Diagnosing Cervical Cancer Using Machine Learning Methods

Derya Yeliz Coşar Soğukkuyu
Institute of Graduate Studies
Altınbaş University
İstanbul, Turkey
0000-0002-0101-897X

Oğuz Ata
Software Engineering
Altınbaş University
İstanbul, Turkey
oguz.ata@altinbas.edu.tr

2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) | 978-1-6654-6835-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/HORA55278.2022.9800033

***Abstract*—Pre-diagnosis of any kind of cancer type has critical impact on person's life. According to World Health Organization Cervical cancer is still one of the most common gynecologic cancers in the world that affects women life. if medical experts focus on early diagnosis of the disease which is detecting symptomatic patients as early as possible, patients will have the best chance for successful treatment since Cervical Cancer is preventable. When cancer treatment is delayed, mortality rate decreases, and treatment becomes more complicated and expensive. Currently, systems based on Artificial Intelligence are used for decision making Machine learning techniques let automated detection of cervical cancer run more quickly and efficiently. In this study a novel ensemble approach is presented to predict the risk of cervical cancer by developing hybrid machine learning model. Multiple performance measurements such as Accuracy, precision score, recall score, and F1 are performed to evaluate the novel model. The results indicate that the proposed novel model can be effectively used to pre diagnosis of Cervical cancer with accuracy 97%.**

Keywords: Cervical Cancer, machine learning, classification, decision systems, artificial intelligence

I. INTRODUCTION

Every year, one in six deaths approximately 10 millions of people die from cancer and women suffer from Cervical cancer which is the fourth most common cancer in the world [1]. This is due to a lack of knowledge about the disease and restricted access to health care. Cervical cancer is one of the most successfully treated cancers once diagnosed, as long as it is detected early and properly handled. Developed countries, on the other hand, have plans in place to enable reliable and successful screening methods, allowing pre-cancerous lesions to be recognized and treated.Recently , there has been various studies for identifying cervical cancer. Sobar et al. [2] applied social science behaviour theory by using naive bayes algorithm to identify the risk of having cervical cancer with 91.67% accuracy. Wu and Zhou [3] used principal component analysis (PCA) techniques for feature selection to construct a machine learning classification model based on SVM for the diagnosis of cervical cancer with accuracy 90.48%. Abdoh et al. [4] diagnosed cervical cancer using random forest classifier with SMOTE and feature reduction techniques such as RFE with accuracy 96.06%.

This paper proposes a novel machine learning models for diagnosis of cervical cancer from demographic information, habits, and historic medical records of 858 patients. The following is a brief of the paper's structure: The approach for obtaining the reviewed papers is described in Section 2. Section 3 includes a comprehensive machine learning technologies presented by this paper. Section 4 contains observations and results, while Section 5 contains findings and recommendations for future research.

II. LITERATURE REVIEW

In this section, literature review is gathered related to Prediction of Cervical Cancer using machine learning models in healthcare sector. Table 1 shows an overview of the literature by application area.

Table 1 Overview of Prediction of Cervical Cancer using machine learning models

Author & Year	***Method***	***Dataset***	***Result***
Su et al. (2016) [5]	Composition of the C4.5 and Logistic Regression	pap-smear images	95.642%
Ashok et al. (2016) [6]	Support Vector Machine (SVM)	pap-smear images	98.5%
Wang et al. (2019) [7]	Support Vector Machine (SVM)	pap-smear images	96%
Ghoneim et al. (2019) [8]	Convolutional Neural Network (CNN)	pap-smear images	99.7%
Adem et al. (2019) [9]	stacked autoencoder	a data set containing 668 samples, 30 attributes and 4 target variables (Schiller, Citology, Biopsy and Hinselmann) from the UCI database	97.25%
Geetha et al. (2019) [10]	Random Forest (RF)	a data set containing 668 samples, 30 attributes and 4 target variables (Schiller, Citology, Biopsy and Hinselmann) from the UCI database	-
Ijaz et al. (2019)[11]	Random Forest (RF)	the repository of UCI collected at Hospital Universitario de Caracas in Caracas, Venezuela 858 instances with 36 features.	97.02%
Lu et al. [12]	an ensemble-based approach	UCI risk factor dataset	83.16%

978-1-6654-6835-0/22/$31.00 ©2022 IEEE

III. METHODOLOGY

A. Dataset description

In this study, a dataset of cervical cancer risk variables publicly collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela has been used which is available since 2017. A total of 858 patients' demographics, behaviours, and medical records are included in the dataset which consists of 36 indicators related to cervical cancer. Many people refuse to answer some questions due to privacy concerns. As a result, there are numerous missing entries in the data. Missing data has been identified and unnecessary columns have been dropped from the dataset. In the dataset only 6.4 % result has positive biopsy result so imbalance in the data has been handled by Oversampling Technique - Smote to overcome the data imbalance.

