Cervical Cancer Prediction through Different Screening Methods using Data Mining

Talha Mahboob Alam[1], Muhammad Milhan Afzal Khan[2], Muhammad Atif Iqbal[3], Abdul Wahab[4], Mubbashar Mushtaq[5]
Computer Science and Engineering Department, University of Engineering and Technology Lahore, Pakistan[1,2,3,5]
School of Systems and Technology, University of Management and Technology Lahore, Pakistan[4]

***Abstract*—Cervical cancer remains an important reason of deaths worldwide because effective access to cervical screening methods is a big challenge. Data mining techniques including decision tree algorithms are used in biomedical research for predictive analysis. The imbalanced dataset was obtained from the dataset archive belongs to the University of California, Irvine. Synthetic Minority Oversampling Technique (SMOTE) has been used to balance the dataset in which the number of instances has increased. The dataset consists of patient age, number of pregnancies, contraceptives usage, smoking patterns and chronological records of sexually transmitted diseases (STDs). Microsoft azure machine learning tool was used for simulation of results. This paper mainly focuses on cervical cancer prediction through different screening methods using data mining techniques like Boosted decision tree, decision forest and decision jungle algorithms as well performance evaluation has done on the basis of AUROC (Area under Receiver operating characteristic) curve, accuracy, specificity and sensitivity. 10-fold cross-validation method was utilized to authenticate the results and Boosted decision tree has given the best results. Boosted decision tree provided very high prediction with 0.978 on AUROC curve while Hinslemann screening method has used. The results obtained by other classifiers were significantly worse than boosted decision tree.**

***Keywords*—Boosted decision tree; cervical cancer; data mining; dcision trees; decision forest; decision jungle; screening methods**

I. Introduction

Cancer is a dangerous disease in which group of abnormal cells develops hysterically by avoiding the usual rules of cell division. Development of cancer takes place when normal cells in a particular portion of the body begin to grow out of control [1]. Each year around 8.2 million people die from cancer which is 13% of total deaths worldwide. In 2017, only 26% of under developing countries reported having screening services available for public. In 90% developed countries treatment services are available compared to less than 26% of low income countries. The expected cancer incidences will reach up to 22 million in 2030 [2, 3]. Millions of early deaths among women is due to lung and breast cancer but cervical cancer is most dangerous because it is only diagnosed in females. Woman’s reproductive system consists of cervix, uterus, vagina and the ovaries. Cervix is the opening to the uterus from the vagina where cervical cancer occurs [4]. Sexually transmitted human papillomavirus (HPV) is the important cause of cervical cancer [5-8]. Cervical cancer occurrence is abundant in low and middle income countries [9]. The important task of cervical cancer is screening. An ideal screening test is the one that is least incursive, easy to achieve, acceptable to subject, cheap and effective in diagnosing the disease process in its early incursive stage when the treatment is easy for ailment. There are four screening methods including cervical cytology also called Pap smear test, biopsy, Schiller and Hinslemann [10]. Cytology screening method is a microscopic analysis of cells scratched from the cervix and is used to detect cancerous or pre-cancerous conditions of the cervix [11]. Biopsy method is a surgical process which includes finding of a living tissue sample for performing diagnosis [12]. The solution of iodine has applied for visual inspection of cervix known as Hinslemann test. Lugol's iodine is used for visual inspection of cervix after smearing Lugol's iodine detection rate of doubtful region over the cervix, this is also known as Schiller test [13].

The size of data is increasing gradually. Expansive, complex and useful datasets have now expanded in all the different fields of science, business and especially in healthcare domain. With these larger data sets, the capacity to mine beneficial hidden knowledge in these huge volume of data is gradually significant in today’s economical world. The method of applying novel techniques for discovering knowledge from data is called data mining [14]. Medical data consists of information regarding patients and symptoms with respect to specific disease. The volume of such type of data is expanded quickly. By utilizing the traditional techniques, it is exceptionally difficult to separate the important information from raw medical data. Due to growth in statistics, mathematics and other domains, it is now possible to extract the meaningful information from raw data. Data mining is helpful where large collections of healthcare data are available [15]. Several data mining techniques like support vector machine (SVM), kernel learning methods as well as clustering techniques were used in healthcare [16]. With the rise of computing methods for disease prediction, WHO and other international organizations are working together for effective screening method to detect the cervical cancer. These initiatives are raising public awareness for effective screening methods for cervical cancer but over the time all these measures have proved to be ineffective because the number of parameters for screening of cervical cancer are still debatable [4, 8, 10]. The methods and techniques have been used for screening of cervical cancer are limited to small number of parameters. The available literature for screening of cervical cancer explores mainly Papanicolaou (Pap) smear test [17], hormonal status, FIGO stage [18] and cervical intraepithelial neoplasia (CIN) [19] but only single parameter was used for screening prediction of cervical cancer. The available data mining techniques using large number of parameters [20-23] were not given effective results. A comparison of studies for screening prediction of cervical cancer along with approaches has presented in Table 1. It was not found effective results in screening prediction of cervical cancer while using huge number of parameters with the help of data mining techniques. As the current techniques are not sufficient, it is necessary to explore the all parameters or symptoms for screening prediction of cervical cancer. Decision tree methods have been used to predict cervical cancer but the demographic and medical attributes were different in previous studies. The aim of this study was to predict the cervical cancer, based on the demographic information, tumor related parameters, sexually transmitted diseases (STD) related parameters and important medical records.

