Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 01, January-2022

Comparative Study between Random Forest and Support Vector Machine Algorithm in Classifying Cervical Cancer

Bhavna Karuparthi[1], Abishek Mahesh[1]
[a]School of Computer Science and Engineering,
Vellore Institute of Technology,
Chennai, India

Abstract— **Cervical cancer, fourth in the most commonly found disease in women, causes deaths worldwide due to lack of adequate access to cervical screening methods such as Schiller, Hinselmann, Cytology, and Biopsy. According to the current data of WHO, Cervical cancer remains the second most frequent cancer in women, with an estimated 570,000 new cases in 2018. The prediction of Cervical cancer at the early stages remained a challenging task. Biomedical research uses Several Machine learning and deep learning algorithms for predictive analysis that identifies the factors and predicts cervical cancer cases. This study aims to use Data Mining techniques using the Support Vector Algorithm and Random Forest Algorithm to predict the Indications of Cervical Cancer Using the Biopsy Test and its technical and social comparison.**

Keywords—Cervical Cancer, Classification, Random Forest, Support Vector Machine

I. INTRODUCTION

Cancer refers to the collection of related diseases which make cells divide uncontrollably. Tumors can arise as a result of this excessive division that can become benign or malignant. The term "cancer case" refers only to malignant tumors. Cervical cancer topped the most dangerous amongst list because it can affect the female reproductive system.

The development of malignant tumors occurs because of the transmission of sexually transmitted human papillomavirus (HPV), and individual differences in susceptibility become the contributing factors in fighting against it; therefore, knowing when to take action becomes very important. For this reason, the health care domain works with complex and valuable datasets that capacitate them to predict the hidden knowledge in the massive volume of data.

The Random Forest Algorithm has the advantage in terms of performance measures such as accuracy, precision, recall, and F1 score (harmonic mean of the precision and recall) compared to the Support Vector Machine Algorithm, thus classifying the indications of cervical cancer using the biopsy test with more accurate true positives and negatives. Comparison of the results and performance of the Random Forest Algorithm and Support Vector Machine Algorithm based on the diagnosis reports of the biopsy test is essential due to the effectiveness of diagnosis methods for cervical cancer.

II. ALTERNATE TECHNOLOGY

The increase in technologies has led to various statistical algorithms that facilitate the extraction of information from raw data sets, especially in healthcare. Few commonly used data mining techniques include Support vector machine (SVM), Random Forest algorithm, Naïve Bayes, and Decision trees. The algorithm mentioned above helps predict Cervical cancer formed in the cervix tissues; however, SVM continues to be the most used.

Vapnic (1997) proposed a problem involving data classification and regression (Vapnic et al., 1997, p.2). The SVM algorithm implements a unique feature that ignores outliers and finds the hyperplanes with the maximum margin. This technique works by analyzing the hyperplanes with maximum margins to classify data from different categories. Initially, each data item plotted on the n-dimensional space where the n stands for the total number of features, which on Constant training improves the division of the original data into different groups via their labels. The kernel trick allows it to transform the low dimensional input to high dimensional input space, useful in non-linear separation problems; however, when large datasets indicate that the target classes overlap, the algorithm fails to predict information. SVM requires more comprehensive training to predict true positives and negatives compared to other algorithms.

III. SUPPORT

A. *Technical Details*

Cervical cancer prevention requires better understanding of the factors responsible for the formation. Medical data sets have become available for learning over the years; however, they remain imbalanced at certain times as the total number of patients outweighs non-patients. Random Forest Algorithm (RF), proposed by Breman, deals with the unbalanced data sets and increases the performance. The fundamental concept behind RF revolves around the wisdom of crowd approach which provides a stable and generalized results due to learning from diverse set of data. The uncorrelated models operate as a committee to outperform any individual constituent models.

This supervised machine learning algorithm uses the Classification and Decision Tree (CART) technique which splits the tree nodes using the top-down approach (Abdoh, 2018). The algorithm searches for the best feature among the random subset of features instead of finding the important feature while breaking each node; Thus, resulting in wide diversity and a better model. This continues till the end of the leaf node without pruning. Each tree classifies the target variables independently and votes for the final tree class. The overall classification stresses the majority obtained tress voting. When building each tree, bagging and feature randomness produce an uncorrelated forest of trees whose prediction occurs more accurately than any tree as shown in Fig 1.

