Hindawi
Scientific Programming
Volume 2021, Article ID 5540024, 10 pages
https://doi.org/10.1155/2021/5540024

Research Article

Cervical Cancer Diagnosis Model Using Extreme Gradient Boosting and Bioinspired Firefly Optimization

Irfan Ullah Khan, Nida Aslam, Rawan Alshehri, Seham Alzahrani, Manal Alghamdi, Atheer Almalki, and Maryam Balabeed

Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam 31441, Saudi Arabia

Correspondence should be addressed to Nida Aslam; naslam@iau.edu.sa

Received 24 February 2021; Accepted 27 June 2021; Published 19 July 2021

Academic Editor: Antonio J. Peña

Copyright © 2021 Irfan Ullah Khan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Cervical cancer is frequently a deadly disease, common in females. However, early diagnosis of cervical cancer can reduce the mortality rate and other associated complications. Cervical cancer risk factors can aid the early diagnosis. For better diagnosis accuracy, we proposed a study for early diagnosis of cervical cancer using reduced risk feature set and three ensemble-based classification techniques, i.e., extreme Gradient Boosting (XGBoost), AdaBoost, and Random Forest (RF) along with Firefly algorithm for optimization. Synthetic Minority Oversampling Technique (SMOTE) data sampling technique was used to alleviate the data imbalance problem. Cervical cancer Risk Factors data set, containing 32 risks factor and four targets (Hinselmann, Schiller, Cytology, and Biopsy), is used in the study. The four targets are the widely used diagnosis test for cervical cancer. The effectiveness of the proposed study is evaluated in terms of accuracy, sensitivity, specificity, positive predictive accuracy (PPA), and negative predictive accuracy (NPA). Moreover, Firefly features selection technique was used to achieve better results with the reduced number of features. Experimental results reveal the significance of the proposed model and achieved the highest outcome for Hinselmann test when compared with other three diagnostic tests. Furthermore, the reduction in the number of features has enhanced the outcomes. Additionally, the performance of the proposed models is noticeable in terms of accuracy when compared with other benchmark studies for cervical cancer diagnosis using reduced risk factors data set.

1. Introduction

Cervical cancer is one of the commonly occurring types of cancer in females and mostly develops during their midlives (35 years–44 years) [1]. This type of cancer can be fatal as it does not show clear symptoms in its early stages. Symptoms usually appear in late stages, where it could have spread to other organs like bones, liver, lymph nodes, and lungs. One of the early signs of cervical cancer is when the tube that carries urine from the kidney is blocked. Other late symptoms that can appear are vaginal bleeding, pelvic pain, weight loss, and leg pain [2].

The risk factors that lead to the development of cervical cancer are hormones containing medicines, birth control pills, smoking, and the number of pregnancies. However, it is believed that human papilloma virus (HPV) is the major factor in developing cervical cancer [2]. HPV is a common sexually transmitted infection; it is usually harmless, but sometimes it may lead to cancer [3]. HPV infection becomes at a higher risk of getting cervical cancer. Furthermore, the probability of getting cervical cancer increases if one possesses more than one risk factor. As the cancer does not show signs in its early stages, regular checkups are required especially for those who have the risk factors. In the developing countries, lack of medical equipment and the cost of conducting checkups could also be a burden. With the advent and advancement of machine learning, it has become possible to find robust solutions for early diagnosis of cancer cases using data-driven approaches.

Various studies have contributed to the field of cervical cancer diagnosis using several classification techniques by

using different types of data such as clinical-based, image, and genetic-based data. In our study, we used clinical cervical risk factor data. Two similar studies were conducted by Wu and Zhou [4] and Abdoh et al. [5]; they performed the comparative analysis of two feature selection techniques, namely, recursive feature elimination (RFE) and Principal Component Analysis (PCA). The first study used Support Vector Machine (SVM), and the other study used Random Forest (RF). Both studies used the same number of features. Although the data suffered from imbalance, an oversampling was applied to the data in [4] and SMOTE was used in [5]. Both studies identified two risk factors to be removed such as time since the first and last diagnosis of STDs (sexually transmitted diseases), due to a lot of missing entries. Furthermore, the study [4] discovered that less computational cost was an advantage given by both SVM-PCA and SVM-RFE, whereas high computational cost is a limitation to the SVM model. Moreover, STDs, intrauterine device (IUD), hormonal contraceptives, and first sexual intercourse were identified as the highly relevant features [5]. Overall, the outcome of both the studies showed that using 30 features produced highest results. Furthermore, it was found that the SMOTE-RF model performed well for all targets.

Similarly, Lu et al. [6] and Karim and Neehal [7] used ensemble models to estimate the risk of cervical cancer. Both studies performed data cleaning mechanism to replace missing values. The former study used an ensemble classifier with voting strategy using a combination of a private and public data set. The private data set contains 472 records taken from Chinese hospital. The public data set was obtained from the UCI repository; 14 features were used. The private data set was collected using questionnaire. The results revealed that voting ensemble classifier produced better results when compared to Linear Regression, Decision Tree (DT), Multilayer Perceptron (MLP), SVM, and K-NN classifiers. On the other hand, Karim and Neehal study used DT, MLP, and SVM using sequential Minimal Optimization (SMO) and K nearest neighbor (KNN) techniques. Experiments showed that SMO has a better performance in terms of accuracy, precision, recall, and F-measure. Similarly, Ul-Islam et al. [8] used DT, RF, Logistic Model Tree, and ANN for cervical cancer detection. Apriori algorithm was used to identify features that strongly relate to cancer. The study found that age, number of sexual partners, hormonal contraceptives, number of pregnancies, and first sexual intercourse are significant risk factors. Results indicated that RF produced best outcome when compared to the other models.

