Identification of the Risk Factors of Cervical Cancer Applying Feature Selection Approaches

Mohiuddin Ahmed*, Mir Md. Jahangir Kabir†, Monika Kabir‡, Md. Mehedi Hasan§
Department of Computer Science and Engineering
Rajshahi University of Engineering and Technology
Rajshahi, Bangladesh
Email: *mohiuddin.nirob.mn@gmail.com, †mmjahangir.kabir@gmail.com, ‡monikakabir11@gmail.com, §mmehedihasann@gmail.com

***Abstract*—A risk factor is any attribute that enhances the probability of getting an injury or disease. Cervical cancer has become one of the leading causes of death around the world. The factors that increases the likelihood of cervical cancer are considered as the risk factors of cervical cancer. If the risk factors of cervical cancer can be identified, the reason behind this disease can be understood easily. Thus, through proper treatment, mortality rate can be minimized and the quality of care can be increased. The aim of this research is to identify the risk factors of cervical cancer and find a model that provides good performance in the prediction of cervical cancer. In this research, Recursive Feature Elimination (RFE) and Random Forests-based ensemble method have been applied to identify the factors that have much impact in the prediction of cervical cancer. Three classifiers such as Support Vector Machine (SVM), Multilayer Perceptron, and Logistic Regression have been used to evaluate the performance. SVM with Random Forests-based ensemble method has outperformed than others with the accuracy of 91.04%, specificity of 91.94% , and AUC score of 0.89 for an independent test set.**

Keywords - Risk Factors, Ensemble Method, Feature Importance, Random Forests, Support Vector Machine (SVM), Logistic Regression, Multilayer Perceptron.

I. Introduction

According to the fact sheet provided by the World Health Organization [1], cervical cancer appears to be the cancer, most commonly found among women residing in various developing countries worldwide.

In 2012, with an staggering 530,000 new cases of cancer have been discovered around the world. It covers 7.5% of all female cancer deaths. In developing countries, this situation is worse. Around 445,000 new cases(84% of the new cases that have been diagnosed worldwide) of cervical cancer have been diagnosed in the developing and under developed countries. This makes it the second most common cancer among women in those regions. In that year, approximately 270,000 women died from this fatal cervical cancer. Among them, 85% of women lived in under developed and developing countries.

In order to reduce these unfortunate deaths, we need to find the risk factors associated with the disease. Identifying the key factors is very crucial. Since there are many features associated with a disease, feature reduction technique can play a vital role in order to find the major factors related to cervical cancer. In this paper, RFE and Random Forests-based ensemble method have been used to select most important features from the datasets and find the risk factors.

After finding the risk factors, three classifiers namely SVM, Multilayer Perceptron, and Logistic regression have been applied on the dataset and the performances have been compared. So the main contributions of this research are identifying the risk factors of cervical cancer and finding a model that performs better in the prediction of cervical cancer.

The rest of this paper is embodied as follows: in section II a brief overview of the previous works is provided. Section III provides a description of the dataset used and the proposed methodology. Experimental results are reported in section IV. Finally, in section V we come to a conclusion.

II. Related Work

In last few years, different researchers studied cervical cancer data. The first contribution came from Fernandes et al. [2]. They tried to show the impact of transform learning method on sharping the accuracy. Cost-sensitive classification was used by Fatlawi et al. [3]. They applied Decision Tree classifier. The good results have been reported by Wu et al. [4]. They used Support Vector Machine with different approaches. They obtained about 94% accuracy. Researchers deliberated cervical cancer data set from different angles. For instance, Ghebre et al. [5] outlined the control of cervical cancer in HIV-positive women. Dasari et al. [6] conducted their study to develop novel biomarkers so that early diagnosis can be performed from specific proteins, enzymes and metabolites.

In previous works, Random Forests-based ensemble method was not used to calculate the relative importance of features of cervical cancer data set. In this paper, Random Forests-based ensemble method has been applied to select the critical factors of cervical cancer. Moreover, RFE has also been used as our another feature selection technique to identify the risk factors of cervical cancer. This paper has also provided a comparison among three classifiers such as SVM, Multilayer Perceptron, and Logistic Regression.