TABLE I. COMPARISON OF EXISTING TECHNIQUES

Reference	**Data Set**			**Technique**	**Results**
	Repository	**Attributes**	**Instances**		
[20]	Universitario de Caracas Hospital patients	28	858	Hybrid method using deep learning	AUC = 0.6875
[24]	NCBI	61	160	CART Algorithm	Accuracy = 83.87%
[25]	Chung Shan Medical University Hospital Tumor Registry	38	75	Naïve Bayes	Accuracy =78.93 %
				SVM	Accuracy =78.67 %
				Random Forest Tree	Accuracy =80.18 %
[21]	Bucheon St Mary’s Hospital, Republic of Korea	15	731	SVM	Accuracy =74.41%
[17]	Chung Shan Medical University Hospital Tumor Registry	12	168	MARS	Accuracy =86.00%
[18]	State Hospital in Rzeszow	10	107	GEP	AUROC=0.72
				MLP	AUROC=0.67
				PNN	AUROC=0.56
				RBFNN	AUROC=0.48
[26]	Universitario de Caracas Hospital patients	18	858	Transfer Learning with Partial observability	RMSE=35.11
[27]	Clinical data from patients treated surgically in 1998–2001.	23	102	PNN	AUROC=0.818
				MLP	AUROC=0.659
				GEP	AUROC=0.651
				SVM	AUROC=0.478
				LRA	AUROC=0.559
				RBFNN	AUROC=0.640
				k-Means	AUROC=0.406

II. Related Work

Kelwin Fernandes et al. [20] presented an automated method for predicting the effect of the patient biopsy for the diagnosis of cervical cancer by using medical history of patients. Their technique allows a joint and fully supervised optimization method for high dimensional reduction and classification. They discovered certain medical results from the embedding spaces and confirmed through the medical literature. R. Vidya and G. M. Nasira [24] predicted cervical cancer using random forest with K-means learning and implemented the techniques in MATLAB tool. These experiments were performed with the help of NCBI dataset to construct decision tree using classification methods. Yulia et al. [25] predicted cervical cancer using Pap smear test results. The Pap smear test results were divided into two categories: cancerous and non-cancerous patients. Three classification methods Naïve Bayes, support vector machine and random forest were used to compute the results in which random forest tree was given better results. Jimin kahng et al. [21] predicted the cervical cancer development using SVM. Weka was used to train and test the data set as well as analyze relationships between attributes. Chang et al. [17] predicted the recurrence of cervical cancer in patients using MARS (Multivariate Adaptive Regression Splines) and C5.0 algorithm. MARS powerfully estimated the relationship between a dependent variable and set of descriptive variables in a pair wise regression. C5.0 used greedy method in which a top down approach was used to build the decision tree and then trained the data with the help of significant attributes. Maciej Kusy et al. [18] presented neural networks to predict adverse events in cervical cancer patients. MLP is a type of neural network where the input signal is fed forward through a number of layers. MLP contains input layer, hidden layer and output layer. The GEP classifier delivered efficient results in the prediction of the adverse events in cervical cancer as compare to other methods. Kelwin Fernandes et al. [26] used transfer learning technique for cervical cancer screening. Their study consists on linear predictive models. Positive results were obtained in most experiments as compared to other methods. Bogdan Obrzut et al. [27] utilized computational intelligence methods for prediction for cervical cancer patients. The probabilistic neural network (PNN) was a very efficient method for predicting overall survival in cervical cancer patients treated with radical hysterectomy.

III. Methodology

Our methodology consists of three main steps; the first step is data set selection. The second step includes preprocessing in which the original data is prepared for classification. The last step contains building effective classification based model for prediction.

A. Dataset

Publicly available dataset have been utilized [28] which was obtained from the UCI repository, in this research**.** The dataset contains 858 patients and 36 attributes which includes the patient age, number of pregnancies, contraceptives usage, smoking patterns and chronological records of sexually transmitted diseases (STDs).

B. Data Preprocessing

Data mining fundamentally depends on the quality of data. Raw data generally vulnerable to noisy data, missing values, outliers and inconsistency. So, it is vital for selected data to be processed before being mined. Preprocessing the data is an essential step to enhance data efficiency. Data preprocessing is one of the most vital data mining step which deals with data preparation and transformation of the dataset which make knowledge discovery more efficient. There are following steps which were used to preprocess data in this study for the experiments.

Step 1: Ignoring some instances and attributes which makes the data consistent because of high ratio of missing values. This method is very effective because there were several instances and attributes with missing values in the dataset which has been used. Some attributes in this dataset like STDs: Time since first diagnosis and STDs: Time since last diagnosis, in which more than 80% data was missing so these attributes were deleted. Two attributes STDs:cervical condylomatosis and STDs:AIDS has constant value so these were also deleted.

Step 2: There were many attributes with missing values like number of pregnancies, hormonal conceptive etc. whereas missing values denoted in data as “?” then replace these values with median values of respective class. The median value was computed as following [29].

$$Median\colon X = Sort(x), Median = X_{\frac{n}{2}}(Half\ below, Half\ above)$$

Step 3: The other important task was outlier detection in data. An outlier is a data object that deviates significantly from the rest of the objects. In this study, two attributes like age and number of partners contains outliers. To solve this issue defining lower and upper threshold limits, these outliers were replaced with median value.

Step 4: Normalization is scaling technique of data preprocessing. There were several methods of normalization i.e. Min-Max, Z-score and decimal scaling normalization [30]. Decimal scaling normalization was applied by using following method [31].

$V^i = v/10^j V^i$, V and j denotes the scaled values, range of values and smallest integer respectively.

In this study, all integer values of all attributes like age, hormonal conceptive etc. are scaled between [0-10] and Boolean attributes like smokes, HPV,STD etc. are scaled [0,1].

Step 5: After data cleaning, cervical cancer data set consists of 734 instances and 32 attributes. This data is imbalanced because only 70 instances are cancerous and 663 are non-cancerous diagnosed patients. To overcome this problem of imbalanced data, Synthetic Minority Oversampling Technique (SMOTE) has been used. This is a statistical method for increasing the number of instances in dataset in a balanced way. The module works by producing new instances from existing minority cases that supplied as input. By using SMOTE, majority instances do not change. The new instances are not just copies of existing minority classes because the algorithm takes samples of the feature space for each target class and its nearest neighbors which generate new instances that associate the features of the target class. This method makes the samples more generic [32]. x_i is a minority class and searches the nearest neighbors and one neighbor is randomly selected as x' then random numbers between [0,1] $\boldsymbol{\partial}$ selected. The new sample x_{new} was created as:

$$x_{new} = x_i + (x' - x_i) \times \boldsymbol{\partial}$$

SMOTE outperforms random oversampling method because it also avoids over fitting problem [33]. Using SMOTE function the total instances have increased. After SMOTE, minority class has oversampled from 70 to 563 instances.

