Recall-based Machine Learning approach for early detection of Cervical Cancer

Apoorva Gupta
Department of Biotechnology
Delhi Technological University
Delhi, India
gupta.apoorva1399@gmail.com

Ashutosh Anand
Department of Biotechnology
Delhi Technological University
Delhi, India
ashutoshanand586@gmail.com

Yasha Hasija
Department of Biotechnology
Delhi Technological University
Delhi, India
yashahasija06@gmail.com

***Abstract*— With frequent advancements in development of algorithms and need to incorporate them with clinically synthesized medical information is the paramount of modern-day bioinformatics. This aspect of computational study is of great healthcare significance as deduced results could be farfetched to generalize conclusions in regular medicine practice and diagnosis thus fastening up the process of detection. This paper tries to generalize cervical cancer detection approach with random forest regression technique. Unlike other papers which focus on accuracy and precision, this paper emphasizes on recall-based approach and beneficial tenets this approach over former ones. Four diagnostic tests used for early stage detection of Cervical cancer are Hinselmann's test, Schiller's test, Biopsy and Cytology. Each test is studied individually and analysis was made on the basis of confusion matrix, recall score and receiver operator curve (ROC). The basic aim during the entire development was to achieve higher recall scores with reduced false positive values.**

Keywords— SMOTE, Expectation Maximization, Recall, ROC curve, Oversampling, Schiller, Cytology, Biopsy, Hinselmann, Cervical Cancer, Random Forest, Classification, SHAP

I. INTRODUCTION

As per the Indian Council of Medical Research (ICMR) estimates, there would be 1.04 lakh cases of cervical cancer in India in 2020. With a woman dying every 8 minutes due to this invasive disease the burden on existing healthcare regimes is immense to prevent its widespread. The disparity arising due to stark variation in socio-economic indicators of nations like every other community ailment has a significant effect on the impact assessment and policy frameworks undertaken to deal with it. Trends revealed from a survey done by the world cancer research fund further consolidate this notion with sub-Saharan African nations of Malawi, Swaziland, and Zambia being the worst-hit nations pan globe [1]. Although the results deduced from more developed regions show less prominence but a global outlook still indicates that the road for its complete eradication is still not half covered.

Oncologically cervical cancer involves the development of invasive malignant tumors in lower squamous regions of female genitalia. The prime reason in the majority of such instances is Human Papillomavirus Infection (HPI) that could have been acquired during sexual activity. Other factors may include environmental triggers or lifestyle complications like smoking or early age sexual activity. Considering the mortality rate especially in developing and underdeveloped countries the early-stage diagnosis of cervical cancer primarily becomes important because unregulated spread in high prominence zones could have irreversible demographic implications. Some diagnosis methods followed by present-day oncologists include:-

- Biopsy - One of the techniques applied for the diagnosis of cancers in general. In it, a tissue sample from the affected region (here lower narrow portion of the uterus) with a similar kind of sample from the unaffected region is taken. The contrast in results obtained from two different tissue analyses is used to predict cancer prevalence.
- Cytology - It is a preliminary test basically employed to detect pre-invasive cancerous lesions.
- Hinselmann Test - It is a visual diagnostic technique in which a colposcope is used for a magnified examination of the female genitalia. In the case of cervical cancer, premalignant and malignant lesions are largely observed
- Schiller Test - It is a biochemical diagnostic method. Iodine solution is applied to the cervical region. Cells in normal cervical mucosa contain glycogen hence give brown stain which hitherto in the case of cancerous cells is not observed. Positive Schiller test results are generally followed by histological analysis or biopsy.

All the above tests are either done alone or in combination for the accurate detection of cervical cancer. Generally, these are performed as confirmatory tests especially Schiller and Biopsy but after abnormal pap smear test results.