III. Methodology

A. Dataset Description

In our study, a dataset, collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela, has been considered.

This dataset has 36 attributes containing the demographic information, historical medical records, and habits of 858 patients [2]. There are some missing values because of the privacy concerns of some patients. Although this dataset is multiclass, we have only considered the class attribute 'Biopsy' in this research.

B. Data Preprocessing

The data has been preprocessed in the following way: The dataset is class imbalanced. SMOTE (Synthetic Minority Over-sampling Technique) has been applied to make the class balanced. For two attributes named 'STDs: Time since first diagnosis' and 'STDs: Time since last diagnosis, approximately 92% of the data is missing. So, these factors have been dropped and not been considered in our study. The samples having at least one null value for an attribute have been dropped. After data preprocessing, we have ended up with 1132 rows, 33 features, and one response variable 'Biopsy'. All 34 attributes including the class attribute obtained after data preprocessing are listed on Table I.

Finally, data preprocessing has come to an end with 1132 observations and 34 attributes including the class variable 'Biopsy'. After data preprocessing, the data has been passed into two independent processes where two feature selection techniques have been applied on the preprocessed data independently. Recursive Feature Elimination (RFE) and Random Forests-based ensemble method have been used to select the critical features.

- *Recursive Feature Elimination (RFE)* - RFE has been used to select the features that can be considered as the risk factors of cervical cancer.

- *Random Forests-Based Ensemble Method* - Random Forests-based ensemble method has been used to calculate the relative importance of every feature. Then a ranking of features has been obtained. The features having higher relative importance have been considered as the risk factors of cervical cancer.

After feature selection, three classifiers namely SVM, Multilayer Perceptron, and Logistic Regression have been applied in order to evaluate the performances. At last the results have been compared to find the expected outcome. The overview of our total methodology is shown in Fig. 1.

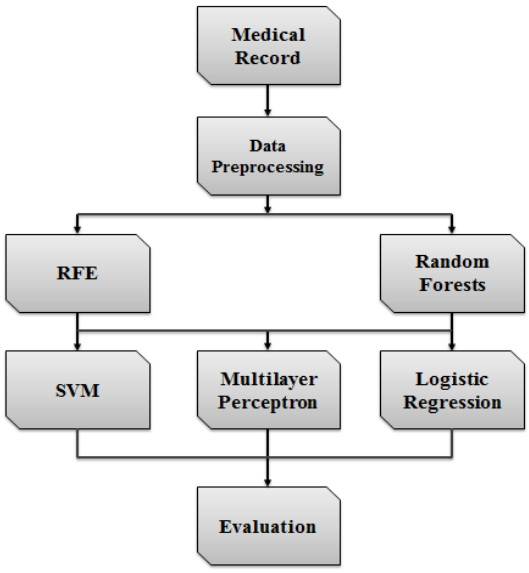

Fig. 1. Overview of the proposed methodology.

IV. Experimental Analysis

The data, after being preprocessed, has been passed through the intermediate phase where two different feature selection techniques have been applied independently on the data.

A. Recursive Feature Elimination (RFE)

80% of the data for training and 20% for testing purposes have been considered respectively. On the preprocessed data, we have applied SVM-RFE to identify the features that have much impact in the prediction of cervical cancer. Three classifiers namely SVM, Multilayer Perceptron, and Logistic Regression have been applied to build a model in order to obtain better accuracy, sensitivity, specificity, and AUC score. The number of selected features, accuracies, sensitivities, specificities, and AUC scores obtained after applying three classifiers are listed on Table II.

From Table II, it can be noticed that the number of selected features for SVM, Multilayer Perceptron, and Logistic Regression are 3, 3, and 8 respectively. Selected features for both SVM and Multilayer Perceptron are STDs:syphilis, STDs:genital herpes, and Schiller. In case of Logistic Regression, the selected features are Hormonal Contraceptives, STDs:syphilis, STDs:genital herpes, STDs:HIV, Dx, Hinselmann, Schiller, Citology. It is noticeable that the selected features- STDs:syphilis, STDs:genital herpes, and Schiller are common to all three cases.