C. Classification Models

A supervised method for classification is decision trees, which is very popular because most of biomedical data mining tasks have already used decision trees for efficient prediction [18]. Three decision tree methods were used in this study as follows.

1) Boosted decision tree: The transformation of a weakened classifier to a vigorous or strong classifier is the key role of boosting. A weak classifier is generally a poor performance prediction model which leads to low accuracy due to high misclassification rate. Boosted method works perfect when majority vote of all weak learners for each prediction combines in such way that final prediction results are effective. Each iteration for a weak learner is added in base learner which trained with respect to the error of the whole ensemble. When weak learner is added iteratively in an ensemble then it delivers the precise classification. A learning method consecutively tries new models to provide an extra accuracy of the class variable which leads to gradient boosting. The negative gradient of the loss function is correlated with each new model which tends to minimize the error. Friedman [34] presented a complete detail associated with boosted decision tree.

Step 1: $h_m(x)$ fit a decision tree to pseudo residuals. J_m Represents the number of leaves and input space divided into disjoint regions R$_{1\,m}$… Rj_mm which predicts a constant value in each region. The output can be written as:

$$h_m(x) = \sum_{J=1}^{J_m} B_{jm}\, 1R_{jm}(x)$$

B_{jm} Denotes the predicted value in R_{jm} region.

Step 2: B_{jm} has multiplied with γ_m which deceases the error rate by minimizing the loss function the value of model is updated $F_m(x)$.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x), \quad \gamma_m = arg_{\gamma^{min}} \sum_{i=1}^{n} L(y_{i,}\, F_{m-1}(x_i) + \gamma^{hm}(x_i))$$

Step 3: when the new updated value has determined then previous value is discarded. The new function is written as:

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm}\, 1R_{jm}(x), \quad \gamma_{jm} = arg_{\gamma^{min}} \sum_{X_I € R_{jm}} L\left(y_{i,F_{m-1}}(x_i) + \gamma\right)$$

Terminals nodes or leaves are denoted by J in the tree. The accuracy of boosted decision tree will improve if number of leaves and size of tree also increases but over fitting problem and longer processing time may occur.

2) Decision forest: The other algorithm to perform classification by utilizing ensemble learning method is known as decision forest. Ensemble methods are generalized rather than depend on a single model. A generalized model generates multiple associated models and merging them which gives better results. Mostly, ensemble models offer efficient accuracy as compared to single decision tree. Decision forest differs from random forest method, in random forest method the individual decision trees might only use some randomized portion of the data or features. There were many methods to ensemble decision trees but voting is one of the effective method for making results in an ensemble model [35]. Decision forest works by constructing multiple decision trees and then voting on the most popular output class. By utilizing the whole data set and different starting points, set of classification trees are constructed. Decision forest outputs non-normalized frequency of histograms of labels for each decision tree. Probabilities of each label is determined by aggregation method which sums the histograms then normalizes the results. Final decision of the ensemble is based on trees in which high prediction confidence depends on high weight. Criminisi [36] presented a complete detail associated with decision forest.

Step 1: Forest training is done by optimizing the parameters of the weak learner at each split node j and θ denotes the parent set and split parameters.

$$\theta_j = argmax_{\theta € T_i} I(S_j, \theta)$$

Step 2: The objective function or loss function denoted as I which takes the value of information gain. $H(s_j)$ Described as Entropy of example set parent node, $\frac{|S^{i_j}|}{|S_j|}$ denotes the weighting left/right children and $H(S^{i_j})$ represents entropy of example sat child nodes.

$$I(S_j, \theta) = H(s_j) - \sum_{T € [L,R]} \frac{|S^{i_j}|}{|S_j|} H(S^{i_j})$$

Step 3: The entropy of generic set of training points were denoted by S and $p(c)$ represents labels of normalized empirical histogram resultant to the training points in S.

$$H(S) = -\sum_{c € C} P(c) log\, p(c)$$

This method contrasts from random forest method like some random features of data may only use by decision tree instead of complete features.

3) Decision jungles: A large number of applications was developed by using decision forests and trees in data science but these methods have some limitations like while given large amount of data the number of nodes in decision trees will develop exponentially with depth. Decision jungles method compares two new node merging algorithms that jointly optimize both the features and the structure of the directed acyclic graph (DAGs) powerfully. DAGs have same structure as decision trees except the nodes can have multiple parents. Node splitting and node merging is determined by objective function and entropies of weighted sum at leaves. The training of DAGs is done level by level by combining objective function over both structure of DAGs and split function. At each level, the algorithm jointly learns the features and branching structure of the nodes. This is done by minimizing an objective function defined over the predictions. Decision jungles require radically less memory while considerably improved generalization. Shotton [37] presented a comprehensive detail related to decision jungles.

Step 1: Set of parent nodes, and a set of child nodes were denoted byN_p andN_c. θ_iDenotes the parameters of split feature function for parent node *i*€N_p and S_i denotes the set of labeled that reach node *i*. The set of instances that reach any child node *j*€N_c is.