Early-stage prediction techniques with a high specificity could prevent the further widespread of carcinogenic cells due to effective treatment availability at an early stage and lesser malignancy of tumor leading to reduced complications that otherwise increase drastically. At the same time will supplement the present physiological and biochemical diagnostics in yielding faster and effective throughputs. With the healthcare sector becoming more data-driven concordant with recent advancements in Machine Learning (ML) algorithms the scope of development of proposed alternative techniques is very wide. They could act as a possible catalyst to accelerate the process. In oncology, ML-based approaches could enable us to find a needle in a haystack as they rule out possibilities of human-based errors and biases. Classification is the most important ML method that can be employed in the diagnosis, detection and management of cancer. In this paper random forest classifier algorithm is used on the acquired dataset with statistical implications and related visualizations.

II. BACKGROUND

A. Data Imbalance

The class imbalance is one of the major challenges faced in classification problems and has received a lot of attention in ML. A dataset having binary classes is bound to be imbalanced when one of the classes has more number of instances than the other one. The class with fewer instances is under-represented in the dataset. This class imbalance can have a drastic effect on the classification algorithm when applied to real-world problems where it becomes disastrous to wrongly classify the instances from the under-represented class. One such example is the early diagnosis of a disease as such datasets usually have few instances of patients having the disease. This under-representation of patients having the disease translates to the Machine Learning algorithm misclassifying the patients [2] [3].

Numerous solutions, both at data and algorithmic level, have been recommended to solve data imbalance problem, which is encountered quiet often. At the data level, several data re-sampling techniques such as random over- and under-sampling, and directed over- and under-sampling are employed [4]. Whereas at the algorithmic level, assigning weights to classes, probability estimate adjustment is some of the proposed solutions [5].

B. Random Oversampling

Random oversampling is a simple, non-heuristic technique to increase the instances of the minority class in order to balance the data by replicating the positive examples [6]. However, this might result in overfitting of the ML model as it just makes identical copies of the instances of minority class. Synthetic Minority Over-Sampling Technique (SMOTE), introduced by Chawla et al. [4], creates new instances of the minority class through interpolation between the already positive instances that are close together. These new instances enable the classifier to build larger decision regions near the minority class.

Oversampling the values in a dataset enhance the computational cost of the algorithm used for training and learning purposes. However, as shown by Batista et al. [6] and Barandela et al. [7], applying oversampling is recommended when minority class has fewer number of samples in the dataset as compared to the majority class, i.e., majority class/minority class ratio is high.

C. Random Forest Classifier

Random forest classifier is a powerful ensemble ML algorithm that has been proven to be very effective in pattern recognition and high-dimensional classification problems. Random forests were first introduced by Ho [8], Amit and Geman [9], and Breiman [10]. Random forest is a collection of decision trees and can be viewed as a classifier constituting various methods. The basic principle of random forests is to construct multiple binary trees using random bootstrap samples coming from a training set. Each tree in the forest makes classification on randomly selected sub-sample of the training data to yield a classification result. Then the forest selects the classification with maximum votes as the final result.

Random forests are based on the bootstrap aggregation concept of Breiman and the random feature selection concept of Ho. Therefore, individual trees instead of being trained on the whole dataset are trained on a subset of the dataset. For example, let the dataset has M instances then, by bootstrap manner, ⅔ of the instances are selected at random for individual tree and the rest are considered out-of-the bad observations to evaluate the error. Moreover, a random feature is chosen to be the decision node at each node. Therefore, for n number of features, the size of the feature selected at each split is either $\sqrt{n}$ or $\sqrt{n/2}$ [11].

D. Recall value in early detection

The rationale behind every classification-based machine learning project is to deduce higher accuracy results. For this to bring at evaluation different techniques are deployed from simpler ones like accuracy and precision values to complex ones like F1 score and recall values. The ambiguity arises in the selection of the technique depending and factors affecting the final results whether it be the nature of the dataset, its premise/theme, or every ML algorithm is hardlined with one of these techniques. Considering the data dynamics and veracity the later proposition could be sidelined easily however to deconstruct the former aspect we need to understand these techniques briefly in relation to our dataset:-

- Accuracy - It is the fraction of total right predictions positive and negative from the total aggregated sample space.
- Precision - It is the fraction of right positive cancer predictions from total aggregated sample space whether the patient has cervical cancer or not.
- Recall - It is the fraction of right positive predictions from the total aggregated sample space of patients having cervical cancer.