Al-Wesabi et al. [9] conducted a comparison between different machine learning classifiers such as Gaussian Naïve Bayes (GNB), KNN, DT, LR, and SVM. The outcome of the classifiers was not satisfactory due to the data imbalance. To resolve this problem, undersampling, oversampling, and SMOTETomek were applied. Oversampling had the best result among all three methods. Moreover, a Sequential Feature Selector was applied with both forward and backward versions. Both the Sequential Forward Feature Selector (SFS) and Sequential Backward Feature Selector (SBS) enhanced the performance of the prediction with an accuracy of 95%. After selecting the common features between DT and KNN, the accuracy exceeded 97% for the DT. The results revealed that age, first sexual intercourse, number of pregnancies, smoking, hormonal contraceptives, and STDs: genital herpes were the main predictive features.

Similarly, several studies have been made using deep learning and transfer learning for cervical cancer diagnosis. Fernandes et al. [10] and Adem et al. [11] used deep learning and showed significant outcome in terms of diagnosis accuracy. The study [10] used a loss function that provides a supervised optimization of dimensionality reduction and classification models. The study indicated that it can be useful in examining records of patients if the Biopsy and perhaps other testing results are absent and are capable of classifying successfully whether they have cervical cancer or not. On the other hand, the researchers in [11] used a deep neural network model with softmax function to classify the data sets. The performance of the softmax function with stacked autoencoder was compared with the other machine learning methods (DT, KNN, SVM, Feed Forward NN, and Rotation Forest models). It was found that the softmax function with a stacked autoencoder model produced better outcome classification rate of 97.8%.

Similarly, Fernandes et al. [12] applied transfer learning with partial observability for cancer screenings. The limitation of the study was that several patients were resisting answering some questions for privacy concerns. Challenges were also faced in defining quality as there are multiple readings and it started relying on human preference. Therefore, as an alternative of an ordinal scale, a simple binary scheme was used. Nevertheless, the model performance was considerable.

Conclusively, the finding made after the above-mentioned literature is that the data set found at UCI repository had several missing values; therefore, previous studies have removed at least 2 features. Missing values were due to patient's concerns regarding their privacy. After removing 2 features due to huge missing value, SVM-PCA seemed to provide satisfactory performance. However, SMO and SMOTE-RF were amongst the best performing models. Another approach to deal with the imbalance in UCI cervical risk factor data set was using oversampling. Deep learning proved to be effective, especially where the Biopsy and possibly other screening results are absent. Age, first sexual intercourse, number of pregnancies, smoking, hormonal contraceptives, IUD, STDs, STDs: genital warts, or HPV infections were identified as the top key features. The significant outcomes made by the machine learning classifiers motivate the need for further investigation and enhancement of the outcomes for the prediction of cervical cancer.

In this study, three ensemble-based classifiers extreme Gradient Boosting, Ada Boost, and RF are used to classify cervical cancer. Cervical Cancer Risk factor data set from UCI machine learning repository was collected at "Hospital Universitario de Caracas" in Caracas, Venezuela [13]. In addition to the importance of correctly classifying cancerous and noncancerous cases, it is also essential to identify key risk factors that contribute to developing cancer. Nature-inspired Firefly feature selection and optimization algorithm

was applied. Furthermore, the Synthetic Minority Oversampling Technique (SMOTE) is used to balance the classes of the data as it suffers greatly from imbalanced problem.

The paper is organized as follows: Section 2 presents material and methods. Section 3 contains experimental setup and results. The comparison of the proposed model with the existing studies using the same dataset is discussed in Section 4. Finally, Section 5 contains the conclusion.

2. Material and Method

2.1. Dataset Description. The cervical cancer risk factors data set used in the study was collected at "Hospital Universitario de Caracas" in Caracas, Venezuela and is available on the UCI Machine Learning repository [13]. It consists of 858 records, with some missing values, as several patients did not answer some of the questions due to privacy concerns. The data set contains 32 risk factors and 4 targets, i.e., the diagnosis tests used for cervical cancer. It contains different categories of feature set such as habits, demographic information, history, and Genomic medical records. Features such as age, Dx: Cancer, Dx: CIN, Dx: HPV, and Dx features contains no missing values. Dx: CIN is a change in the walls of cervix and is commonly due to HPV infection; sometimes, it may lead to cancer if it is not treated properly. However, Dx: cancer variable is represented if the patient has other types of cancer or not. Sometimes, a patient may have more than one type of cancer. In the data set, some of the patients do not have cervical cancer, but they had the Dx: cancer value true. Therefore, it is not used as a target variable.

Table 1 presents a brief description of each feature with the type. Cervical cancer diagnosis usually requires several tests; this data contains the widely used diagnosis tests as the target. Hinselmann, Schiller, Cytology, and Biopsy are four widely used diagnosis tests for cervical cancer. Hinselmann or Colposcopy is a test that examines the inside of the vagina and cervix using a tool that magnifies the tissues to detect any anomalies [3]. Schiller is a test in which a chemical substance called iodine is applied to the cervix, where it stains healthy cells into brown color and leaves the abnormal cells uncolored, while cytology is a test that examines body cells from uterine cervix for any cancerous cells or other diseases. And Biopsy refers to the test where a small part of cervical tissue is examined under a microscope. Most Biopsy tests can make significant diagnosis.

2.2. Dataset Preprocessing. The data set suffers from a huge number of missing values; 24 features out of the 32 contained missing values. Initially, the features with the huge percentage of missing values were removed. STDs: Time since first diagnosis and STDs: Time since last diagnosis features were removed since they have 787 missing values (see Table 2), which is more than half of the data. However, the data imputation was performed for the features with fewer numbers of missing values. The most frequent value technique was used to impute the remaining missing values. Additionally, the data set also suffers from huge class imbalance. The data set target labels were imbalanced with 35 for the Hinselmann, 74 for Schiller, 44 for Cytology, and 55 Biopsy out of the 858 records as shown in Figure 1. SMOTE was used to deal with class imbalance. SMOTE works by oversampling the minority class by generating new synthetic data for minority instances based on nearest neighbors using the Euclidean Distance between data points [14]. Figure 1 shows the number of records per class labels in the data set.