It is also noticeable from Table II that Logistic Regression has outperformed than others in terms of AUC score.

B. Random Forests-Based Ensemble Method

80% of the data for training and 20% for testing purposes have been considered respectively. On the preprocessed data, we have applied Random Forests-based ensemble method to calculate the relative importance of every feature. Thus a ranking of features has been obtained. Then the top features that provide the best AUC score have been selected and considered as the risk factors. This process has been performed in case of all three classifiers independently. Fig.2, Fig.3, and Fig.4 show the ranking of relative importance of features obtained applying Random Forests-based ensemble method in case of SVM, Multilayer Perceptron, and Logistic Regression respectively. On the reduced dataset containing the selected features, the performances of three classifiers namely SVM, Multilayer Perceptron, and Logistic Regression

TABLE I
THE LIST OF OBTAINED ATTRIBUTES AFTER DATA PREPROCESSING

Feature Name	Type	Feature Name	Type
Age	Integer	STDs:syphilis	Boolean
Number of sexual partners	Integer	STDs:pelvic inflammatory disease	Boolean
First sexual intercourse (age)	Integer	STDs:genital herpes	Boolean
Number of pregnancies	Integer	STDs:molluscum contagiosum	Boolean
Smokes	Boolean	STDs:AIDS	Boolean
Smokes (years)	Boolean	STDs:HIV	Boolean
Smokes (packs/year)	Boolean	STDs:Hepatitis B	Boolean
Hormonal Contraceptives	Boolean	STDs:HPV	Boolean
Hormonal Contraceptives (years)	Integer	STDs: Number of diagnosis	Boolean
IUD	Boolean	Dx:CIN	Boolean
IUD (years)	Integer	Dx:HPV	Boolean
STDs	Boolean	Dx	Boolean
STDs (number)	Integer	Hinselmann	Boolean
STDs:condylomatosis	Boolean	Schiller	Boolean
STDs:cervical condylomatosis	Boolean	Citology	Boolean
STDs:vaginal condylomatosis	Boolean	Dx:Cancer	Boolean
STDs:vulvo-perineal condylomatosis	Boolean	Biopsy	Boolean

TABLE II
NUMBER OF SELECTED FEATURES, ACCURACIES, SENSITIVITIES, SPECIFICITIES, AND AUC SCORES OBTAINED APPLYING THREE CLASSIFIERS FOR RFE FEATURE SELECTION APPROACH

	SVM	Multilayer Perceptron	Logistic Regression
Features	3	3	8
Accuracy	90.30%	89.55%	90.30%
Sensitivity	80.00%	80.00%	80.00%
Specificity	91.13%	90.32%	91.13%
AUC Score	0.86	0.86	0.89

TABLE III
NUMBER OF SELECTED FEATURES, ACCURACIES, SENSITIVITIES, SPECIFICITIES, AND AUC SCORES OBTAINED APPLYING THREE CLASSIFIERS FOR RANDOM FORESTS-BASED ENSEMBLE METHOD

	SVM	Multilayer Perceptron	Logistic Regression
Features	6	4	1
Accuracy	91.04%	89.55%	90.30%
Sensitivity	80.00%	80.00%	80.00%
Specificity	91.94%	90.32%	91.13%
AUC Score	0.89	0.88	0.86

