An Empirical Study of Cervical Cancer Diagnosis using Ensemble Methods

Enamul Karim
Department of CSE
Daffodil International University
Dhaka, Bangladesh
enamul.cse@diu.edu.bd

Nafis Neehal
Department of CSE
Daffodil International University
Dhaka, Bangladesh
nafis.cse@diu.edu.bd

***Abstract*—Cervical Cancer, being one of the most pressing issues now-a-days, needs to be addressed properly. With a view to achieving an accurate diagnosis method for Cervical Cancer by screening the risk factors, different machine learning approaches have been taken over time. But by analyzing the performances of most of state-of-the-art approaches, it was inferred that there is still room for improvement by developing a more accurate model. Hence, in this paper an approach using ensemble methods with SVM as the base classifier has been taken. The ensemble method with Bagging technique achieved an accuracy of 98.12% with very high precision, recall and f-measure value.**

***Index Terms*—Ensemble methods, Bagging, Machine learning, cervical cancer, risk factors**

I. Introduction

Cervical cancer has turned out to be a major concern around the world. According to the World Health Organization (WHO), it is the fourth most common cancer in women. Cervical cancer represents almost 7.9% of the total female cancer that occurred in 2012 [1]. It arises in the cervix cells in uterus and has the potential to invade other parts of the body [2]. A number of risk factors are associated with this type of cancer. It is very pivotal to make a prediction based on the risk factors for the detection of the disease. An early identification can play a crucial role in the treatment process. Proper medical care, after an early detection can increase the chances of survival to a great extent [1]. Different stages of Cervical Cancer has been shown in Fig 1.

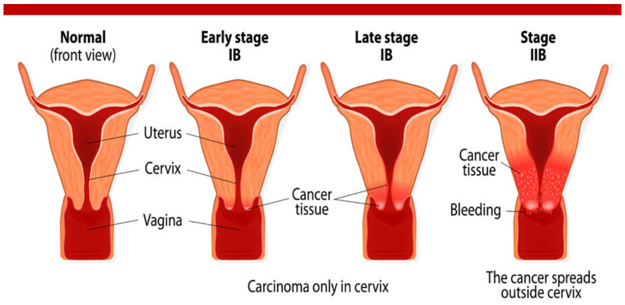

Figure 1. Cervical Cancer [3]

In this study, different machine learning algorithms were carried out based on the risk factors in order to make a comparative analysis in the intervention of the disease. Four classification algorithms such as SVM, Decision tree, Multilayer Perceptron and K-nearest neighbors were used for the purpose in **WEKA** data mining tool. Classification accuracy was then improved by applying different ensemble methods. Bagging and boosting were performed on the data-set to acquire better accuracy.

The organization of the remaining paper is as follows - Earlier works related to cervical cancer is presented in Section II, Section III represents the overall methodology. Section IV shows the performances of the experiment done and Section V summarizes the results with a short discussion. Finally, Section VI concludes the paper with an overall discussion.

II. Literature Review

Cervical Cancer has been a red-flagged issue for a very long time now. Proper diagnosis of cervical cancer has become a very challenging and demanding domain of research. Various machine learning based attempts has been made in order to achieving a classification model with high accuracy for screening cervical cancer.

Table I
Literature Review

Author	Accuracy	Classifier
Kelwin et. al. (2017) [4]	68.30%	SVM
Hayder et. al. (2017) [5]	42.9%	Decision Tree
Wesabi et. al. (2018) [6]	97.5% 49.0% 88.2% 87.6%	Decision Tree Logistic Regression SVM KNN
Wen et. al. (2017) [7]	94.13% 94.03% 94.03%	SVM SVM-RFE SVM-PCA
Ramit et. al. (2017) [8]	94.92%	Random Forest
Fahri et. al. (2018) [9]	95.89% 95.89% 97.26%	kNN MLP Bayes Net

978-1-7281-3445-1/19/$31.00 ©2019 IEEE

All the above mentioned papers in TABLE I have used the same dataset. Our empirical result is also based on the same dataset with 36 attributes containing the medical records of 858 patients. In [4], a transfer learning based approach has been taken. The authors proposed a strategy that facilitates target and source models to share the same coefficient signs for reducing the amount of labeled data from each expert/modality. In [5], a 3-staged cost sensitive classifier for cervical cancer screening was proposed. The classifier was a decision tree with cost selectivity and the model was evaluated with 10-fold cross validation.