2.3. Firefly Feature Selection. Dimensionality reduction is one of the effective ways to select the features that improve the performance of the supervised learning model. In the study, we adopted nature-inspired algorithm Firefly for selecting the features that better formulate the problem. Firefly was proposed by Yang [15] and was initially proposed for the optimization. Metaheuristic Firefly algorithm is inspired by fireflies' and flash lightening capability of a fly. It is a population-based optimization algorithm to find the optimal value or parameter for a target function. In this technique, each fly is pulled out by the glow intensity of the nearby flies. If the intensity of the gleam is extremely low at some point, then the attraction will be declining. Firefly used three rules; that is, (a) all the flies should be of the same gender; (b) the criteria of attractiveness depend upon the intensity of the glow; (c) target function will generate the gleam of the firefly. The flies with less glow will move towards the flies with brighter glow. The brightness can be adjusted using objective function. The same idea is implemented in the algorithm to search the optimal features that can better fit the training model. Firefly is more computationally economical and produced better outcome in feature selection when compared with other metaheuristic techniques like genetic algorithms and particle swarm optimization [16]. The time complexity of firefly is $O(n^2t)$ [17]. It uses the light intensity to select the features. Highly relevant features are represented as the features with high intensity light.

For feature selection, initially, some fireflies will be generated, and each fly will randomly assign the weights to all features. In our study, we generated 50 number of flies ($n = 50$). The dimension of the data set is 30. Furthermore, the lower bound was set to -50, while the upper bound is equal to 50. The maximum generations were 500. Additionally, α (alpha) was initially set to 0.5 and in every subsequent iteration, we used the (1) and (2) to update α (alpha) value.

$$X = \lim_{i=1 \longrightarrow 500} 1 - 10^{-(4/0.9)^{1/i}}, \tag{1}$$

$$\alpha = (1 - X) \times \alpha. \tag{2}$$

However, the gamma (γ) was set to 1. The number of features selected using Firefly for Hinselmann was 15, for Schiller 13 features, for Cytology 11 features, and 11 features for Biopsy, respectively.

2.4. Ensemble-Based Classification Methods. Three ensemble-based classification techniques such as Random Forest, Extreme Gradient Boosting, and Ada Boost were used to

Table 1: Statistical description of the data set.

Feature type	Feature name	Hinselmann		Schiller		Cytology		Biopsy	
		Mean (μ) ± Std (σ)		Mean (μ) ± Std (σ)		Mean (μ) ± Std (σ)		Mean (μ) ± Std (σ)	
		Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
Demographic	Age	26.7 ± 7.7	26.8 ± 8.5	29.6 ± 11	26.6 ± 8.2	26.2 ± 8.4	26.9 ± 8.5	28.6 ± 8.9	26.7 ± 8.5
	Smokes	0.2 ± 0.4	0.1 ± 0.4	0.2 ± 0.4	0.2 ± 0.3	0.1 ± 0.3	0.1 ± 0.4	0.2 ± 0.4	0.1 ± 0.3
	Smokes (years)	2.5 ± 7.2	1.1 ± 3.9	2.4 ± 6.2	1.1 ± 3.8	1.1 ± 3.4	1.2 ± 4.1	2.2 ± 6.2	1.1 ± 3.9
	Smokes (packs/year)	0.7 ± 2.6	0.4 ± 2.2	0.6 ± 1.9	0.4 ± 2.2	0.5 ± 2.3	0.4 ± 2.2	0.7 ± 2.3	0.4 ± 2.2
	Number of sexual partners	2.2 ± 0.9	2.5 ± 1.7	2.5 ± 1.2	2.5 ± 1.7	2.7 ± 1.3	2.5 ± 1.7	2.5 ± 1.3	2.5 ± 1.7
	First sexual intercourse (age)	16.8 ± 2.0	16.9 ± 2.8	17 ± 2.5	16.9 ± 2.8	16.9 ± 2.9	16.9 ± 2.8	17.1 ± 2.6	16.9 ± 2.8
	Number of pregnancies	2.4 ± 1.4	2.5 ± 1.7	2.6 ± 1.7	2.2 ± 1.4	2.1 ± 1.4	2.2 ± 1.4	2.3 ± 1.3	2.2 ± 1.4
	Hormonal contraceptives	0.7 ± 4.5	0.7 ± 0.5	0.6 ± 0.5	0.7 ± 0.5	0.7 ± 0.5	0.7 ± 0.5	0.7 ± 0.5	0.7 ± 0.5
Habit	Hormonal contraceptives (years)	2.9 ± 4.8	1.9 ± 3.5	3.2 ± 5.2	1.9 ± 3.4	3.3 ± 6.4	1.9 ± 3.4	3.3 ± 5.4	1.9 ± 3.4
	IUD	0.2 ± 0.4	0.1 ± 0.3	0.2 ± 0.4	0.1 ± 0.3	0.1 ± 0.3	0.1 ± 0.3	0.2 ± 0.4	0.1 ± 0.3
	IUD (years)	0.6 ± 1.5	0.4 ± 1.8	0.9 ± 2.9	0.4 ± 1.7	0.5 ± 1.7	0.4 ± 1.8	0.7 ± 2.0	0.4 ± 1.8
	STDs	0.2 ± 0.4	0.1 ± 0.3	0.2 ± 0.4	0.1 ± 0.3	0.2 ± 0.4	0.2 ± 0.3	0.2 ± 0.4	0.1 ± 0.3
	STDs (number)	0.3 ± 0.9	0.2 ± 0.5	0.4 ± 0.8	0.1 ± 0.5	0.3 ± 0.7	0.1 ± 0.5	0.3 ± 0.8	0.1 ± 0.5
	STDs: condylomatosis	0.1 ± 0.3	0.1 ± 0.2	0.1 ± 0.3	0.0 ± 0.2	0.1 ± 0.3	0.0 ± 0.2	0.1 ± 0.3	0 ± 0.2
	STDs: cervical condylomatosis	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0
	STDs: vaginal condylomatosis	0 ± 0	0.0 ± 0.1	0 ± 0	0 ± 0.1	0 ± 0	0 ± 0.1	0 ± 0	0 ± 0.1
History	STDs: vulvo-perineal condylomatosis	0.1 ± 0.3	0.0 ± 0.2	0.1 ± 0.3	0 ± 0.2	0.1 ± 0.3	0 ± 0.2	0.1 ± 0.4	0 ± 0.1
	STDs: syphilis	0.0 ± 0.2	0.0 ± 0.1	0.0 ± 0.2	0 ± 0.1	0 ± 0	0 ± 0.1	0 ± 0	0 ± 0.1
	STDs: pelvic inflammatory disease	0 ± 0	0.0 ± 0.0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0
	STDs: genital herpes	0 ± 0	0.0 ± 0.0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0.1	0 ± 0
	STDs: molluscum contagiosum	0 ± 0	0.0 ± 0.0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0
	STDs: AIDS	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0
	STDs: HIV	0.1 ± 0.3	0 ± 0.1	0.1 ± 0.3	0 ± 0.1	0.1 ± 0.3	0 ± 0.1	0.1 ± 0.3	0 ± 0.1
	STDs: Hepatitis B	0 ± 0	0.0 ± 0.0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0
	STDs: HPV	0 ± 0	0.0 ± 0.0	0 ± 0	0 ± 0	0 ± 0	0 ± 0.1	0 ± 0	0 ± 0.1
Genomics	Dx: CIN	0 ± 0	0.0 ± 0.1	0.0 ± 0.1	0 ± 0.1	0 ± 0	0 ± 0.1	0.1 ± 0.2	0 ± 0.1
	Dx: HPV	0.1 ± 0.3	0.0 ± 0.1	0.1 ± 0.3	0 ± 0.1	0.1 ± 0.3	0 ± 0.1	0.1 ± 0.3	0 ± 0.1
	Dx	0.1 ± 0.3	0.0 ± 0.2	0.1 ± 0.3	0 ± 0.2	0.1 ± 0.3	0 ± 0.2	0.1 ± 0.3	0 ± 0.1
	Dx: cancer	0.1 ± 0.3	0.0 ± 0.1	0.1 ± 0.3	0 ± 0.1	0.1 ± 0.3	0 ± 0.1	0.1 ± 0.3	0 ± 0.1