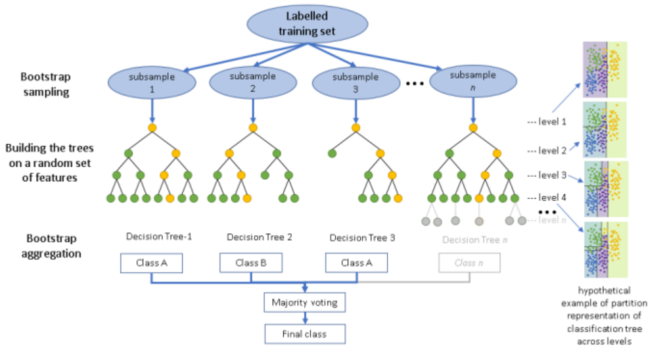


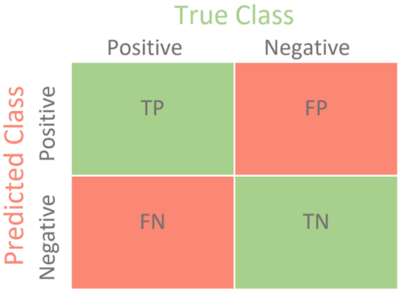
Fig 1. Random forest classifier

Construction of the Random Forest Algorithm requires quantitative number of bootstrap samples (N), the total number of features (M), sample size (M), and next node (k). The algorithm initially generates an N number of bootstrap samples from the dataset. Each node takes a random sample of size m where m < M. It then constructs a split for the m attributes selected and calculates the k node using the best break point. Splitting the tree continues until only one leaf node remains and the tree concludes until the algorithm finishes training on each bootstrapped node separately. The researchers can use either the Gini index formula or entropy formula to decide how nodes branch in the decision tree. The Gini index formula mentioned in Equation 1 uses the class and probability of each node to identify its index. The entropy formula stated in (2) uses the likelihood of a particular outcome to determine how the nodes should branch. Finally, the algorithm uses the tree's classification voting to collect the prediction data from the (n) trained trees, and using the highest voted feature, and it builds the final RF model. (Abdoh, 2018)

$$Gini\ Index = 1 - \sum_{i=1}^{c} (P_i)^2 \qquad (1)$$

$$Entropy = \sum_{i=1}^{c} -(P_i)(log_2(P_i)) \qquad (2)$$

The fully constructed RF model assesses accuracy based on the number of correctly classified samples to the number of total pieces using both regression and classification tasks where the classification of cervical cancer is done based on the preprocessed data (Geetha, 2019,para. 12). Upon deriving the results, the confusion matrix, as shown in Fig 2., is used for visualization and comparison of the performance.



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Fig 2. Confusion Matrix

A confusion matrix contains a tabular summary of the classifier's number of correct and incorrect predictions used to judge the performance of the classification model. The confusion matrix consists of four essential characteristics (numbers) used to define the measurement metrics of the classifier.

i. True positive (TP) represents the number of patients who have been appropriately classified to have cervical cancer
ii. True negative (TN) represents the number of correctly classified patients who are healthy.
iii. False positive (FP) represents the number of misclassified patients with cervical cancer, but actually, they are healthy.
iv. False negative (FN) represents the number of patients misclassified as healthy, but actually, they are diagnosed with cervical cancer.

Performance indicators like accuracy, precision, recall, and F1 score are further calculated using the four essential characteristics. They have various uses, including examining and comparing the performance of the SVM algorithm and Random Forest algorithm.

Equation 3 shows the accuracy as a ratio of all the correctly classified samples to the total number of test samples.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \qquad (3)$$

Precision (also called positive predictive value) is the ratio of true positive to the total positive instances detected by the model; as shown in (4), the ratio of true positive to the summation of true positive and false negative shown in (5). F1 score is described as the harmonic mean of the precision and recall, as shown in (6).

$$Precision = \frac{TP}{TP + FP} \qquad (4)$$

$$Recall = \frac{TP}{TP + FN} \qquad (5)$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (6)$$

TABLE I. EXPERIMENTAL RESULTS

Algorithms	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
Support Vector Machine	95.16	95.50	90.95	95.15
Random Forest	98.92	98.92	98.93	98.92

Note: Table 1 shows the two algorithms' performance in accuracy, precision, recall, and F1 score. (Ijaz, 2020)

Comparing the overall performance of Random forest and the Support Vector Machine algorithms in Table 1, Random Forest algorithm outweighs the Support Vector Algorithm and proves to provide better results in predicting cervical cancer. The performance analysis of these machine learning algorithms has helped improve the quality of service for the society and has contributed the medical industry fairly.