In [6], diverse classification techniques on cervical cancer dataset have been applied. The result shows the advantage of feature selection approaches to the best predicting of cervical cancer disease. The results also showed that first sexual intercourse, age, hormonal contraceptives, number of pregnancies, smokes, and STDs:genital herpes are the main predictive features for cervical cancer. In [7] three SVM based approaches, namely - Simple SVM, SVM-RFE and SVM-PCA were introduced for cervical cancer classification. According to their result, SVM-PCA method was the most accurate method.

In [8] a penalty function was introduced to the already existing fitness function for Binary Firefly Algorithm. This addition radically reduced to an optimal subset which gained an increased classification accuracy compared to the traditional deep learning and information gain methods. In [9] a comparative study of the performance of popular classification algorithms like Multilayer Perceptron, BayesNet and k-Nearest Neighbor was done.

All the results from the aforementioned papers are presented in Table 1.

III. Methodology

The overall workflow had been divided into some major steps. Each and every steps are shown in Figure 2 and discussed in details afterwards.

A. Data Collection

The data-set used for the classification was collected from the UCI machine learning repository. This data-set was originally collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela. The data-set comprises demographic information, habits, and historic medical records of 858 patients. Several patients decided not to answer some of the questions because of privacy concerns. So definitely there were some missing values present in the dataset which was handled in the later step.

The classifications were carried out on the medical information of 858 different patients with 36 features. Table I shows the information of the available attributes. Four attributes such as **Schiller**, **Hinselmann**, **Biopsy** and **Cytology** were selected as the class attributes for classification.

Table II
Attributes [10]

Attribute Name	**Type**	**Attribute Name**	**Type**
Age	Integer	First sexual intercourse (age)	Integer
Number of sexual partners	Integer	STDs:molluscum contagiosum	Boolean
STDs:pelvic inflammatory disease	Boolean	STDs:genital herpes	Boolean
Number of pregnancies	Integer	STDs:AIDS	Boolean
Smokes	Boolean	Smokes (years)	Integer
Smokes (packs/year)	Integer	STDs:HPV	Boolean
STDs:HIV	Boolean	STDs:Hepatitis B	Boolean
Hormonal Contraceptives	Boolean	Hormonal Contraceptives (years)	Integer
STDs: Time since first diagnosis	Integer	STDs: Number of diagnosis	Integer
Intrauterine Devices (IUDs)	Boolean	STDs: Time since last diagnosis	Integer
IUDs (years)	Integer	Dx:Cancer	Boolean
Sexually Transmitted Diseases (STDs)	Boolean	Dx:CIN	Boolean
STDs:condylomatosis	Boolean	STDs:pelvic inflammatory Dx	Boolean
STDs (number)	Integer	Dx:HPV	Boolean
STDs:vaginal condylomatosis	Boolean	STDs:vulvo-perineal condylomatosis	Boolean
STDs:cervical condylomatosis	Boolean	STDs:syphilis	Boolean
Hinselmann	Boolean	**Biopsy**	Boolean
Schiller	Boolean	**Cytology**	Boolean

Each patient can be classified into four different classes at the same time since it is a multilabel and multiclass problem. The classifications were done on WEKA data mining tool. A cross-validation with 10 folds were conducted on the dataset to evaluate the performances of the different algorithms used for comparison.

B. Pre-processing

The dataset collected contained a lot of missing values. After analysis, it was seen that two attributes named ***STDs: Time since last diagnosis*** and ***STDs: Time since first diagnosis*** contained the highest number of missing data. So, they were removed from the dataset and classifications were done based on the other 34 attributes. The missing values of these 34 attributes were filled up by the median of the corresponding column in java.

