Original Article

Cervical Cancer Prediction by Merging Features of Different Colposcopic Images and Using Ensemble Classifier

Abstract

Background: Cervical cancer is a significant cause of cancer mortality in women, particularly in low-income countries. In regular cervical screening methods, such as colposcopy, an image is taken from the cervix of a patient. The particular image can be used by computer-aided diagnosis (CAD) systems that are trained using artificial intelligence algorithms to predict the possibility of cervical cancer. Artificial intelligence models had been highlighted in a number of cervical cancer studies. However, there are a limited number of studies that investigate the simultaneous use of three colposcopic screening modalities including Greenlight, Hinselmann, and Schiller. **Methods:** We propose a cervical cancer predictor model which incorporates the result of different classification algorithms and ensemble classifiers. Our approach merges features of different colposcopic images of a patient. The feature vector of each image includes semantic medical features, subjective judgments, and a consensus. The class label of each sample is calculated using an aggregation function on expert judgments and consensuses. **Results:** We investigated different aggregation strategies to find the best formula for aggregation function and then we evaluated our method using the quality assessment of digital colposcopies dataset, and our approach performance with 96% of sensitivity and 94% of specificity values yields a significant improvement in the field. **Conclusion:** Our model can be used as a supportive clinical decision-making strategy by giving more reliable information to the clinical decision makers. Our proposed model also is more applicable in cervical cancer CAD systems compared to the available methods.

Keywords: *Aggregation strategy, artificial intelligence, cervical cancer, ensemble classifier, machine learning*

Submitted: 22-Feb-2020 **Revised:** 15-Mar-2020 **Accepted:** 02-May-2020 **Published:** 24-May-2021

Elham Nikookar[1], Ebrahim Naderi[2], Ali Rahnavard[3]

[1]Department of Computer Engineering, Faculty of Engineering, Shiahd Chamran University of Ahvaz, [2]Department of Computer Engineering, University of Applied Science and Technology, Ahvaz, Iran, [3]Computational Biology Institute, Department of Biostatistics and Bioinformatics, Milken Institute School of Public Health, The George Washington University, Washington D.C., United States

Address for correspondence:
Mrs. Elham Nikookar, Department of Computer Engineering, Faculty of Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran.
E-mail: e.nikookar@scu.ac.ir

Introduction

Cervical cancer is a type of cancer in which the abnormal cell growth occurs on the surface lining of the cervix. These cells have the potential to invade the surrounding tissues and organs. Its symptoms may include abnormal vaginal bleeding, pelvic pain, or pain during sexual intercourse.[1] According to Fernandes *et al*.,[2] this disease occurs in more than half a million cases per year, and it kills more than a quarter of a million people in the same period. Although cervical cancer can be prevented through regular screening methods, it remains a significant cause of mortality, particularly in low-income countries. This phenomenon has motivated many experts in various fields of science, such as medical and computer science, to reduce the mortality rates by contributing an innovative approach for cervical cancer prevention.[3-5]

Applications of artificial intelligence techniques in the medical domain have become a hot topic in recent years. There are a number of intelligent systems that have been proposed to facilitate the decision-making process for physicians.[6-8] In the case of cervical cancer, precancerous cervical cancer examination is effective to reduce cervical cancer incidence and mortality as it has several preventive actions. Prediagnosis methods of cervical cancer are a diagnostic procedure for patients, who have symptoms of cervical cancer, and a screening examination for patients, who can be infected in future. The screening strategies for the detection of precancerous cervical lesions include cytology, colposcopy, and the gold-standard biopsy. From a computer-aided diagnosis (CAD) system point of view, a digital image-processing toolbox provides physicians with advanced screening and prediagnosis methods for cervical cancer predetection.

How to cite this article: Nikookar E, Naderi E, Rahnavard A. Cervical cancer prediction by merging features of different colposcopic images and using ensemble classifier. J Med Sign Sens 2021;11:67-78.

Access this article online

Website: www.jmssjournal.net

DOI: 10.4103/jmss.JMSS_16_20

Quick Response Code:

The focus of this study is on the colposcopy imaging procedure, which allows a physician to closely examine the cervix and the tissues of vagina and vulva. Many premalignant and malignant lesions in these areas have discernible characteristics that can be detected through the examination. Different modalities of colposcopy include Hinselmann, Greenlight, and Schiller, which are used for the treatment. Each of the modalities has its strengths. The Hinselmann modality is the oldest colposcopic examination whereby a stereoscopic magnified view of the illuminated cervix is obtained with a stand-mounted binocular instrument,[9,10] and the resulting image is used to check the state of cervical cancer. In Greenlight modality, the colposcopist assesses vessel patterns that can indicate the existence of more advanced cancerous or precancerous lesions.[11] In Schiller modality, better visualization of abnormal areas in the cervix can be obtained to assist the physicians to decide whether a cancerous lesion is detectable or not.[11] The resulting images of these modalities are a reference for the assessment to predict the possible occurrence of cervical cancer. The judgments of several physicians are taken into account to enhance the quality of the assessment, and the consensus is considered as a final decision of the assessment.