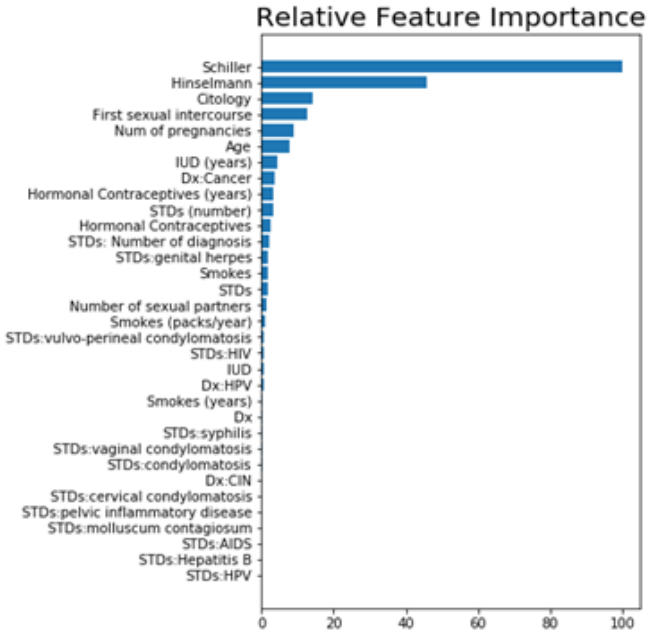

Fig. 2. Relative importance of 33 features obtained applying Random Forests-based ensemble method in case of SVM.

have been evaluated to build a model in order to obtain better accuracy, sensitivity, specificity, and AUC score. The number of selected features, accuracies, sensitivities, specificities, and AUC scores obtained after applying three classifiers for Random Forests-based ensemble method are listed on Table III.

In case of SVM, the selected 6 features are Schiller, Hinselmann, Citology, First sexual intercourse (age), Number of pregnancies, and Age. Schiller, Hinselmann, Citology, and Age are the selected 4 features in case of Multilayer Perceptron. In case of Logistic Regression, Schiller is the only selected feature. It is noticeable that the selected feature- Schiller is common to all three cases.

It can be noticed from Table III that SVM has outperformed than others in terms of accuracy and specificity. In case of RFE, the performance of Logistic Regression is better. Fig. 5 shows a comparison between Logistic Regression and SVM.

From Fig. 5, it is clearly seen that SVM with Random Forests-based feature selection has outperformed than Logistic Regression with RFE in terms of accuracy and specificity. So the selected features in case of SVM- Schiller, Hinselmann, Citology, First sexual intercourse (age), Number of pregnancies, and Age can be considered as the risk factors of cervical cancer.

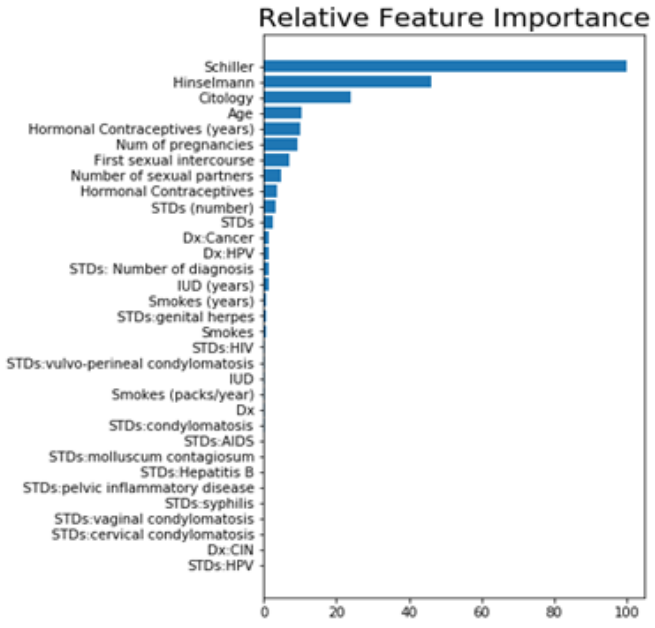

Fig. 3. Relative importance of 33 features obtained applying Random Forests-based ensemble method in case of Multilayer Perceptron.

