy test, it may come to the verge of a cure. For making prognostication on biopsy test we used a dataset provided by the hospital Universitario de Caracas. Analytical and statistical survey on the dataset has led us to a conjecture of making predictions on biopsy using the following six classifiers which are LR, random forest, NB, SVM, SGD, and kNN, and finally done by comparing performance analysis. Among them, multinomial NB has stood out as the best performer with an accuracy of 97.76 percent.**

Keywords*— comparative analysis, cervical cancer, classifiers, machine learning, medical data, random forest, SVM, kNN, SGD, NB.

I. Introduction

Cervical Cancer is the deadliest disease and the fourth most common among women in the world. This mainly develops in a woman's cervix. Human papillomavirus infection is the most common cause of cervical cancer (HPV). In the early stage, there are no symptoms at all but later may include vaginal bleeding, pelvic pain, or pain during intimate. This cancer can be spread to other cells also. If this cancer can be found early, it can be treatable and curable. The percent of recovery is high of early detection, screening, and careful precautions. For diagnosing the correct stage of infection some papers proposed a prediction model. A biopsy in the cervix is the method to remove tissue from the cervix to test for abnormal or cervix cancer. It's always done after colposcopy. This presents a procedure to classify the biopsy result using ML classifier algorithms which are Logistic Regression (LR), Random Forest, Naïve Bayes (NB), Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), k-Nearest Neighbor (kNN). After doing the classifier, we compared the classifiers. Here we are going to have a look at some previous work on cervical cancer. Here on the 1st work Support vector machine has been used on Herlev pap-smear image Dataset. For segmentation active contour models have been used here This paper comes with a 95% accuracy [1]. Then another work is based on the behavior for predicting this cancer. This uses the Decision tree, Random Tree, and XGBoost for prediction using social behavior. This comes with an accuracy of 93.33% [2]. Now here a novel ensembled approach which used a data correction methodology has been proposed to increase the accuracy of prediction [3]. The next one works by considering risk factors of cervical cancer dataset comparing classifier techniques i.e. NB, C4.5 Decision Tree, kNN, Sequential Minimal Optimization, Random Forest Decision Tree, Multilayer Perceptron Neural Network, and Simple Logistic Regression [4].

II. Dataset

A. Dataset Description

The dataset we are using here is for cervical cancer. So, the features information is focused on detecting this which will be more viable. Dataset was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela [5]. Demographic Information, habits, and historic medical records of 858 patients are included in this dataset. Due to privacy concerns, there are few patients who are not willing to share the answer to some questions, these are known as missing values. It mainly focuses on the prediction of indicators of cervical cancer. Demographic Information, habits, and historic medical records were covered by features. This contains 36 features. One of the features is the Biopsy. This column represents that whether the biopsy report is positive or negative which are respectively one (1) and zero (0). The meaning of this is if the patient has cervical cancer or not and this one will be our output column and others will be for training.

B. Dataset Feature Analysis From Graph

Now we are going to have a statistical presentation of the features from the dataset. How the data are distributed throughout the rows and columns and if they are having any correlation among the features.

As we have depicted earlier that it has 36 feature columns here among these columns 35 will be used for training the model and the Biopsy column is our output. As the dataset is having 858 rows, it will be tough to see the whole dataset here and we won't be able to understand anything either. So, for having a gross idea of the dataset we are going to see here the

978-1-6654-2363-2/21/$31.00 ©2021 IEEE

statistical presentation of some feature columns, and from this, we will have some understanding of the data presented in the columns. It is also going to be cognizable which data correlate or not. Before using the data for presentation, we have first labeled it and this process also gets rid of any redundant values. So, the values we will see here will be the labeled ones instead of actual ones.

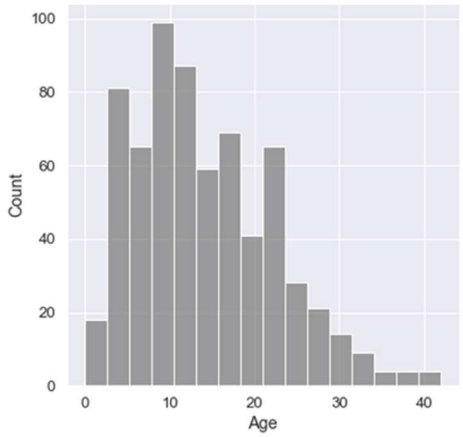

Fig. 1. Distributed age plot.

First, we have represented here the age in the x-axis which is the age of each patient observed here. Here, we can see in Fig. 1 That most of the patients diagnosed here are young, and very few are older.