train the model. The description of these techniques is discussed in the section below.

2.5. Random Forest. Random Forest (RF) was first proposed by Breiman in 2001 [18]. Random forest is an ensemble model that uses decision tree as individual model and bagging as ensemble method. It improves the performance of decision tree by adding many trees to reduce the overfitting in the decision tree. RF can be used for both classification and regression. RF generates a random forest that contains decision trees and gets a prediction from each one of them and then selects the best solution with the maximum votes [19].

When training a tree, it is important to measure how much each feature decreases the impurity, as the decrease in the impurity indicates the significance of the feature. The tree classification result depends on the impurity measure used. For classification, the measures for impurity are either Gini impurity or information gain and for regression, and the measure for impurity is variance. Training decision tree consists of iteratively splitting the data. Gini impurity decides the best split of the data using the formula.

$$G = 1 - \sum_i p_i^2, \tag{3}$$

where p (i) is the probability of selecting a datapoint with class; i.e., Information gain (IG) is also another measure to decide the best split of the data depending on the gain of each feature. The formula that calculates the information gain is given in the following equation:

$$\begin{aligned} \text{Entropy} &= -\sum_i p_i \log_2 p_i, \\ \text{IG}(\text{parent}, \text{child}) &= \text{Entropy}(\text{parent}) - [p1(c1) * \text{entropy}(c1) + p(c2) * \text{entropy}(c2) + \cdots]. \end{aligned} \tag{4}$$

Table 2: Missing records per attribute in the data set.

Feature name	Missing values
First sexual intercourse (age)	7
Smoking	13
Smokes (years)	13
Smokes (packs/year)	13
Number of sexual partners	26
Number of pregnancies	56
Hormonal contraceptives	108
Hormonal contraceptives (years)	108
IUD	117
IUD (years)	117
STDs	105
STDs (number)	105
STDs: condylomatosis	105
STDs: cervical condylomatosis	105
STDs: vaginal condylomatosis	105
STDs: vulvo-perineal condylomatosis	105
STDs: syphilis	105
STDs: pelvic inflammatory disease	105
STDs: genital herpes	105
STDs: molluscum contagiosum	105
STDs: AIDS	105
STDs: HIV	105
STDs: hepatitis B	105
STDs: HPV	105
STDs: time since first diagnosis	787
STDs: time since last diagnosis	787

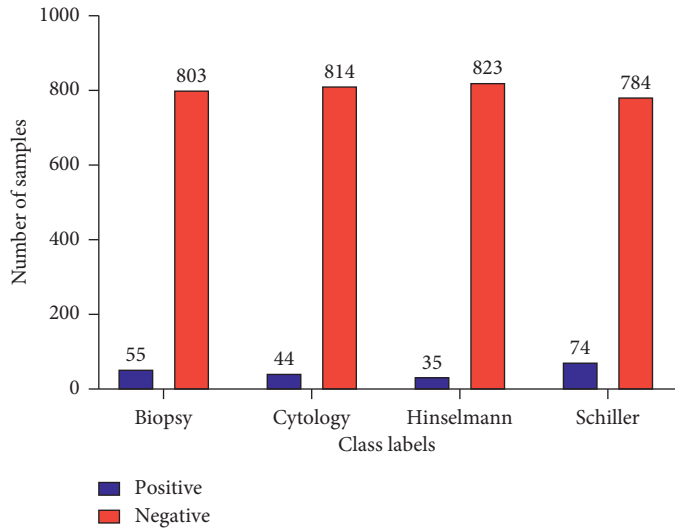

Figure 1: Number of records per class label in the data set.