Evidently recall value thus obtained is a key to higher specificity and should be leveraged over accuracy because it involves wrong predictions as well which are of no prediction significance in this case. So, this leaves us with precision and it should also be outweighed because the former itself involves actual cervical cancer patients clearing the idea that we should not miss any actual positive case with negative prediction over an actual negative case with positive prediction. Thus, the higher the recall value higher the model specificity. Fringing benefits from it is the reduction in mental stress that one undergoes after listening about cancer, costs in medicare, and undergoing cumbersome diagnostic tests [12].

E. ROC Curve

The receiver operator curve abbreviated as ROC is a machine learning-based graphical visualization method area under which (AUC) is used for the determination of binary classifier's tendency to categorize between required classifiers correctly like a correct positive cervical prediction on one side and correct negative on the other [13]. Higher AUC value accounts for more accurate binary differentiation with lesser overlapping. For projects involving multi-algorithm analysis, they also help in the identification of the best algorithm for the used dataset. It can be done by plotting algorithm wise ROCs and selecting the one with maximum AUC. The plot constitutes true positive rates (sensitivity) and false-positive rates (1-specificity) to summarize confusion matrices. Visually ROC curves more peaked towards the upper left-hand corner tend to have a higher discriminant tendency [14].

III. Material and methods

A. Data Source and Features

Data is obtained from cervical cancer (risk factors) dataset available at the UCI Centre for Machine Learning and Intelligent Systems machine learning repository [15]. It is a multivariate dataset with 858 instances of cervical cancer patients reported at 'Hospital Universitario de Caracas' in Caracas, Venezuela. It acquired information on 36 different attributes ranging from demographic, lifestyle, and physiological to diagnostic features. Bottleneck arose due to missing data as several patients preferred not to share personal information due to privacy concerns. Thus, to deal with missing data a new imputation method based on the concept of expectation maximization was used.

Expectation maximization imputations are better than mean as it preserves the characteristic relationship of missing variables with associated features unlike mean which just makes simple replacement based on average from a single column. Statistically, it finds the maximum likelihood for a variable/missing value in a dataset where the value of related other variables is latent. Since data has missing values associated with variables having a distinct origin and features this algorithm is used.

Further, data was highly imbalanced with only 18 1's in Dx: Cancer (bool value for YES to cancer) column accounting for only 2.09% of true positive value. This imbalance could have led to erroneous values of specificity predictors. SMOTE was incorporated as a random oversampling technique to reduce the imbalance bias in the results. It further complimented the specificity of our alternative approach as well thus, serving the twin benefits. Dx: Cancer column acts as the label but, the final classification of whether the patient has cancer or not is made on the basis of four tests: - Hinselmann, Schiller, Cytology, and Biopsy. Therefore, the dataset was split into four parts, one for each diagnostic test.

B. Machine Learning Workflow

The execution starts with importing the dataset, since data has several missing values and data imbalance, expectation maximization and SMOTE is used to go away with it respectively. After preprocessing the training data is fed to a random forest regression algorithm. Predicted values are evaluated with test values with confusion matrix. Recall scores are obtained and further evaluated. The 5% reduction rule is used to remove two columns as they had maximum data cells missing before starting with processing. A unique method using Shapley Additive Explanations (SHAP) library is used. This recently developed model helps in comprehensive understanding of used machine learning models and output generated from it. The integration of this aspect enabled us to figure out the column attributes that have major involvement in forming the prediction thus affecting the values of ultimate indicators. Thus, the model for SHAP is developed separately for identification of most significant attributes and the results associated with each diagnostic test are evaluated.