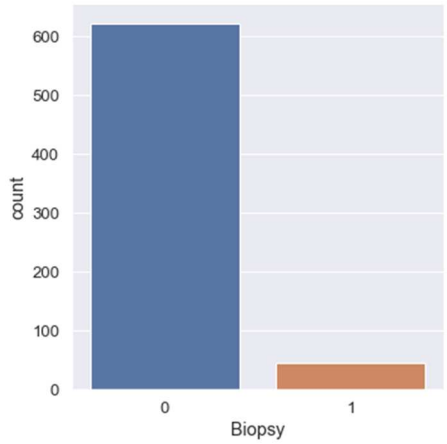

Fig. 2. Biopsy positive (1) or negative (0) plot.

As we have already said that Biopsy is a Boolean valued column. Test value one means cervical cancer positive and if we get zero value this means that it's negative. This representation presented in Fig. 2 on the biopsy test shows that most of the patients tested here are negative. Now if we look at Fig. 1 we can see that most of them are young age, so it makes some sense why the biopsy test positives are low, and negatives are so high. Because we know that in most cases cervical cancer develops in aged females.

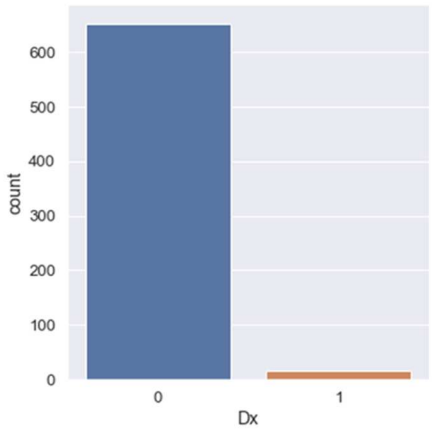

Fig. 3. Dx positive (1) or negative (0) Plot.

Here, in Fig. 3 the Dx represents that if there is the presence of Human Papilloma Viruses (HPV) or Cervical Intraepithelial Neoplasia (CIN) or both. So, in this case, we can see that the positive which is one is very low among all diagnosed individuals.

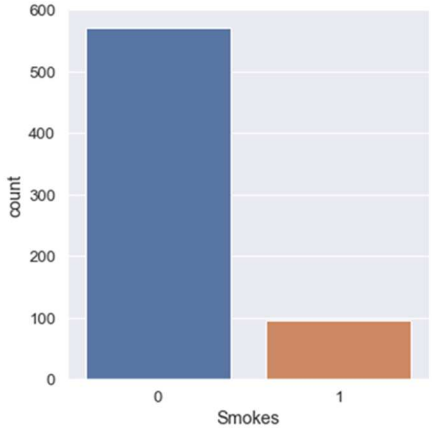

Fig. 4. Smokes (1) or not-smokes (0) plot.

Now, this is a dataset of people with cervical cancer that means all are female as we know. So, when we are talking about if they do smoke or not in that case we can see in Fig. 4 that most of them don't do this.

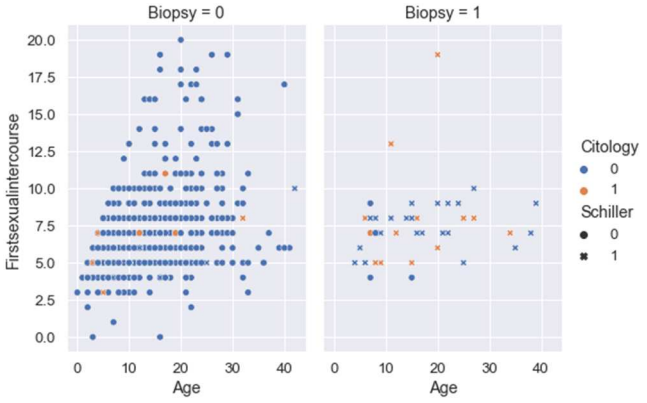

Fig. 5. Age and first sexual intercourse plotting with Biopsy, Cytology & Schiller.

From Fig. 5 we can see that first sexual intercourse is placed on the y-axis and this indicates that in most of the cases when it's done at a young age there is a higher chance of

developing cervical cancer and the x-axis age shows that most of the biopsy test positive detected patients are from middle-aged to higher. Schiller test positive and negative is indicated by cross a circular value respectively which shows that in most cases a positive identified person has come out positive in the biopsy. Color represents the cytology negative or positive respectively by blue and orange. For cytology negative test there is a much higher chance of getting the biopsy report negative.