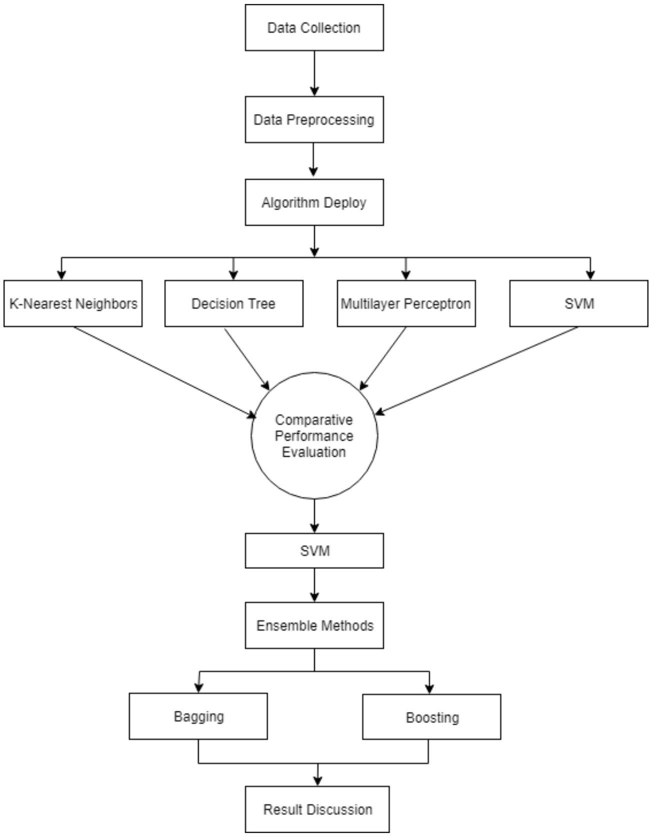

Figure 2. Overall Work flow.

C. *Classification Algorithms*

Decision Tree:

Decision tree is a fast learning algorithm that enables us to classify and predict a target variable. It is a tree like structure in which every non-terminal nodes represent a condition and the terminal nodes contain class labels [5].

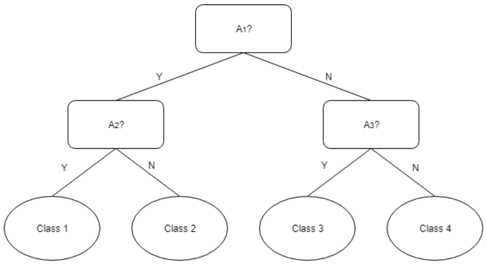

Figure 3. Decision Tree Induction.

Multilayer Perceptron:

Multilayer Perceptron (MLP) is a type of neural network model that makes the use of supervised learning technique where the input sets are mapped into proper output sets. It consists of different layers of nodes in which one layer is connected to the next layer. An MLP typically has 3 types of layers, namely - Input Layer, Hidden Layer and Output Layer. Every node, other than the input nodes, is a processing unit called neuron that has a nonlinear activation function. In order to train the network, back propagation is typically used in this algorithm. [9]

SVM:

SUPPORT vector machine is a strong classifier for solving regression problems [11]. It has a very easy implementation. A hyperplane is chosen to separate all the points in the input variable with respect to their classes. Sequential Minimal Optimization (SMO) is one of the ways to solve the SVM problem. According to [12], SMO is fast and has good scaling properties.

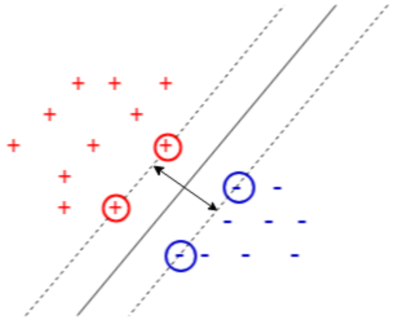

Figure 4. Optimal hyperplane separating two classes.

K-Nearest Neighbor:

The k-NN is one of the supervised learning techniques for solving classification problems. One crucial point is that the features are to be determined in advance. According to this algorithm, the distance between the desired item to be classified and the previous items is determined and the nearest k class is taken. Identifying the number of neighbors is a very important optimization problem in this algorithm [9].