2.6. Extreme Gradient Boosting. eXtreme Gradient Boosting (XGBoost) is a tree-based ensemble technique [20]. XGBoost can be used for classification, regression, and ranking problems. XG boosting is a type of gradient boosting. Gradient Boosting (GB) is a boosting ensemble technique that makes predicators sequentially instead of individually. GB is a method that produces a strong classifier by combining weak classifiers [21]. The goal of the GB is building an iterative model that optimizes a loss function. It pinpoints the failings of weak learners by using gradients in the loss function [21]:

$$y = ax + b + e, \tag{5}$$

where e denotes the error term. The loss function measures how good is the model at fitting the underlying data. The loss function depends on the optimization goal, for regression is a measure of the error between the true and predicated values, whereas, for classification, it measures the how good is a model at classifying cases correctly [21]. This technique takes less time and less iterations, since predictors are learning from the past mistakes of the other predictors. The

GB works by teaching a model C to predict values of the form

$$\mathcal{Y}' = C(x). \tag{6}$$

By minimizing a loss function, e.g., MSE:

$$\frac{1}{n}\sum_{i}\left(y_i' - y_i\right)^2, \tag{7}$$

where i iterates over a training set of size n of true values of the target variable yy′ = estimated values of $C(x)$ y = true values & n = number of instances in y.

Considering a GB model with M phases and m as a single phase being $(1 \le m \le M)$, to improve some deficient model F_m, a new estimator $h_m(x)$ is added. Therefore,

$$h_m(x) = \mathcal{Y} - F_m(x). \tag{8}$$

Estimator h will be fitted to $\mathcal{Y} - F_m(x)$, which is the difference between the true value and the predicated value, i.e., the residual. Thus, we attempt to adjust the errors of the previous model (F_m) [22].

XGBoost is better than Ada boost in terms of speed and performance. It is highly scalable and runs 10 times faster as compared to the other traditional single machine learning algorithms. XGBoost handles the sparse data and implements several optimization and regularization techniques. Moreover, it also uses the concept of parallel and distributed computing.

2.7. AdaBoost. Adaptive Boosting (AdaBoost) is a meta-learner originally proposed for the binary classification proposed by Freund and Schapire [23]. It is an ensemble technique to build a meta classifier by combining several weak classifiers using progressive learning.

AdaBoost uses the concept of boosting data sampling technique; adaptive sampling was used to assign high weights to the misclassified events. The misclassified samples will be selected in the next iteration to better train the model, and the final prediction was made using weighted voting. AdaBoost has reduced error rate, has a better effect on the prediction as compared to bagging [24], and uses decision tree stumps. Initially, all the samples in the data set have equal weights. Let x be the number of samples in the data set, and let y be the target. The target is a binary class represented by 0 and 1. The first decision tree stump will use some records from the data set, and predictions will be performed. After the initial prediction, the weights to the sample will be updated. More weights will be assigned to the data samples that were misclassified. The samples with the high weights will be selected in the next iteration. The process will be continued, unless the error rate is completely reduced, or a certain target level is achieved.

AdaBoost contains two main steps, combination and step forward using sequential iterative approach. All the instances in the training set have equal weights in the first iteration. However, in subsequent iterations, the weights are changed based on the error rates. The instances with error have increased weights. For the binary class classification problem containing T training samples is represented in the following equation:

$$\{(x_i, y_i)\}_{i=1}^{T}, \text{with}, \quad y_i \in \{0, 1\}. \tag{9}$$

Let C be the linear combination of weak classifiers. The combination of the classifiers is represented as

$$C(x) = \sum_{n=1}^{N} w_n c_n(x), \tag{10}$$

where N is the number of weak classifiers, w represents the weights, and $C(x)$ represents weak classifiers. In every next iteration, the classifier is trained based on the performance of the classifier in previous iteration.

$$C(x)_t = C(x)_{t-1} + w_n c_n(x), \tag{11}$$

where $C(x)_t$ represents the classifier in t iteration. $C(x)_{t-1}$ is the performance of the classifier at $t-1$ iteration.

The weights can be calculated using the following equation:

$$w_n = \frac{1}{2}\ln\left(\frac{1-\epsilon_n}{\epsilon_n}\right), \tag{12}$$

ϵ_n represents the error rate of the weak classifier.

2.8. Optimization Strategy. This section discusses optimization strategy to find the best hyperparameters combination that produces the highest targeted outcomes. Firefly optimization algorithm was used for parameter tunning. The details of Firefly are discussed in Section 2.3. Table 3 presents the hyperparameter values of Random Forest for all the four targets, For RF "gini" index criterion was used. Table 4 represents the parameters used for XGBoost. Gbtree booster was used with the random state of 42 and the learning rate of 0.05. Similarly, Table 5 presents the optimal feature vales for AdaBoost. Furthermore, Figures 2–4 represent the Grid Search optimization graph for Random Forest, Extreme Gradient Boosting, and AdaBoost classifier.