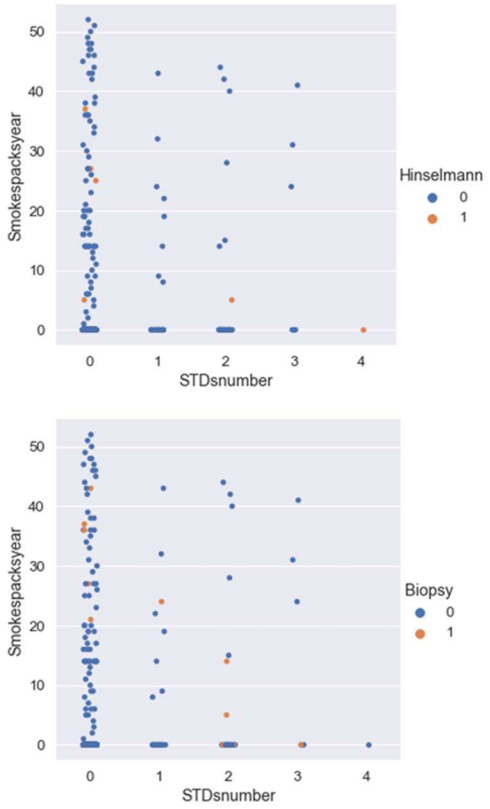

Fig. 6. No. of sexually transmitted disease with smokes (packs/year) and Biopsy plot.

Here the Fig. 6 represents the number of sexually transmitted diseases (STD) present in the patient on the x-axis and smokes (packs/year) the first one represents if the Hinselmann test positive or not and the second one represents Biopsy test positive or not. Here, with a higher smoking rate, there is a possibility of having cancer, and having STD there is also some possibility of developing cancer.

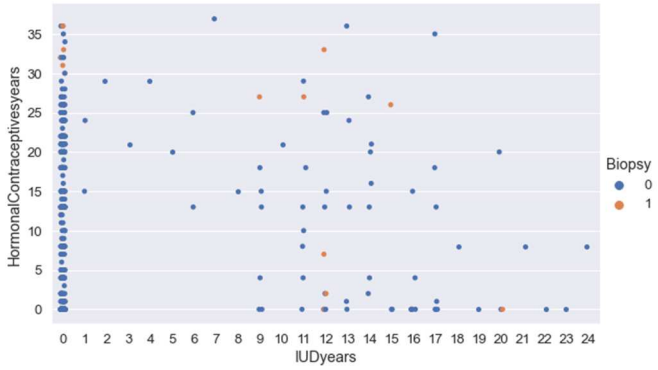

Fig. 7. Intrauterine device with oral contraceptives and biopsy plot.

In Fig. 7 there on the x-axis and y-axis, the usage time of the intrauterine device (IUD) and oral hormonal contraceptives have been presented respectively and color indicates the biopsy as previous. By observing this we can say that long time use of any of these methods has a higher chance of resulting in being positive in biopsy test meaning having cervical cancer.

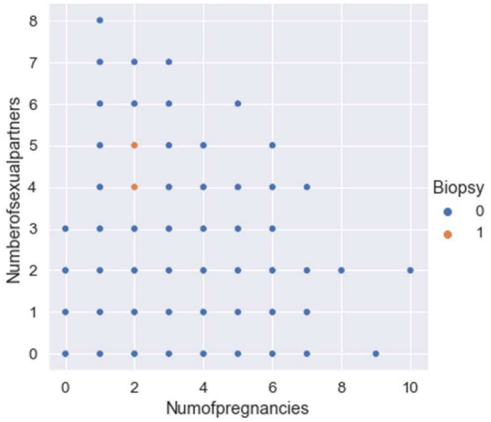

Fig. 8. No. of pregnancies with no. of sexual partners plot.

Fig. 8 shows values of the number of pregnancies and number of sexual partners respectively in the x-axis and y-axis. The number of pregnancies does not show that much of an impact, but a higher number of sexual partners have some impact on getting cancer positive in the cervix.

III. Methodology

The working procedure, shown in Fig. 9, we are going to perform here, which is a prediction on the biopsy test of cervical cancer. As we have already seen the dataset contains 858 rows and 34 training feature columns, which are patients' various kinds of data for example some personal data or some from diagnosis. But, if there are some missing values in the dataset, first we must clean the dataset so that there are no null values. Then we need to normalize the values of the dataset. By normalization, we transform the data into common scaled values. For here we need to scale down the values from zero to one as we are going to apply classifier algorithms. By analyzing the features, these need to be split into input features for training the model and the output data. In this dataset, the biopsy is our output data which is represented by Boolean value and the rest all are for input. After training the model we are going to be needing some test values for performance analyzing purposes. So that's why we have defined a part of the dataset as the training values and other as test values. For training Logistic regression (LR), Naive Bayes (NB), Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), k-Nearest Neighbor (kNN), and Random Forest (RF) is going to be used to make predictions both on training data and test data. Prediction on both data will help us to make a better comparison and have a deeper understanding of these evaluated output values. We are going to perform a detailed comparative analysis on outputs.