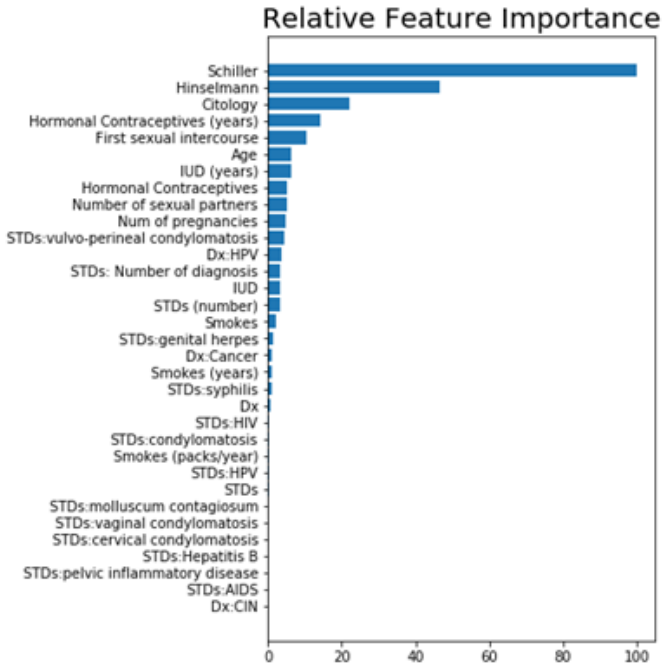

Fig. 4. Relative importance of 33 features obtained applying Random Forests-based ensemble method in case of Logistic Regression.

V. Conclusion and Future Work

In this research, we have worked on identifying risk factors of cervical cancer. In order to select the critical factors two feature selection approaches- RFE and Random Forests-based ensemble method have been used. To obtain a better predicting

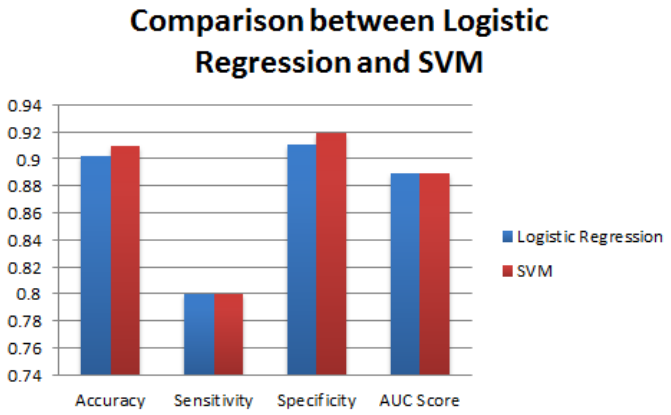

Fig. 5. Comparison between Logistic Regression and SVM

model that can be used to identify the risk factors, three classifiers namely SVM, Multilayer Perceptron, and Logistic Regression have been applied and the performances have been evaluated. After analyzing the performances, it has been concluded that SVM with Random Forests-based ensemble method has outperformed than others in terms of accuracy, specificity, and AUC score. This model has identified 6 risk factors of cervical cancer namely Schiller, Hinselmann, Citology, First sexual intercourse (age), Number of pregnancies, and Age. These risk factors are only the outcome of a model without considering any knowledge of medical specialists. So, we will be discussing about our findings to the medical specialists in order to gain information from the real medical system and implement the suggested idea. Our study will be conducted on another three classes of the cervical cancer dataset. Different sorts of feature reduction approaches can also be applied. Our approach will be applied on different datasets in future.

References

[1] W. H. Organization *et al.*, "Human papillomavirus (hpv) and cervical cancer fact sheet," 2016.

[2] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Transfer learning with partial observability applied to cervical cancer screening," in *Iberian conference on pattern recognition and image analysis*. Springer, 2017, pp. 243–250.

[3] H. K. Fatlawi, "Enhanced classification model for cervical cancer dataset based on cost sensitive classifier," *International Journal of Computer Techniques*, vol. 4, no. 4, pp. 1–8, 2017.

[4] W. Wu and H. Zhou, "Data-driven diagnosis of cervical cancer with support vector machine-based approaches," *IEEE Access*, vol. 5, pp. 25 189–25 195, 2017.

[5] R. G. Ghebre, S. Grover, M. J. Xu, L. T. Chuang, and H. Simonds, "Cervical cancer control in hiv-infected women: Past, present and future," *Gynecologic oncology reports*, vol. 21, pp. 101–108, 2017.

[6] S. Dasari, R. Wudayagiri, and L. Valluru, "Cervical cancer: Biomarkers for diagnosis and treatment," *Clinica chimica acta*, vol. 445, pp. 7–11, 2015.