$S_j(\{\theta_i\}.\{l_i\}.\{r_i\})$=[$\cup_{i€N_p s.t.t_{i=j}} S_i^L\ (\theta_i)$] ∪ [$\cup_{i€N_p s.t.t_{i=j}} S_i^R\ (\theta_i)$]

Step 2: The objective function E related with the current level of the DAG is a function of $\{S_j\}$*j*€N_c. The difficulty of learning the parameters of the decision DAG as a joint minimization of the objective over the split parameters $\{\theta_i\}$ and the child assignments $\{l_i\}.\{r_i\}$ were resolved. The task of learning the current level of a DAG can be written as:

$\underset{\{\theta_i\}.\{l_i\}.\{r_i\}}{min} E(\{\theta_i\}.\{l_i\}.\{r_i\}$

Step 3: The information gain objective needs the minimization of the total weighted entropy of instances, defined as:

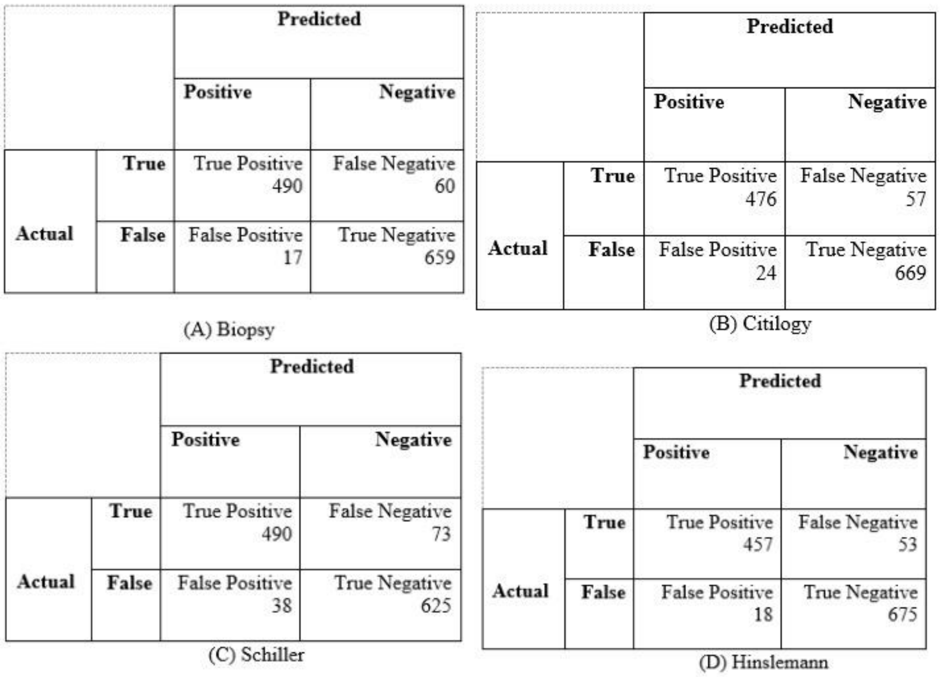
$E(\{\theta_i\}.\{l_i\}.\{r_i\})$=$\sum_{j€N_c}$ ∣S_j∣ H (S_j)

$E(\{\theta_i\}.\{l_i\}.\{r_i\})$ Presents features and branches for all parent nodes i, $\sum_{j€N_c}$ ∣S_j∣ presents sum over child nodes and number of examples at j, H (S_j) denotes entropy of examples that reach child node j.

Step 4: To solve the minimization problem cluster search method was used which substitutes among optimizing the branching variables and the split parameters but optimizes the branching variables more globally.

IV. Results and Discussion

In this study numerous methods have been examined and three methods that have the best performances has been presented. 10 fold cross validation method was used in the evaluation of the proposed methods. Cross validation method was used because it uses the entire training dataset for both training and evaluation, instead of some portion [38]. Among 858 patients, 124 patients have huge number of missing values due to privacy concerns and the remaining 734 were considered. Using SMOTE method, imbalanced dataset problem was overcome and instances were increased. The new balanced dataset consists of 32 attributes and 1226 patients in which cancer patients were 563 and non-cancer patients were 663 as shown in Fig. 1 of confusion matrix. The median value of patients' age was 26 years (range, 13-84). The median number of sex partners was 2 (range, 1–10). The median of first sexual intercourse age was 17 (range, 10-32) and median of number of pregnancies was 2 (range, 0-10).



		Predicted	
		Positive	Negative
Actual	True	True Positive 490	False Negative 60
	False	False Positive 17	True Negative 659

(A) Biopsy

		Predicted	
		Positive	Negative
Actual	True	True Positive 476	False Negative 57
	False	False Positive 24	True Negative 669

(B) Citilogy

		Predicted	
		Positive	Negative
Actual	True	True Positive 490	False Negative 73
	False	False Positive 38	True Negative 625

(C) Schiller

		Predicted	
		Positive	Negative
Actual	True	True Positive 457	False Negative 53
	False	False Positive 18	True Negative 675

(D) Hinslemann

Fig. 1. Confusion Matrix Obtained by using Different Models.

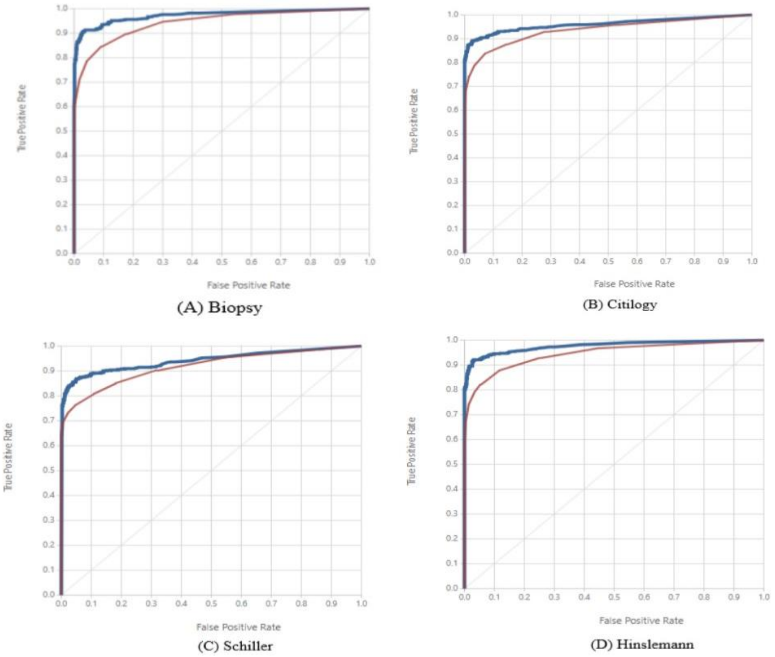
There were four screening methods (target attributes) in the data set labeled as biopsy, cytology, Schiller and hinslemann. These four screening methods have been used to diagnose cancer and each screening method was trained with same dataset but individually. Boosted decision tree outperformed all other methods as shown in Table 2. The hinslemann screening method also outperformed other methods as AUROC curve performance is 0.978 which was slightly higher from Biopsy but significant higher from cytology and Schiller. The AUROC curve has also given better results on boosted decision tree i.e. 0.974 on biopsy, 0.959 on cytology and 0.943 on Schiller target attribute. The complete performance of proposed models has given in Fig. 3 and performance on AUROC curve has shown in Fig. 2.