IV. SOCIAL IMPACTS

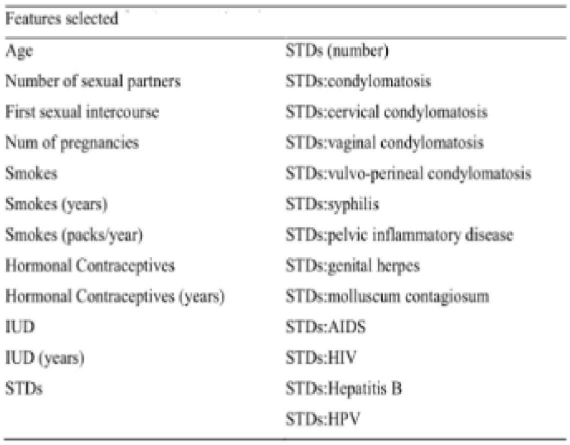
Over the past few years, the medical fraternity has a significant constraint to meet growing patient demands. The complexities of a global pandemic have been added to the challenges and the workloads of medical practitioners' remained significantly heavier than in the past. Such heavy workloads limited access to patient care, minimizing the impact they can have in their career – this remains a place where the machine learning algorithms such as Random Forest and SVM can step in and make a difference.

Cancer continues to be the primary health problems responsible for mortality in the world. The number of women who have had cervical cancer has increased continuously, and it has globally become the highest cause of cancer related deaths in women. Cervical cancer ranked fourth after breast cancer, this cancer took the lives of 311,000 women in 2018. Almost 79% of women younger than 45 had cervical cancer. (Arbyn, 2020, para. 3) Women of the eastern, western, middle, and southern Africa had cervical cancer as the leading cause of cancer-related deaths. Eswatini recorded the highest incidence, with 6.5% of women developing cervical cancer before 75. Countries like China and India recorded 106,000 and 97,000 cases, respectively. (Arbyn, 2020, para. 3)

About 90% of the 270,000 deaths by cervical cancer deaths in 2015 occurred from low-income and middle-income countries due to a lack of formalized screening programs (Yimer, 2021, para. 6). Though incidence and mortality vary with geographic location, high-incomed countries saw a decline in the age-standardized incident rates over the past 30 years. Vaccines that protect against common cancer-causing types of human papillomavirus can significantly reduce the risk of cervical cancer. Advanced medication and treatment methodologies have helped decrease cervical cancer in countries with good healthcare facilities.

Predicting cancer formation at early stages becomes very important to proceed with better medication and treatment. The clinics' screening methods facilitate the prediction and provides the patients with required outputs, but the amount of time consumed for prediction becomes a demerit. Comprehensive prevention, early diagnosis, effective screening, and treatment programs have reduced the mortality rate. Machine learning algorithms come to the rescue by speeding up the diagnosis and producing more accurate results, benefiting the health care industry. According to Derek Driggs, co-author of a paper from the lab of Carola-Bibiane Schönlieb at the University of Cambridge, the best way to help would be by setting rigid research standards that could help people develop models that could actually be useful to clinicians. (Scudellari, 2021).

The initial step towards running the algorithm involves the collection of various statistical information from the patients. The statistical and demographic features provided in Figure 3 helps the algorithm classify the type of cervical cancer.



Features selected	
Age	STDs (number)
Number of sexual partners	STDs:condylomatosis
First sexual intercourse	STDs:cervical condylomatosis
Num of pregnancies	STDs:vaginal condylomatosis
Smokes	STDs:vulvo-perineal condylomatosis
Smokes (years)	STDs:syphilis
Smokes (packs/year)	STDs:pelvic inflammatory disease
Hormonal Contraceptives	STDs:genital herpes
Hormonal Contraceptives (years)	STDs:molluscum contagiosum
IUD	STDs:AIDS
IUD (years)	STDs:HIV
STDs	STDs:Hepatitis B
	STDs:HPV

Fig 3. Selected parameters

The medical practitioners make use of the dependent features for cancer prediction, which decreases their work time and increases focus on the wellness and wellbeing of the patients. Most medical practitioners entered the sector to provide patient care but faced challenges and complex regulations. Also, excessive work has contributed significantly to burnout, causing exhaustion and practitioners quitting the medical industry entirely. These machine learning algorithms open up new ways of managing these complexities, minimizing these pressures, and reducing practitioner burnout (Ghassemi, 2021). On the other hand, patients benefit from the rapid detection of the disease. These algorithms impact is beyond expected and has a vibrant and exciting future.