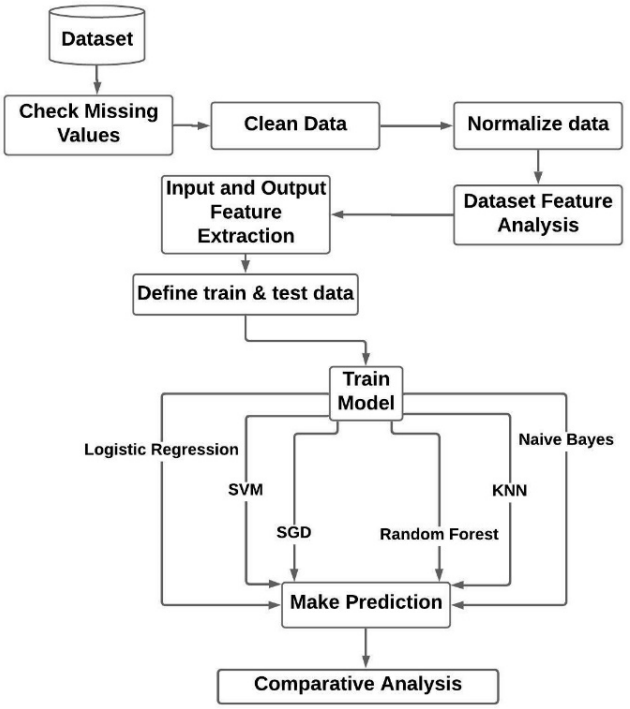

Fig. 9. Methodology.

IV. Result Comparison

From the Table. I we can see that we have used six machine learning algorithms for making predictions that the patient has been tested positive or negative in the biopsy test which are Random forest, k-Nearest Neighbor (kNN), Support Vector Machine (SVM), Logistic Regression (LR), Stochastic Gradient Descent (SGD), Naïve Bayes (NB). We have used the multinomial Naïve Bayes here. After that, we have evaluated the True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN) for all the algorithms presented in Table. I. Now from these, we have calculated the precision, recall, f1-score, and accuracy for comparing the performance which we can be found in the Table. II. Here, from accuracy we see that multinomial Naïve Bayes is the best performer with 97.76% and then followed by SGD, LR, SVM, kNN, and Random Forest. Precision represents the percentage of real true values from all true values and recall is the sensitivity that means the rate of picking genuine true values. F-score depicts the balance between these two parameters. The high value of this means low false positive and false negative values. Here, we can see that the f1-score is highest in multinomial NB then followed by LR, SGD, kNN, SVM, and Random Forest. As most of the data are categorical here and we have got the best performance from the multinomial Naïve Bayes.

TABLE I. True Positive, False Positive, False Negative, True Negative for All Algorithms.

Classifier	TP	FP	FN	TN
Random Forest	127	0	6	1
kNN	126	1	4	3
SVM	125	2	3	4
LR	126	1	3	4
SGD	125	2	2	5
NB	126	1	2	5

TABLE II. Precision, Recall, F1-score and Accuracy (Acc) for All Classifiers.

Classifier	precision	recall	f1-score	Acc
Random Forest	100	95.48	97.69	95.52
kNN	99.21	96.92	98.05	96.26
SVM	98.42	97.65	98.03	96.26
LR	99.21	97.67	98.43	97.01
SGD	98.42	98.42	98.42	97.014
NB	99.21	98.43	98.82	97.76

V. Conclusion and Future Scope

If cervical cancer reaches a severe state, then it's too difficult to cure. So, from the diagnosed data of patients, if we could make accurate predictions, it would have been possible to provide early treatment. Here from the data set analysis, we have seen that the data set is a bit imbalanced in terms of the biopsy because most of the cases are of negative patients and very few are of positive. There were lots of missing values because some patients don't want to share all personal info at basic diagnosis. Then again by looking at accuracy, we can't tell that which features are more impactful, and we may apply explanation techniques for this purpose. We may train on other datasets for further better understanding.

References

[1] A. Arora, A. Tripathi, and A. Bhan, "Classification of Cervical Cancer Detection using Machine Learning Algorithms," 2021 6th International Conference on Inventive Computation Technologies (ICICT), 2021, pp. 827-835.

[2] L. Akter, Ferdib-Al-Islam, M. M. Islam, M. S. Al-Rakhami, M. R. Haque, "Prediction of Cervical Cancer from Behavior Risk Using Machine Learning Techniques," SN COMPUT. SCI. 2, 2021, 177.

[3] J. Lu, E. Song, A. Ghoneim, M. Alrashoud, "Machine learning for assisting cervical cancer diagnosis: An ensemble approach," Future Generation Computer Systems, 2020, Volume 106, Pages 199-205.

[4] N. Razali, S. A. Mostafa, A. Mustapha, M. H. A. Wahab, N. A. Ibrahim, "Risk Factors of Cervical Cancer using Classification in Data Mining," J. Phys.: Conf. Ser., 2020, vol 1529, 022102.

[5] Repository U M L 2019 Retrieved from https://archive.ics.uci.edu/ml/index.php