3. Experimental Setup and Results

The model was implemented in Python language 3.8.0 release using Jupyter Notebook environment. Ski-learn library was used for the classifiers along with other needed built-in tools, while separate library (xgboost 1.2.0) was used for XGBoost ensemble. There is K-fold cross validation with $K = 10$ for partitioning the data into training and testing. Five evaluation measures such as accuracy, sensitivity (recall), specificity (precision), positive predictive accuracy (PPA), and negative predictive accuracy (NPA) were used. Sensitivity and specificity are focused more during the study due to the application of the proposed model. Accuracy denotes the percentage of correctly classified cases, sensitivity measures the percentage of positives cases that were classified as positives, and specificity refers to the percentage of negative cases that were classified as negatives. Moreover, the criteria for the selection of the performance evaluation

Table 3: Random Forest optimized parameters set for all four targets using Firefly.

Parameters	Optimal values obtained
n_estimators	100
max_features	Log2
criterion	gini
max_depth	15
min_samples_split	5
min_samples_leaf	1

Table 4: XGBoost optimized parameters set for all four targets using Firefly.

Parameters	Optimal values obtained
Booster	gbtree
Random_state	42
nthread	8
learning_rate	0.05
gamma (i.e. min_split_loss)	0.1
max_depth	3

Table 5: AdaBoost optimized parameters set for all four targets using Firefly.

Parameters	Optimal values obtained
base_estimator	None
n_estimators	600
learning_rate	1.0

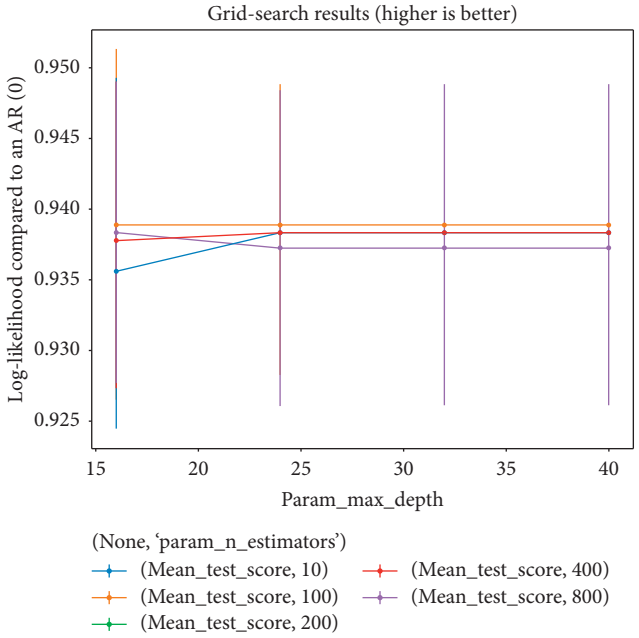

Figure 2: Grid search optimization for random forest.

measures depend upon the measures used in the benchmark studies. Two sets of experiments were conducted for each target using selected features by using Firefly feature selection algorithm and 30 features for four targets. The SMOTE technique was applied to generate synthetic data. The results of model are presented in section below.

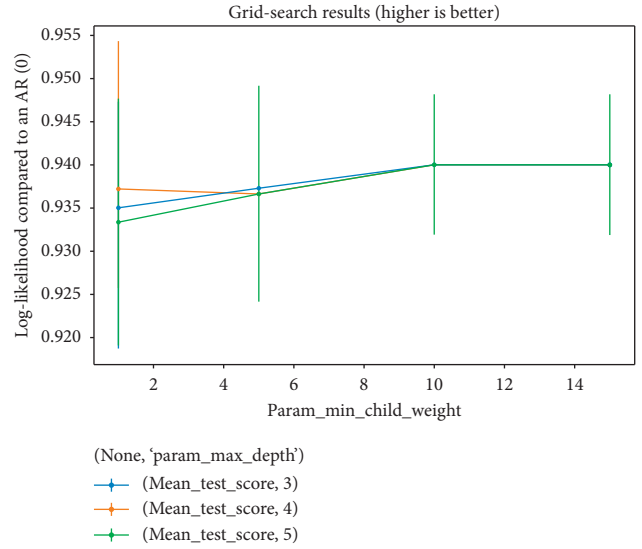

Figure 3: Grid search optimization for extreme gradient boosting.

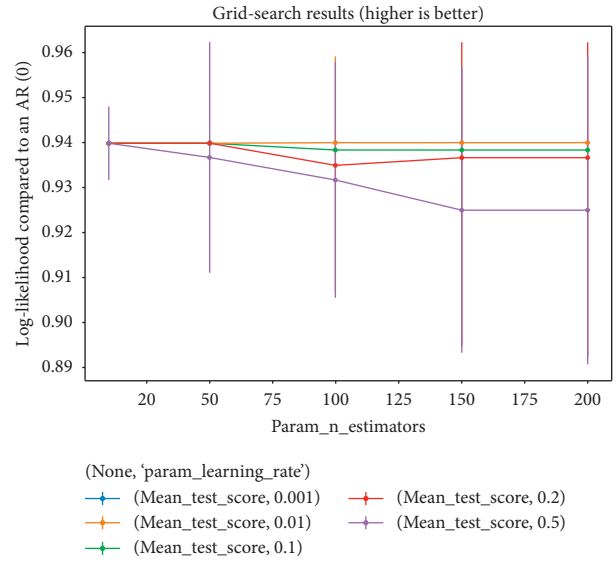

Figure 4: Grid search optimization for AdaBoost.

3.1. Hinselmann. Table 6 presents the accuracy, sensitivity, specificity, PPA, and NPA for the RF, AdaBoost, and XGBoost models, respectively, using SMOTE for Hinselmann test target class. The number of selected features for Hinselmann was 15. XGBoost outperformed the other classifiers for both feature sets. However, the performance of XGBoost with selected feature is better when compared with 30 features. The model produces an accuracy of 98.83, sensitivity of 97.5, specificity of 99.2, PPA of 99.17, and NPA of 97.63, respectively.