CONCLUSION

Cervical cancer is considered the most communal malignant disease amongst women and about 300,000 women die because of cervical cancer yearly. Hence, it is important to understand the risk factors of cervical cancer to diagnose it in women faster. The two algorithms – Random Forest and support vector machine help in analyzing and evaluating the features of cervical cancer to produce diagnosis results. These machine learning algorithms have shown desirable outputs where random forest outweighed the support vector machine with better accuracy. The observations have shown that the usage of these algorithms decreased the time taken for cancer diagnosis and helped both patients and doctors predict the presence of cervical cancer precisely. The comparison of the algorithms benefits the data scientists' and other medical organizations whose research area deals with finding better solutions for detecting various cancers. As researched by processor Hu, AI is able to predict cancers faster and with better accuracy than doctors, as depicted in his study in 2019. This leaves us with the question- will AI replace doctors in the future? (Hu, 2019)

REFERENCES

[1] Abdoh, S. F., Rizka, M. A., & Maghraby, F. A. (2018). Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. IEEE Access, 6, 59475-59485. 10.1109/ACCESS.2018.2874063

[2] Alam, T. M., Khan, M. M. A., Iqbal, M. A., Abdul, W., & Mushtaq, M. (2019). Cervical cancer prediction through different screening methods using data mining. IJACSA) International Journal of Advanced Computer Science and Applications, 10(2). 10.14569/IJACSA.2019.0100251

[3] Arbyn, M., Weiderpass, E., Bruni, L., de Sanjosé, S., Saraiya, M., Ferlay, J., & Bray, F. (2020). Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. The Lancet Global Health, 8(2), e191-e203. https://doi.org/10.1016/S2214-109X(19)30482-6

[4] Geetha, R., Sivasubramanian, S., Kaliappan, M., Vimal, S., & Annamalai, S. (2019). Cervical cancer identification with synthetic minority oversampling technique and PCA analysis using random forest classifier. Journal of medical systems, 43(9), 1-19. https://doi.org/10.1007/s10916-019-1402-6

[5] Ghassemi M. (2021, February). How machine learning enhances healthcare. [Video]. TED. https://www.ted.com/talks/marzyeh_ghassemi_how_machine_learning_enhances_healthcare

[6] Hu, L., Bell, D., Antani, S., Xue, Z., Yu, K., Horning, M. P., ... & Schiffman, M. (2019). An observational study of deep learning and automated evaluation of cervical images for cancer screening. JNCI: Journal of the National Cancer Institute, 111(9), 923-932. https://doi.org/10.1093/jnci/djy225

[7] Ijaz, M. F., Attique, M., & Son, Y. (2020). Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. Sensors, 20(10), 2809.https://doi.org/10.3390/s20102809

[8] Kahng, J., Kim, E.-H., Kim, H.-G., & Lee, W. (2015). Development of a cervical cancer progress prediction tool for human papillomavirus-positive Koreans: A support vector machine-based approach. Journal of International Medical Research, 22(7), 518–525. https://doi.org/10.1177/0300060515577846

[9] Kumar R. (July 23). Random forest-a sturdy algorithm. [Image]. https://medium.com/nerd-for-tech/random-forest-sturdy-algorithm-d60b9f9140d4

[10] Kurniawati Y. E, Permanasari A. E & Fauziati. S. (2016). Comparative study on data mining classification methods for cervical cancer prediction using pap smear results, 2016 1st International Conference on Biomedical Engineering. (IBIOMED). 10.1109/IBIOMED.2016.7869827

[11] Mohajon J. (2020, May 28). Confusion matrix for your multi-class machine learning model. [Image]. https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

[12] Scudellari M. (2021, March 29). Machine learning faces a reckoning in health research [Article]. IEE3E. https://spectrum.ieee.org/machine-learning-faces-a-reckoning-in-health-research

[13] Sumana R.K. (2021, october). Early prediction of cervical cancer using machine learning algorithms. [Image]. https://www.researchgate.net/publication/355284115_Early_Prediction_of_Cervical_Cancer_Using_Machine_Learning_Algorithms

[14] W. Wu and H. Zhou, "Data-Driven Diagnosis of Cervical Cancer With Support Vector Machine-Based Approaches," in IEEE Access, vol. 5, pp. 25189-25195, 2017,10.1109/ACCESS.2017.2763984.

[15] X. Deng, Y. Luo and C. Wang. (2018, November), "Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods," 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), Nanjing, China, 10.1109/CCIS.2018.8691126.

[16] Yimer, N. B., Mohammed, M. A., Solomon, K., Tadese, M., Grutzmacher, S., Meikena, H. K., ... & Habtewold, T. D. (2021). Cervical cancer screening uptake in Sub-Saharan Africa: a systematic review and meta-analysis. Public Health, 195, 105-111. https://doi.org/10.1016/j.puhe.2021.04.014