IV. Results

To obtain the results, we first cleaned the dataset as explained in the earlier sections. Then the dataset thus obtained was fed to a Random Forest Classifier, whose attributes were set to their default values, in order to see how each diagnostic test performed in combination with a machine learning algorithm. We chose recall as the measure to gauge the success of the model. In addition to accurately detecting the patients having cervical cancer, the detector should also report the least number of false-negative cases because if an early disease detection algorithm is reporting a high number of false negatives then it defeats the core concept of early detection of the disease. Table 1. shows the confusion matrix and the recall score obtained on the test dataset. Biopsy has the highest recall score (0.996) and reports a minimum number of false-negative cases (1) whereas Cytology has the least recall score (0.912) and reports a maximum number of false-negative cases (22). However, all of the diagnostic tests have a recall score of greater than 0.9.

Further SHAP library is deployed to identify factors affecting the diagnosis of cancer most and the conclusions drawn from evaluation are very interesting. The top 3 factors responsible for development or diagnosis of cervical cancer are 'Dx:CIN'(Abnormal cells that grow on cervix epithelium. These cells are not carcinogenic but may turn invasive leading to cancer), 'Dx:HPV(Human Papillomavirus infection)' and 'First sexual intercourse' in case of Hinselmann's test, biopsy and cytology. Apparently the first 2 factors remained the same for Schiller's test as well but third factors came out to be 'Dx'. For more information, the Github repository of the developed code could be accessed [16].

Table 1. Confusion matrix and recall score for various cervical cancer diagnostic tests

S.no	Diagnostic Test	Confusion Matrix	Recall Score
1.	Hinselmann Test	$\begin{pmatrix} 251 & 0 \\ 20 & 231 \end{pmatrix}$	0.920
2.	Schiller Test	$\begin{pmatrix} 251 & 0 \\ 7 & 244 \end{pmatrix}$	0.972
3.	Cytology	$\begin{pmatrix} 251 & 0 \\ 22 & 229 \end{pmatrix}$	0.912
4.	Biopsy	$\begin{pmatrix} 251 & 0 \\ 1 & 250 \end{pmatrix}$	0.996

Figure 1. shows the ROC curve as well as the area under the ROC curve for both predictions using a random model and predictions using Random Forest Classifier. From Fig. 1. it is observed that Biopsy has the maximum AUROC value (0.998) whereas Cytology has the minimum AUROC value (0.956).

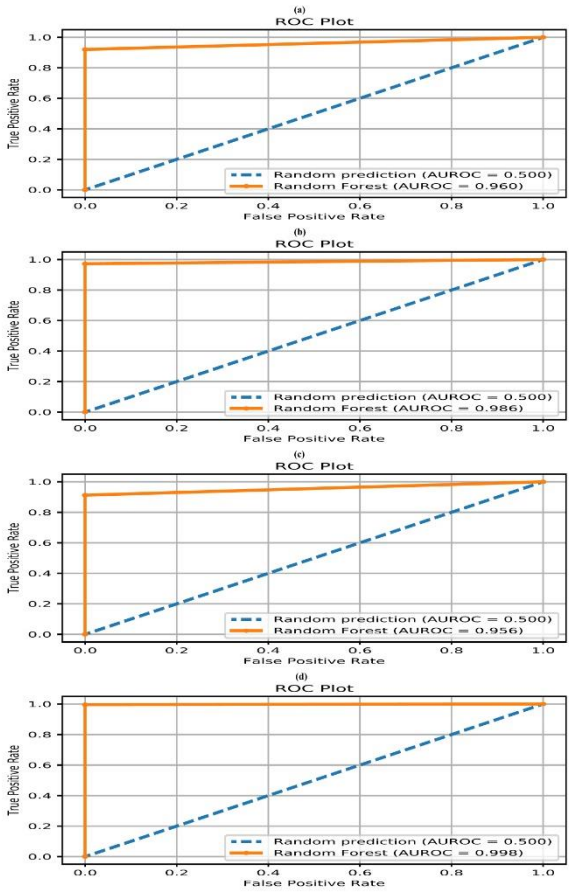

Fig. 1. ROC Curves along with AUROC values obtained from classifying the patients using both random model and Random Forest Classifier (a)Hinselmann (b)Schiller (c)Cytology (d)Biopsy