Ensemble Methods Ensemble methods are hybrid algorithms that combine different machine learning algorithms into one predictive model and decrease bias (boosting), variance (bagging) or improve predictions (stacking). Most ensemble methods use a single base learning algorithm to produce homogeneous base learners.

1) Boosting: Boosting is one of the ensemble methods which is used for improving the performances of a weak classifier. Sequential learning of the predictors is used here. Initially, the whole data set is used for learning. Subsequently, learning occurs through the training sets based on the previous performances. The weights of the misclassified instances are increased so that the possibility of appearing in the training set of the following predictor gets higher [13]

2) Bagging: Bagging is an ensemble methods which is used to improve unstable classification problems. If the position of a training point changes marginally, weak classifiers can become unstable. Bagging can be applied to different classification algorithms. It is a very useful technique for large data sets in which detecting a good model is difficult due to the complexity and scale of the problem [13].

IV. Performance Evaluation

Performances were determined based on accuracy, precision, recall and f-measure. Table II shows the equations used for calculating the desired performance measures.

Table III
Performance Evaluation Matrix [14]

Measure	Equation
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Recall	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
F-Measure	$\frac{2XRecallXPrecision}{Recall+Precision}$

Using 10 folds cross validation, different classification algorithms were conducted on the data-set to obtain the performance measures which is represented by the following graphs.

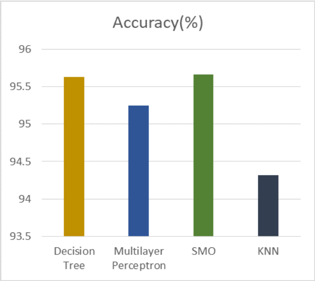

Figure 5. Accuracy.

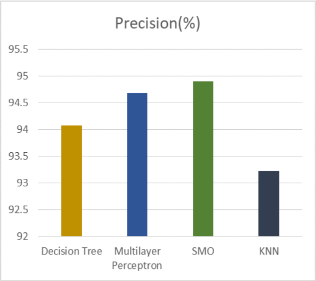

Figure 6. Precision.

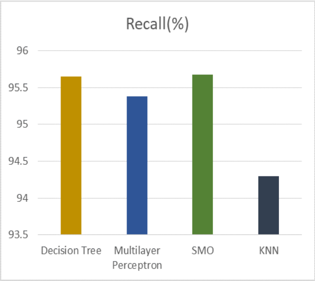

Figure 7. Recall.

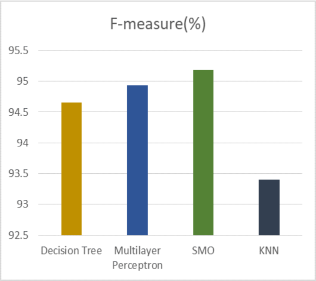

Figure 8. F-measure.

After analyzing the results obtained, it is seen that SMO offers us the best performance. SMO is superior to all other classification algorithms used in respect of accuracy, precision, recall and f-measure. Two ensemble methods namely AdaBoostM1 and bagging were conducted using 10 fold cross validation and SMO as the classifier.

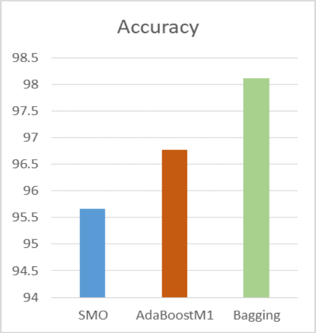

Figure 9. Accuracy.

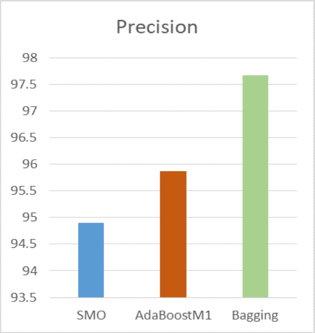

Figure 10. Precision.

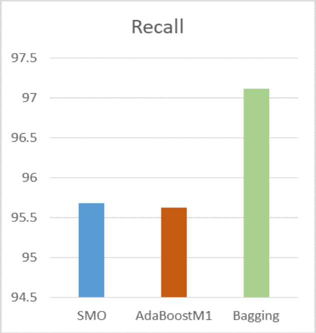

Figure 11. Recall.