3.2. Schiller. Table 7 presents the outcomes for the Schiller test. Like Hinselmann target, XGBoost with selected features outperformed that of Schiller, respectively. However, the outcomes achieved by the model for Schiller are lower when compared with Hinselmann target class. The performance of

Table 6: Performance of ensemble classifiers using 30 and selected feature for Hinselmann target.

Features	Model	Accuracy	Sensitivity	Specificity	PPA	NPA
30	RF	96.36	97.05	95.72	95.44	97.23
	AB	90.08	90.34	89.84	89.21	90.91
	XGB	97.37	96.72	98	97.93	96.84
Selected	RF	97.57	97.12	98.01	97.93	97.23
	AB	94.94	95	94.88	94.61	95.26
	XGB	**98.38**	**97.55**	**99.2**	**99.17**	**97.63**

Table 7: Performance of ensemble classifiers using 30 and selected feature for Schiller target.

Features	Model	Accuracy	Sensitivity	Specificity	PPA	NPA
30	RF	93.84	94.54	93.12	93.36	94.35
	AB	86.62	85.32	88.13	89.21	83.91
	XGB	92.36	93.99	90.76	90.87	93.91
Selected	RF	95.97	95.49	96.48	96.68	95.89
	AB	89.6	86.92	92.89	93.78	85.22
	XGB	96.98	95.9	96.92	97.1	95.65

RF and XGBoost is similar with selected feature for Schiller with a minor difference. The number of features selected by Firefly for Schiller was 13.

3.3. Cytology. Table 8 presents the outcome of all the classifiers for the cytology diagnosis tests. Like Hinselmann and Schiller diagnostic test, XGBoost outperformed Cytology test as well with selected features. For specificity and accuracy, similar outcomes were achieved using 30 and selected features. Similarly, the performance of RF is similar in both 30 and selected features. The number of features selected by Firefly feature selector for Cytology was 11.

Table 8: Performance of ensemble classifiers using 30 and selected feature for Cytology target.

Features	Model	Accuracy	Sensitivity	Specificity	PPA	NPA
30	RF	95.91	94.09	97.87	97.95	93.88
	AB	93.66	91.12	96.52	96.72	90.61
	XGB	96.32	94.49	98.3	98.36	94.29
Selected	RF	95.91	94.44	97.47	97.54	94.29
	AB	92.64	90.94	94.47	94.67	90.61
	XGB	**96.93**	**95.26**	**99.74**	**98.73**	**95.1**

3.4. Biopsy. Similarly, performance was not drastically different, yet using all the features resulted in a higher accuracy than when using SMOTE with selected features for Biopsy as shown in Table 9. XGB obtained the highest accuracy of 97.1 with all features. However, for other measures, the performance of the XGBoost is better with the selected features. Similar performance was achieved for all measures when classified using RF for both feature sets 30 and selected, respectively. The number of selected features used for Biopsy target class was 11.

Table 9: Performance of ensemble classifiers using 30 and selected feature for Biopsy target.

Features	Model	Accuracy	Sensitivity	Specificity	PPA	NPA
30	RF	96.68	95.45	97.32	97.88	95.53
	AB	93.57	93.99	93.17	92.8	94.31
	XGB	97.1	95.49	97.52	97.46	95.53
Selected	RF	96.27	95.8	96.72	96.61	95.93
	AB	89.21	88.02	90.42	90.25	88.21
	XGB	**96.68**	**95.83**	**98.74**	**98.73**	**95.93**

Overall, after comparing all the four-diagnostic tests, Hinselmann test achieved the better outcome and can be used for the diagnosis of cervical cancer as shown in Table 10. As per the outcome achieved in the proposed study, Hinselmann diagnosis test has better performance when compared from other cervical cancer diagnosis tests like Schiller, Biopsy, and Cytology, respectively. Similar findings have been made in Abdoh et al. [5] and Wu and Zhou [4] study.

Table 10: Results of proposed model for 4 diagnosis tests for cervical cancer.

Diagnosis test	Accuracy	Sensitivity	Specificity	PPA	NPA
Hinselmann	**98.38**	**97.55**	**99.2**	**99.17**	**97.63**
Schiller	96.98	95.9	96.92	97.1	95.65
Biopsy	96.68	95.83	98.74	98.73	95.93
Cytology	96.93	95.26	98.73	98.77	95.1

4. Comparison with Existing Studies

The study used three ensemble techniques AdaBoost, extreme Gradient Boosting, and Random Forest. Furthermore, the proposed study is the pioneer in using bioinspired algorithm for feature selection and optimization for cervical cancer diagnosis. To explore the significance of our proposed study, the outcome of the study was compared with the benchmark studies. The criteria for the benchmark studies selection were based on data set used for the diagnosis of cervical cancer. Table 11 contains the comparison of the proposed technique with the benchmark studies in the literature. The best outcomes in the benchmark studies were achieved using 30 features. However, some of the outcomes in the previous studies were achieved with the reduced features. The number in the brackets next to some of the outcomes represents the number of features.

Therefore, based on Table 11, the proposed study outperforms the two studies in the benchmark interms of accuracy with reduced risk factors. However, the achieved sensitivity and NPA are less than those of Wu and Zhou [4] but higher than those of Abdoh et al. [5]. The number of features in Wu et al. study is 30, while the proposed study used reduced risk factors. The specificity and PPA of the proposed study are higher than those of the benchmark studies except for the Schiller diagnosis test.

In nutshell, the main contributions of the current study are applying bioinspired algorithm for feature selection and for model optimization for cervical cancer risk factors. The

Table 11: Comparison of the proposed study with benchmark studies.