V. Conclusions

Our entire study comprehensively covered a major aspect in development of bioinformatics that is early detection of a disease in tested patients (cervical cancer here). Bioinformatic throughputs like these should be taken up at large scale. Since availability of data is the fuel to such studies, policy level interventions for major data storage in healthcare regimes are required. Bottlenecking arises due to lack of efficient data sources both in terms of volumes and versatility. Another main deduction from this study is further promotion of the idea of more applicability of studies focused on recall scores thus reducing false positive cases. The algorithm developed with assistance from SHAP library consolidated the fact that presence of CIN in cervix epidermis and HPV infection are the prime resaons that might lead to the development of cervical cancer and associated complications. Interestingly the age of first sexual intercourse came out to be the another most important factor for such tumour developments.

Chronic diseases like cervical cancer have serious ramifications not only on patients but also on society and rapid surge in positive instances aggravate the already stressed healthcare and diagnostic infrastructure. Adoption of such studies at mass level with better quality of data inputs could provide eminent relief to these shortcomings. Collaborative inter-governmental initiatives in such directions could also be a good step in waging the gap arising due to financial and infrastructural constraints in developing and underdeveloped countries. The persistence of early sexual intercourse as one of the prime factors further directs as to undergo deliberations to make sex education mandatory with impetus to personal and reproductive health. Thus, making such primary studies a foundation for transforming modern estates of healthcare will not only alleviate untimely loss of life but also enhance the socio-economic output of the nations.

References

[1] “Cervical cancer statistics,” 2018. https://www.wcrf.org/dietandcancer/cancer-trends/cervical-cancer-statistics.

[2] V. García, J. S. Sánchez, R. A. Mollineda, R. Alejo, and J. M. Sotoca, “The class imbalance problem in pattern classification and learning.”

[3] D. Ramyachitra and P. Manikandan, “IMBALANCED DATASET CLASSIFICATION AND SOLUTIONS: A REVIEW,” International Journal of Computing and Business Research.

[4] N. v. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[5] F. Provost and T. Fawcett, “Robust Classification for Imprecise Environments,” Machine Learning, vol. 42, no. 3, pp. 203–231, 2001, doi: 10.1023/A:1007601015854.

[6] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735.

[7] R. Barandela, R. M. Valdovinos, J. S. Sánchez, and F. J. Ferri, “The Imbalanced Training Sample Problem: Under or over Sampling,” 2004, pp. 806–814.

[8] Tin Kam Ho, “Random decision forests,” in Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1, pp. 278–282, doi: 10.1109/ICDAR.1995.598994.

[9] Y. Amit and D. Geman, “Shape Quantization and Recognition with Randomized Trees,” Neural Computation, vol. 9, no. 7, pp. 1545–1588, Oct. 1997, doi: 10.1162/neco.1997.9.7.1545.

[10] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[11] O. Okun and H. Priisalu, “Random Forest for Gene Expression Based Cancer Classification: Overlooked Issues,” in Pattern Recognition and Image Analysis, pp. 483–490.

[12] R. E. Sharpe et al., “Increased Cancer Detection Rate and Variations in the Recall Rate Resulting from Implementation of 3D Digital Breast Tomosynthesis into a Population-based Screening Program,” Radiology, vol. 278, no. 3, pp. 698–706, Mar. 2016, doi: 10.1148/radiol.2015142036.

[13] C. E. Metz, “Basic principles of ROC analysis,” Seminars in Nuclear Medicine, vol. 8, no. 4, pp. 283–298, Oct. 1978, doi: 10.1016/S0001-2998(78)80014-2.

[14] K. Hajian-Tilaki, “Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation,” 2013.

[15] K. Fernandes, J. S. Cardoso, and J. Fernandes, “Transfer Learning with Partial Observability Applied to Cervical Cancer Screening,” 2017, pp. 243–250.

[16] A. Gupta and A. Anand, “cervical_cancer_prediction.” 2020, [Online]. Available:- https://github.com/mr-sesquipedalian/cervical_cancer_prediction.git.