Boosted decision tree, decision forest and decision jungle algorithms were used to determine the prediction ability of tested models by computing the accuracy, sensitivity, specificity and AUROC curve. AUROC curve is a best measure to evaluate the performance of classification models [39-42]. The AUROC curve performance of proposed models has shown in Fig. 2.

The AUROC curve is a summary measure of performance that indicates whether on average a true positive is ranked higher than a false positive rate or not. AUROC curve was also used for evaluation of different techniques [18, 27] in biomedical data mining.

TABLE II. AUROC CURVE OBTAINED BY THE ML TECHNIQUES ON THE RISK PREDICTION TASK WITH MULTIPLE SCREENING METHODS: BIOPSY, CYTOLOGY, SCHILLER AND HINSELMANN. PERFORMANCE WAS ALSO EVALUATED IN TERMS OF ACCURACY, SENSITIVITY AND SPECIFICITY

Method	Screening Method (Target Attribute)	Accuracy	Sensitivity	Specificity	AUROC
Boosted Decision Tree	**Biopsy**	0.937	0.891	0.974	**0.974**
Decision Forest		0.880	0.785	0.957	0.943
Decision Jungle		0.863	0.733	0.968	0.929
Boosted Decision Tree	**Cytology**	0.934	0.893	0.965	**0.959**
Decision Forest		0.888	0.790	0.963	0.935
Decision Jungle		0.879	0.735	0.989	0.929
Boosted Decision Tree	**Schiller**	0.909	0.870	0.942	**0.943**
Decision Forest		0.865	0.766	0.948	0.918
Decision Jungle		0.863	0.726	0.978	0.908
Boosted Decision Tree	**Hinslemann**	0.941	0.896	0.974	**0.978**
Decision Forest		0.892	0.793	0.965	0.945
Decision Jungle		0.879	0.730	0.991	0.934



(A) Biopsy

(B) Citilogy

(C) Schiller

(D) Hinslemann

Fig. 2. Comparison of Area under Receiver operating Characteristic (AUROC) Curve between Boosted Decision Tree (Blue Line) and Decision Forest (Red Line) as these Model Gives Best Results. Plots are Shown for the Models with Threshold=5.

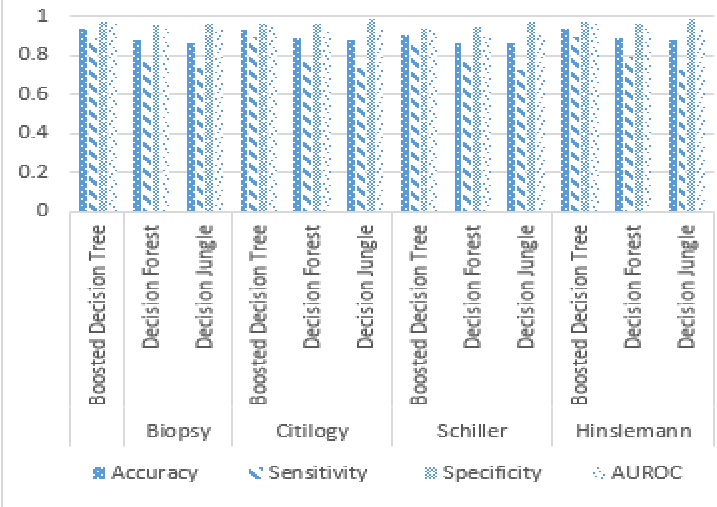


Fig. 3. The Results in Terms of Accuracy, Sensitivity, Specificity and AUROC Curve in the Prediction of Cervical Cancer.

There are 50% of cervical cancer identification in females age (35–54) and around 20% diagnosed more than 65 years old as well as around 15% of between the age of (20 – 30). Median age for diagnosis in cervical cancer is 48 years. Cervical cancer is significantly unusual in females, younger than age 20. In any case, several young females end up infected with different sorts of human papilloma infection (HPV), which can expand their danger of getting cervical cancer in future. Young females with early abnormal changes who don't have regular checkup are at high risk of cervical cancer when they reach at the age of 40 [43-45]. The main risk factor for cervical cancer growth is HPV. Sexual relation with infected persons is another risk factor for HPV. Different parameters with respect to sexual relation like sexual relation with multiple persons are also danger factor for females which leads to cervical cancer. Sexually dynamic females (sexually obsessed) have never been in danger of cervical cancer as compare to those who have multiple sexual partners [46,47]. Smoking is related with a higher risk for precancerous fluctuations in the cervix and development to invasive cervical cancer, particularly for women infected with HPV. Women with weak immune system are more prone to getting HPV [48].