Target class	Model	Accuracy	Sensitivity	Specificity	PPA	NPA
Hinselmann (15)	Abdoh et al. [5]	93.97	**100**	89.96	84.97	**100**
	Wu and Zhou [4]	97.6	96.65	98.54	98.48	96.78
	Proposed study	**98.38**	97.55	**99.2**	**99.17**	97.63
Schiller (13)	Abdoh et al. [5]	90.18	**98.73**	84.63	80.75	**99.03**
	Wu and Zhou [4]	95.01	93.24	**97.58 (12)**	**97.29 (12)**	93.81
	Proposed study	**96.39**	95.9	96.92	97.1	95.65
Cytology (11)	Abdoh et al. [5]	92.75	**100**	87.92	83	**100**
	Wu and Zhou [4]	96.94	95.58 (8)	99.01	98.94	95.76 (8)
	Proposed study	**96.98**	95.83	**99.74**	**98.73**	95.93
Biopsy (11)	Abdoh et al. [5]	94.13	**100**	90.21	86.07	**100**
	Wu and Zhou [4]	96.06	94.94 (6)	97.76 (11)	97.58 (11)	94.91
	Proposed study	**96.93**	95.26	**98.73**	**98.77**	95.1

proposed model enhanced the outcomes when compared with the previous studies related with cervical cancer risk factors data set. Despite the above-mentioned advantages, the study suffers from some limitations: the data set suffers from huge imbalance, and augmented data was generated using SMOTE. Moreover, the current study was based on open-source data set, and further testing is required to use other real and open-source data sets.

To alleviate the above-mentioned limitations, there is a need for validating the model on real data set from the hospital.

5. Conclusion

This study presents an investigation of several ensemble techniques such as Random Forest, AdaBoost, and Extreme Gradient Boosting for diagnosing cervical cancer. The data set was obtained from the UCI machine learning repository containing 858 records, 32 features, and 4 target variables. The target variables are the diagnosis test used for cervical cancer. Experiments were conducted for each target class separately. Data preprocessing includes imputing missing values and class balancing using SMOTE. Moreover, bioinspired firefly algorithm was used to optimize the models, and to identify the key features. To compare the performance of the models, the experiments were conducted with 30 features and the selected features using SMOTED data. Extreme Gradient Boosting outperformed the other two models for all four target variables. For future work, the model will be validated on multiple data sets. Also, other models that can handle outliers and unbalanced data differently should be investigated.

Data Availability

The study used open-source data set available at https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] American Cancer Society, *Key Statistics for Cervical Cancer*, American Cancer Society, Atlanta, GA, USA, 2021, https://www.cancer.org/cancer/cervical-cancer/about/key-statistics.html.

[2] "Cervical cancer: risk factors," 2020, https://www.cancer.net/cancer-types/cervical-cancer/risk-factors#:%7E:text=The%20most%20important%20risk%20factor,100%20different%20types%20of%20HPV.

[3] "Cervical cancer," 2020, https://www.nccc-online.org/hpvcervical-cancer/cervical-cancer-overview/.

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

[5] S. F. Abdoh, M. Abo Rizka, and F. A. Maghraby, "Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques," *IEEE Access*, vol. 6, 2018.

[6] J. Lu, E. Song, A. Ghoneim, and M. Alrashoud, "Machine learning for assisting cervical cancer diagnosis: an ensemble approach," *Future Generation Computer Systems*, vol. 106, pp. 199–205, 2020.

[7] E. Karim and N. Neehal, "An empirical study of cervical cancer diagnosis using ensemble methods," in *Proceedings of the 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, Dhaka, Bangladesh, December 2019.

[8] A. Ul-Islam, S. H. Ripon, and N. Qaisar Bhuiyan, "Cervical cancer risk factors: classification and mining associations," *APTIKOM Journal on Computer Science and Information Technologies*, vol. 4, no. 1, pp. 8–18, 2019.

[9] Y. M. S. Al-Wesabi, A. Choudhury, and D. Won, "Classification of cervical cancer dataset," in *Proceedings of the 2018 IISE Annual Conference and Expo*, pp. 1456–1461, Orlando, FL, USA, May 2018.

[10] 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, no. 5, pp. e154–21, 2018.

[11] K. Adem, S. Kiliçarslan, and O. Cömert, "Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification," *Expert Systems with Applications*, vol. 115, pp. 557–564, 2019.

[12] K. Fernandes, J. S. Cardoso, and J. Fernandes, "Transfer learning with partial observability applied to cervical cancer screening," *Pattern Recognition and Image Analysis*, Springer, vol. 10255, pp. 243–250, Berlin, Germany, 2017.

[13] "Cervical cancer (risk factors) data set," 2020, https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29.
[14] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, 2018.
[15] X. Yang, *Nature-Inspired Metaheuristic Algorithms*, Vol. 4, Luniver Press, Cambridge, UK, 2nd edition, 2010.
[16] B. H. Nguyen, B. Xue, and M. Zhang, "A survey on swarm intelligence approaches to feature selection in data mining," *Swarm and Evolutionary Computation*, vol. 54, 2020.
[17] M. Anbu and G. S. Anandha Mala, "Feature selection using firefly algorithm in software defect prediction," *Cluster Computing*, vol. 22, no. s5, Article ID 10925, 2019.
[18] L. Breiman, "Random forests," in *Hands-On Machine Learning with R*, pp. 203–219, CRC Press, Boca Raton, FL, USA, 2019.
[19] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?" in *Lecture Notes in Computer Science*, vol. 7376, pp. 154–168, Springer, New York, NY, USA, 2012.
[20] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13–17, pp. 785–794, San Francisco, CA, USA, August 2016.
[21] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in Neurorobotics*, vol. 7, 2013.
[22] J. H. Friedman and J. J. Meulman, "Multiple additive regression trees with application in epidemiology," *Statistics in Medicine*, vol. 22, no. 9, pp. 1365–1381, 2003.
[23] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.
[24] S. B. Kotsiantis, "Supervised machine learning: a review of classification techniques," *Informatica*, vol. 31, pp. 249–268, 2007.