B. Decision Tree Machine Learning Algorithms

Decision Trees are a form of Supervised Machine Learning in which data is regularly separated based on parameter[13]. Two main parameters are decision nodes and leaves. These parameters form the tree. The decisions or final outcomes are represented by the leaves where the data is separated.

C. Random Forest Machine Learning Algorithms

Random Forest is a well-known machine learning algorithm that uses the supervised learning method that can be used for both classification and regression issues[14]. It is based on ensemble learning, which is a method of integrating several classifiers to overcome various problem and increase the model's performance.

D. Model Optimization

In this study, since sensitive medical data has been used, the recall score is very important. Predicting a cancer patient as healthy (non-cancer) is extremely risky, and if done incorrectly, outcome affects badly in the patient's life. When working with imbalanced datasets, the difficulty is that most machine learning techniques will overlook the minority class, resulting in poor performance. Oversampling examples in the minority class is one technique to tackle this problem where minority class samples are duplicated in the training dataset before fitting a model. This can help to balance the class distribution, but it doesn't give the model any extra information. In this study, imbalance of the target variable has been handled by applying the smote oversampling technique.

Machine learning algorithms can run more efficiently (less space or time complexity) and be more effective with fewer features. Irrelevant input features can lead to poor predicted performance in some machine learning methods. For feature selection, Recursive Feature Elimination (RFE) algorithm has been used whereas RFE removes features and then a new model is built from remaining attributes [15].

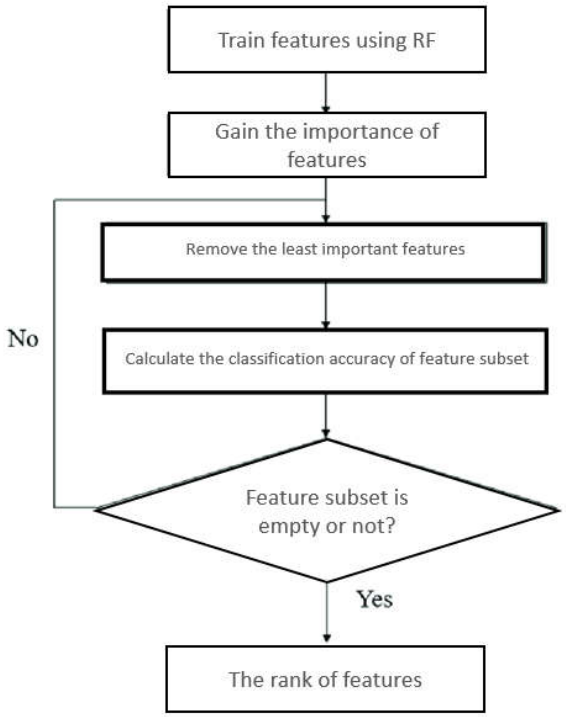

Figure 1:The recursive feature elimination (RFE) method's main procedure.

E. Application and Results

Research is carried out for hospital based in Venezuela evaluating the hospital's demographics, behaviours, and medical records. A decision system based on Artificial Intelligence was created for prediction of cervical cancer. Precision, recall, and F-score (harmonic mean of precision and recall), which are determined below formulas, were used to assess the model's performance.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP + FP + TN + FN}} \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}} \quad (2)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP + FP}} \quad (3)$$

$$\text{F1} - \text{score} = \frac{2*Precision*Recall}{recision+Recall} \quad (4)$$

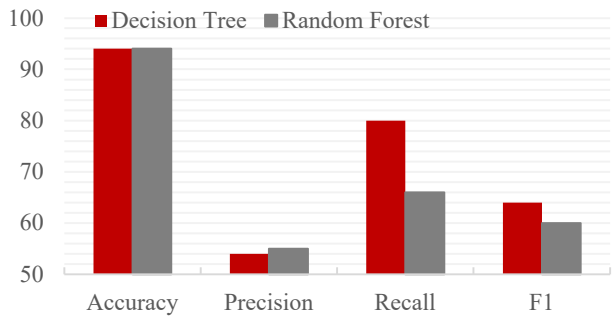

Figure 2Performance measures of each model

According to model results shown in Figure 1, accuracy performance is the same 94% for both decision tree and random forest, but maximum performance based on the experiment findings specially recall shows that patients who were predicted to be healthy had high cancer risk. And this is determined by checking Recall values. And highest recall is achieved with decision tree model with 80%.