This study was exploited late advancements in statistical learning for handling the high dimensional data with numerous features. Other promising areas of research in these conditions were also used ensemble learning methods [49]. Classification algorithms have a wide range of applications which used decision trees other than biomedical domain. Astronomical objects detection [50], fraud detection in banking [51] and financial failure prediction [52] were also utilized decision trees for classification. There were several classification algorithms presented in literature but decision trees were generally utilized because of its simplicity of implementation and ease to understand as compared to other classification methods. Recently, high dimensional classification problems have been abundant due to substantial developments in technology [53]. Generally the problem of large dimensional data modelling has been solved by variable reduction methods in the preprocessing and in the post-processing stage. Several data mining methods like artificial neural networks, support vector machines and k-nearest neighbor method were also used to resolve the high dimensional classification problem [54]. In this study, high dimensional classification problem was resolved by using decision tree methods because only those attributes were considered which showed highest relevance with the screening method (target class). The Hinslemann screening method showed high performance because Hinslemann is also traditional method of screening of cervical cancer which is effective [55-57]. The performance of biopsy screening method was slightly low from Hinslemann screening method. From various studies, it was also found that biopsy screening has huge impact for cervical cancer detection [58, 59]. The use of boosted decision tree was preferred because it focused on misclassified instances and had tendency to increase accuracy. Boosting is one way to decrease the misclassification rate because inside boosting, iteration was introduced [60]. In general, this increased the degree of accuracy in classification. Since, boosted decision tree is an ensemble model in which results from various models are consolidated. The outcome acquired from ensemble model is normally higher to the outcome from any of individual model. In this study, maximum number of leaves per tree were 20 and minimum number of leaves per tree were 10. Learning rate has taken low which is 0.1 but processing time slightly increases because 100 number of tree to ensemble are constructed while boosted decision tree has used. Learning rate and number of trees are higher which leads to better performance but processing time also increased. Boosted decision tree was also used for sentiment analysis of Greek language which efficiently coped with both high dimensional and imbalanced datasets and achieves considerably enhanced then other traditional machine learning methods [61] as well as utilized for cardiovascular risk prediction [62] and risk prediction for inflammatory bowel disease [63]. Due to some limitations, decision forest was not given better results. The main limitation of the decision forests is that real time prediction is slow when a large number of trees are made. These algorithms are fast to train but quite slow to create predictions once they are trained. The accuracy may increases when the number of trees were also increased [64] but also leads slower model for prediction. In most real world applications the decision forest is fast enough but in some situations run time performance is important and other methods would be chosen. Decision forest was also used to understand protein interactions and making predictions based on all the protein domains [65]. The other applications of decision forest were prediction of different types of liver diseases including alcoholic, liver damage and liver cirrhosis [66]. Other than biomedical classification, Decision forest method was applied for academic data analysis [67] as well as classification and forecasting of chronic kidney disease [68]. Decision Jungles were used for feature selection for images with some modification to achieve efficient results with modest training time [69].

V. Conclusion

Nowadays, cervical cancer is a common disease and its screening often involves very time consuming clinical tests. In this perspective, machine learning can deliver efficient methods to speed up the diagnosis procedure. Furthermore in this research work, Data mining methods especially tree based algorithms enable sound prediction for cervical cancer patients. The imbalanced data set problem in which cancerous patients were too small as compared to non-cancerous patients has been resolved by using SMOTE method. The prediction ability of the boosted decision tree measured by the AUROC curve value which outperformed decision forest and decision jungle. The low AUROC curve value for the decision forest and decision jungle methods disqualified them as best predictive classifiers. We believe that with the growing collection of cervical cancer patient's data and the rapidly advancing methods for analyzing this data, we will begin to be able to identify best screening method for cervical cancer patients that will be informative for patient care. In future, this study can be used as a prototype to develop a healthcare system for cervical cancer patients.

References

[1] M. Hejmadi, Introduction to cancer biology: Bookboon, 2009.

[2] N. Kamil and S. Kamil, "Global cancer incidences, causes and future predictions for subcontinent region," Systematic Reviews in Pharmacy, vol. 6, p. 13, 2015.

[3] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2017," CA: a cancer journal for clinicians, vol. 67, pp. 7-30, 2017.

[4] S. Subramanian, R. Sankaranarayanan, P. O. Esmy, J. V. Thulaseedharan, R. Swaminathan, and S. Thomas, "Clinical trial to implementation: Cost and effectiveness considerations for scaling up cervical cancer screening in low-and middle-income countries," Journal of Cancer Policy, vol. 7, pp. 4-11, 2016.

[5] K. U. Petry, "HPV and cervical cancer," Scandinavian Journal of Clinical and Laboratory Investigation, vol. 74, pp. 59-62, 2014.

[6] G. Ronco, J. Dillner, K. M. Elfström, S. Tunesi, P. J. Snijders, M. Arbyn, et al., "Efficacy of HPV-based screening for prevention of invasive cervical cancer: follow-up of four European randomised controlled trials," The lancet, vol. 383, pp. 524-532, 2014.

[7] K. J. Sales, "Human papillomavirus and cervical cancer," in Cancer and Inflammation Mechanisms: Chemical, Biological, and Clinical Aspects, ed: John Wiley & Sons, 2014, pp. 165-180.

[8] W. H. Organization, "WHO guidance note: comprehensive cervical cancer prevention and control: a healthier future for girls and women," 2013.

[9] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: a cancer journal for clinicians, vol. 68, pp. 394-424, 2018.

[10] R. A. Kerkar, "Screening for cervical cancer: an overview."

[11] G. Guvenc, A. Akyuz, and C. H. Açikel, "Health belief model scale for cervical cancer and Pap smear test: psychometric testing," Journal of advanced nursing, vol. 67, pp. 428-437, 2011.

[12] M. T. Galgano, P. E. Castle, K. A. Atkins, W. K. Brix, S. R. Nassau, and M. H. Stoler, "Using biomarkers as objective standards in the diagnosis of cervical biopsies," The American journal of surgical pathology, vol. 34, p. 1077, 2010.