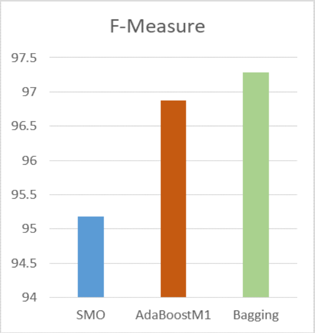

Figure 12. F-measure.

V. Results and Discussion

Four classification algorithms were used in our study. After analyzing the performances, it was seen that SMO had a better performance than the others. So, SMO was selected as the base classifier and two ensemble methods namely AdaboostM1 and bagging were carried out to enhance the performance. AdaboostM1 had a slight increase in performance but bagging had the bigger impact.

Machine learning techniques aim to find a single model that best predicts our desired result. Instead of using a single model as the predictor, ensemble methods take into account different models and average those models to generate a final stable model.

Bagging is a powerful ensemble method which uses multiple models of the same algorithm. It reduces variance by building more advanced models of data sets. It takes different subsets from our training data sets randomly and then uses bootstrap as a subroutine of its learners. Bagging trains the learning models and then take the votes on their output. Thus, by reducing variance, it can reduce the overfitting problem leading to higher stability.

VI. Conclusion

Machine learning classification has received a great attention in the past days. Some of the popular classification algorithms were used in our study. The study was conducted among 858 patients having 36 different attributes. The performances were measured in respect of accuracy, precision, recall and f-measure. The algorithm with the best performance was selected and the ensemble methods were carried out on them to improve the overall accuracy. This accuracy can further be improved with nature inspired optimization algorithms and therefore will remain as a possible future extension for our work. Also, we could try out other non-ensemble based classification models to reach higher accuracy. Finally, this study turns out to be a very crucial one considering its health and social impact and research on this arena should be kept going on in a continued manner.

References

[1] "Cancer," Available:http://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/, [Online; accessed 29-August-2018].

[2] "About cervical cancer," Available:https://www.cancer.org/cancer/cervical-cancer/about/what-is-cervical-cancer.html, [Online; accessed 29-August-2018].

[3] "Cervical cancer," https://www.womenworking.com/cervical-cancer-5-signs-never-ignore/, online; accessed: 2018-09-03.

[4] 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.

[5] H. K. Fatlawi, "Enhanced classification model for cervical cancer dataset based on cost sensitive classifier," *Int. J. Comput. Tech*, vol. 4, pp. 115–120, 2007.

[6] K. Barker, D. Berry, and C. Rainwater, "Classification of cervical cancer dataset."

[7] 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.

[8] R. Sawhney, P. Mathur, and R. Shankar, "A firefly algorithm based wrapper-penalty feature selection method for cancer diagnosis," in *International Conference on Computational Science and Its Applications*. Springer, 2018, pp. 438–449.

[9] M. F. Unlersen, K. Sabanci, and M. Özcan, "Determining cervical cancer possibility by using machine learning methods."

[10] Z. Ceylan and E. Pekel, "Comparison of multi-label classification methods for prediagnosis of cervical cancer," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 5, no. 4, pp. 232–236, 2017.

[11] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to the smo algorithm for svm regression," *IEEE transactions on neural networks*, vol. 11, no. 5, pp. 1188–1193, 2000.

[12] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to platt's smo algorithm for svm classifier design," *Neural computation*, vol. 13, no. 3, pp. 637–649, 2001.

[13] I. Syarif, E. Zaluska, A. Prugel-Bennett, and G. Wills, "Application of bagging, boosting and stacking to intrusion detection," in *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, 2012, pp. 593–602.

[14] Y. J. Huang, R. Powers, and G. T. Montelione, "Protein nmr recall, precision, and f-measure scores (rpf scores): structure quality assessment measures based on information retrieval statistics," *Journal of the American Chemical Society*, vol. 127, no. 6, pp. 1665–1674, 2005.