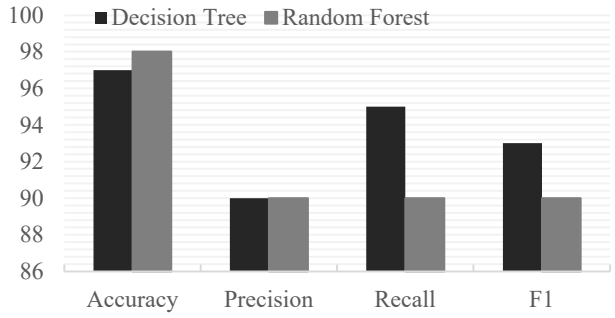

Figure 3:Performance measures of each model after applying feature selection with RFE

According to model after applying RFE feature selection technique results shown in Figure 2, maximum performance achieved with Decision Tree with accuracy 97% even random forest had accuracy score 98%. When all three metrics are considered, the results show that for Recall measurement, the Decision Tree algorithm performs better than other algorithms, with a score of 95% whereas the random forest reached 90%.

IV. Conclusion

This study can be a guideline for prediction of Cervical Cancer for based on machine learning techniques. Findings of this study reveal that even random forest gives higher accuracy, decision tree gives the highest performance in terms of performance indicators of recall and F1. For further directions, patients can consider giving more accurate personal results for dealing missing value. In this regard, supervised machine learning techniques can be used for prediction of Cervical Cancer based on clinical hospital data.

Acknowledgment

I'd want to express my gratitude to my advisor Oğuz ATA, who supported me developing the research and methodology. Your informative remarks encouraged me to improve my thoughts and raise the quality of my work..

References

[1] World Health Organization. (2022, April 10). https://www.who.int/news-room/fact-sheets

[2] Sobar, S.; Machmud, R.; Wijaya, A. Behavior determinant based cervical cancer early detection with machine learning algorithm. Adv. Sci. Lett. 2016, 22, 3120–3123.

[3] Wu, W.; Zhou, H. Data-Driven Diagnosis of Cervical Cancer With Support Vector Machine-Based Approaches. IEEE Access2017,5, 25189–25195

[4] Abdoh, S.F.; Rizka, M.A.; Maghraby, F.A. Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. IEEE. Access 2018, 6, 59475–59485.

[5] G. J. Su, X. Xu, Y. He, J. Song, Automatic detection of cervical cancer cells by a two-level cascade classification system, Anal. Cell. Pathol. 2016 (2016) http://dx.doi.org/10.1155/2016/9535027

[6] B. Ashok, P. Aruna, Comparison of feature selection methods for diagnosis of cervical cancer using SVM classifier, Int. J. Eng. Res. Appl. 6 (2016) 94–99

[7] P. Wang, L. Wang, Y. Li, Q. Song, S. Lv, X. Hu, Automatic cell nuclei segmentation and classification of cervical Pap smear images, Biomed. Signal Process. Control 48 (2019) http://dx.doi.org/10.1016/j.bspc.2018.09.008.

[8] A. Ghoneim, G. Muhammad, M.S. Hossain, Cervical cancer classification using convolutional neural networks and extreme learning machines, Future Gener. Comput. Syst. 102 (2020) http://dx.doi.org/10.1016/j.future.2019.09.015

[9] K. Adem, S. Kiliçarslan, O. Cömert, Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification, Expert Syst. Appl. 115 (2019) http://dx.doi.org/10.1016/j.eswa.2018.08.050.

[10] R. Geetha, S. Sivasubramanian, M. Kaliappan, S. Vimal, S. Annamalai, Cervical cancer identification with synthetic minority oversampling technique and PCA analysis using random forest classifier, J. Med. Syst. 43 (2019) http://dx.doi.org/10.1007/s10916-019-1402-6.

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

[12] J. Lu, E. Song, A. Ghoneim, M. Alrashoud, Machine learning for assisting cervical cancer diagnosis: An ensemble approach, Future Gener. Comput. Syst. 106 (2020) http://dx.doi.org/10.1016/j.future.2019.12.033.

[13] Zhifang, Sun & Yi, Li. (2020). Optimization of Decision Tree Machine Learning Strategy in Data Analysis. Journal of Physics: Conference Series. 1693. 012219. 10.1088/1742-6596/1693/1/012219.

[14] Murphy, Andrew & Moore, Candace. (2019). Random forest (machine learning). 10.53347/rID-67772.

[15] Magboo, Vincent & Magboo, Ma. (2021). Imputation Techniques and Recursive Feature Elimination in Machine Learning Applied to Type II Diabetes Classification. 201-207. 10.1145/3508259.3508288.