[13] H. Ramaraju, Y. Nagaveni, and A. Khazi, "Use of Schiller's test versus Pap smear to increase detection rate of cervical dysplasias," International Journal of Reproduction, Contraception, Obstetrics and Gynecology, vol. 5, pp. 1446-1450, 2017.

[14] N. Jothi and W. Husain, "Data mining in healthcare–a review," Procedia Computer Science, vol. 72, pp. 306-313, 2015.

[15] P. Ahmad, S. Qamar, and S. Q. A. Rizvi, "Techniques of data mining in healthcare: a review," International Journal of Computer Applications, vol. 120, 2015.

[16] T. M. Alam and M. J. Awan, "Domain Analysis of Information ExtractionTechniques," INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY SCIENCES AND ENGINEERING, vol. 9, pp. 1-9, 2018.

[17] C.-C. Chang, S.-L. Cheng, C.-J. Lu, and K.-H. Liao, "Prediction of Recurrence in Patients with Cervical Cancer Using MARS and Classification," International Journal of Machine Learning and Computing, vol. 3, p. 75, 2013.

[18] M. Kusy, B. Obrzut, and J. Kluska, "Application of gene expression programming and neural networks to predict adverse events of radical hysterectomy in cervical cancer patients," Medical & biological engineering & computing, vol. 51, pp. 1357-1365, 2013.

[19] J. M. Yamal, M. Guillaud, E. N. Atkinson, M. Follen, C. MacAulay, S. B. Cantor, et al., "Prediction using hierarchical data: Applications for automated detection of cervical cancer," Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 8, pp. 65-74, 2015.

[20] K. Fernandes, D. Chicco, J. S. Cardoso, and J. Fernandes, "Supervised deep learning embeddings for the prediction of cervical cancer diagnosis," PeerJ Computer Science, vol. 4, p. e154, 2018.

[21] J. Kahng, E.-H. Kim, H.-G. Kim, and W. Lee, "Development of a cervical cancer progress prediction tool for human papillomavirus-positive Koreans: A support vector machine-based approach," Journal of International Medical Research, vol. 43, pp. 518-525, 2015.

[22] Y. Al-Wesabi, A. Choudhury, and D. Won, "Classification of cervical cancer dataset," in Avishek Choudhury, Wesabi, Classification of Cervical Cancer Dataset, Proceedings of the 2018 IISE Annual Conference, Orlando, 2018, pp. 1456-1461.

[23] Y. Qi, Z. Zhao, L. Zhang, H. Liu, and K. Lei, "A Classification Diagnosis of Cervical Cancer Medical Data Based on Various Artificial Neural Networks," in 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018), 2018.

[24] R. Vidya and G. Nasira, "Prediction of cervical cancer using hybrid induction technique: A solution for human hereditary disease patterns," Indian Journal of Science and Technology, vol. 9, 2016.

[25] Y. E. Kurniawati, A. E. Permanasari, and S. Fauziati, "Comparative study on data mining classification methods for cervical cancer prediction using pap smear results," in Biomedical Engineering (IBIOMED), International Conference on, 2016, pp. 1-5.

[26] 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, 2017, pp. 243-250.

[27] B. Obrzut, M. Kusy, A. Semczuk, M. Obrzut, and J. Kluska, "Prediction of 5–year overall survival in cervical cancer patients treated with radical hysterectomy using computational intelligence methods," BMC cancer, vol. 17, p. 840, 2017.

[28] U. M. L. Repository, "Cervical cancer (Risk Factors) Data Set," 2017.

[29] R. F. Woolson and W. R. Clarke, Statistical methods for the analysis of biomedical data vol. 371: John Wiley & Sons, 2011.

[30] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Data preprocessing for supervised leaning," International Journal of Computer Science, vol. 1, pp. 111-117, 2006.

[31] S. Patro and K. K. Sahu, "Normalization: A preprocessing stage," arXiv preprint arXiv:1503.06462, 2015.

[32] 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, 2002.

[33] Z. Zheng, Y. Cai, and Y. Li, "Oversampling method for imbalanced classification," Computing and Informatics, vol. 34, pp. 1017-1037, 2016.

[34] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of statistics, pp. 1189-1232, 2001.

[35] L. Rokach, "Decision forest: Twenty years of research," Information Fusion, vol. 27, pp. 111-125, 2016.

[36] A. Criminisi and J. Shotton, Decision forests for computer vision and medical image analysis: Springer Science & Business Media, 2013.

[37] J. Shotton, T. Sharp, P. Kohli, S. Nowozin, J. Winn, and A. Criminisi, "Decision jungles: Compact and rich models for classification," in Advances in Neural Information Processing Systems, 2013, pp. 234-242.

[38] D. Krstajic, L. J. Buturovic, D. E. Leahy, and S. Thomas, "Cross-validation pitfalls when selecting and assessing regression and classification models," Journal of cheminformatics, vol. 6, p. 10, 2014.

[39] F. Garrido, W. Verbeke, and C. Bravo, "A Robust profit measure for binary classification model evaluation," Expert Systems with Applications, vol. 92, pp. 154-160, 2018.

[40] M. Vihinen, "How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis," in BMC genomics, 2012, p. S2.

[41] D. J. Hand, "Measuring classifier performance: a coherent alternative to the area under the ROC curve," Machine learning, vol. 77, pp. 103-123, 2009.

[42] K. Hajian-Tilaki, "Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation," Caspian journal of internal medicine, vol. 4, p. 627, 2013.

[43] C. Sun, A. J. Brown, A. Jhingran, M. Frumovitz, L. Ramondetta, and D. C. Bodurka, "Patient preferences for side effects associated with cervical cancer treatment," International journal of gynecological cancer: official journal of the International Gynecological Cancer Society, vol. 24, p. 1077, 2014.

[44] I.C.o.E.S.o.C. Cancer, "Cervical cancer and hormonal contraceptives: collaborative reanalysis of individual data for 16 573 women with cervical cancer and 35 509 women without cervical cancer from 24 epidemiological studies," The Lancet, vol. 370, pp. 1609-1621, 2007.

[45] G. Danaei, S. Vander Hoorn, A. D. Lopez, C. J. Murray, M. Ezzati, and C. R. A. c. group, "Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors," The Lancet, vol. 366, pp. 1784-1793, 2005.

[46] F. X. Bosch, A. Lorincz, N. Muñoz, C. Meijer, and K. V. Shah, "The causal relation between human papillomavirus and cervical cancer," Journal of clinical pathology, vol. 55, pp. 244-265, 2002.

[47] S. de Sanjosé, M. Brotons, and M. A. Pavón, "The natural history of human papillomavirus infection," Best practice & research Clinical obstetrics & gynaecology, vol. 47, pp. 2-13, 2018.

[48] E. Mazarico, R. Gómez, L. Guirado, N. Lorente, and E. Gonzalez-Bosquet, "Relationship between smoking, HPV infection, and risk of cervical cancer," Eur. J. Gynaec. Oncol.-ISSN, vol. 392, p. 2936, 2015.

[49] L. Rokach, "Ensemble-based classifiers," Artificial Intelligence Review, vol. 33, pp. 1-39, 2010.

[50] A. Franco-Arcega, L. Flores-Flores, and R. F. Gabbasov, "Application of decision trees for classifying astronomical objects," in Artificial Intelligence (MICAI), 2013 12th Mexican International Conference on, 2013, pp. 181-186.

[51] K. Chitra and B. Subashini, "Data mining techniques and its applications in banking sector," International Journal of Emerging Technology and Advanced Engineering, vol. 3, pp. 219-226, 2013.

[52] N. Öcal, M. K. Ercan, and E. Kadıoğlu, "Predicting Financial Failure Using Decision Tree Algorithms: An Empirical Test on the Manufacturing Industry at Borsa Istanbul," International Journal of Economics and Finance, vol. 7, 2015.

[53] V. Pappu and P. M. Pardalos, "High-Dimensional Data Classification," in Clusters, Orders, and Trees: Methods and Applications: In Honor of Boris Mirkin's 70th Birthday, F. Aleskerov, B. Goldengorin, and P. M. Pardalos, Eds., ed New York, NY: Springer New York, 2014, pp. 119-150.

[54] M. Zekić-Sušac, S. Pfeifer, and N. Šarlija, "A Comparison of Machine Learning Methods in a High-Dimensional Classification Problem," Business Systems Research Journal, vol. 5, pp. 82-96, 2014.

[55] Y. Eraso, "Migrating techniques, multiplying diagnoses: the contribution of Argentina and Brazil to early'detection policy'in cervical cancer," História, Ciências, Saúde-Manguinhos, vol. 17, pp. 33-51, 2010.

[56] M. Aref - Adib and T. Freeman - Wang, "Cervical cancer prevention and screening: the role of human papillomavirus testing," The Obstetrician & Gynaecologist, vol. 18, pp. 251-263, 2016.

[57] I. Löwy, "Cancer, women, and public health: the history of screening for cervical cancer," História, Ciências, Saúde-Manguinhos, vol. 17, pp. 53-67, 2010.

[58] P. Ghosh, G. Gandhi, P. Kochhar, V. Zutshi, and S. Batra, "Visual inspection of cervix with Lugol's iodine for early detection of premalignant & malignant lesions of cervix," The Indian journal of medical research, vol. 136, p. 265, 2012.

[59] K. Petry, J. Horn, A. Luyten, and R. Mikolajczyk, "Punch biopsies shorten time to clearance of high-risk human papillomavirus infections of the uterine cervix," BMC cancer, vol. 18, p. 318, 2018.

[60] A. Niculescu-Mizil and R. Caruana, "Obtaining Calibrated Probabilities from Boosting.".

[61] V. Athanasiou and M. Maragoudakis, "A novel, gradient boosting framework for sentiment analysis in languages where NLP resources are not plentiful: a case study for modern greek," Algorithms, vol. 10, p. 34, 2017.

[62] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," PloS one, vol. 12, p. e0174944, 2017.

[63] Z. Wei, W. Wang, J. Bradfield, J. Li, C. Cardinale, E. Frackelton, et al., "Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease," The American Journal of Human Genetics, vol. 92, pp. 1008-1012, 2013.

[64] S. Fong, W. Song, R. Wong, C. Bhatt, and D. Korzun, "Framework of Temporal Data Stream Mining by Using Incrementally Optimized Very Fast Decision Forest," in Internet of Things and Big Data Analytics Toward Next-Generation Intelligence, ed: Springer, 2018, pp. 483-502.

[65] X.-W. Chen and M. Liu, "Prediction of protein–protein interactions using random decision forest framework," Bioinformatics, vol. 21, pp. 4394-4400, 2005.

[66] A. Singh and B. Pandey, "A New Intelligent Medical Decision Support System Based on Enhanced Hierarchical Clustering and Random Decision Forest for the Classification of Alcoholic Liver Damage, Primary Hepatoma, Liver Cirrhosis, and Cholelithiasis," Journal of healthcare engineering, vol. 2018, 2018.

[67] A. J. Fernández-García, L. Iribarne, A. Corral, and J. Criado, "A Comparison of Feature Selection Methods to Optimize Predictive Models Based on Decision Forest Algorithms for Academic Data Analysis," in World Conference on Information Systems and Technologies, 2018, pp. 338-347.

[68] W. Gunarathne, K. Perera, and K. Kahandawaarachchi, "Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)," in Bioinformatics and Bioengineering (BIBE), 2017 IEEE 17th International Conference on, 2017, pp. 291-296.

[69] S. Baek, K. I. Kim, and T.-K. Kim, "Deep Convolutional Decision Jungle for Image Classification," arXiv preprint arXiv:1706.02003, 2017.