Due to the limited access to physicians and high cost of medical comments, the potential of machine learning that explores the construction of intelligent models to predict the risk of cervical cancer based on the results of colposcopy has been considered since the past few years.[12,13] In these intelligent models, the images extracted from the imaging device are examined by image-processing algorithms and the features of images are extracted. The possibility of occurrence of cervical cancer is then predicted using the extracted features of cervix image and through a supervised classification algorithm that has been built and trained.

Applications of colposcopic imaging modalities, as a tool to evaluate the state of cervical cancer or predict the possibility of its occurrence, have been highlighted in a number of analyses.[5,13-16] However, the simultaneous use of three colposcopic modalities, namely Greenlight, Hinselmann, and Schiller, is an unmet need. In this research, we developed an intelligent model for predicting the occurrence of cervical cancer by constructing an ensemble-based model in which feature vectors of cervix images for three different colposcopy modalities are merged. Their class label is produced by applying best performing aggregation strategy on a combination of expert judgments and different classification algorithms which improved the performance of the model. We investigated different approaches of aggregating judgments and consensus of six physicians for each patient to identify the best strategy of aggregation in the case of performing different colposcopy tests. In addition, in our model, we evaluated different configurations for classifier module to compare the performance of our multi-tiered model with a simple model that uses one colposcopy modality dataset. The proposed model evaluation using the quality assessment of digital colposcopies dataset[17] performs at 96% of sensitivity, 94% of specificity, 91% of F-Score, and competitive 0.94 receiver operating characteristic (ROC) area, which are exceedingly effective compared to a model that uses one colposcopy modality dataset.

In this work, we first investigate and evaluate some of the existing studies in cervical cancer detection and prediction domain, which culminates with an identification of the knowledge gap and inconsistencies in the literature. We elaborate the dataset that is used to train, test, and evaluate our proposed model, following by the proposed model. Finally, we show our prediction results and future direction.

Literature Review

Artificial intelligence techniques have been increasingly used in the medical domain.[5,18-20] Numerous studies have applied these techniques, particularly machine learning and image processing, to solve multifaceted clinical problems in medical field, especially in diagnosing cervical cancer.[5,13-16,21]

Xu *et al*.[5] built an image dataset as a benchmark that was tested using different classifiers for evaluating cervical disease classification algorithms. In their training phase, they used a uniform strategy, called Exhaustive Grid Search,[22] to search for the optimal parameters of each classifier. In their investigation, they applied a type of convolutional neural network (ft-CNN), and they, respectively, achieved 0.8694, 83.42%, 88.3%, and 83.41% of area under the ROC curve (AUC), accuracy, sensitivity, and specificity.

Phoulady *et al*.[13] proposed a model for classifying a cervical cell as normal or cancerous using a large ensemble of segmentations, which separates normal and cancerous cases based on the single feature of mean nuclear volume. They used four basic segments to differentiate the nucleus area of the cell from the background. The segments' vote is applied for the final feature, which is the mean nuclear volume for each particular case. The mean nuclear volume is then used for distinguishing between cancerous and normal cases.

Pfohl *et al*.[16] proposed an algorithm for identifying patient's cervical type, namely TZs (1, 2, and 3), and used the resulting information for cervical cancer treatment. They developed and implemented a CNN-based algorithm for distinguishing between three cervical types using transfer learning pipelines for a fine-tuning deep CNN. They reported that they could improve the performance of the developed model. Based on their experiment, the proposed algorithm achieved 81% of accuracy.

Sarwar *et al*.[15] introduced a hybrid ensemble algorithm to improve the predictive performance of cervical cancer

screening by characterizing and classifying the Pap smear images. They had applied several classification algorithms on their dataset and used the results for constructing an ensemble of ensemble techniques, named as hybrid ensemble algorithm. According to their experiment, the hybrid ensemble algorithm achieved 98.57% of correct classification for the two-class problem and 78.8571% of correct classification for the seven-class problem compared to other algorithms.

Arteta *et al*.[14] proposed a machine learning-based cell detection method that applies to different modalities. For each cervigram, they initially identified a set of candidate cell-like regions; then, each candidate region is evaluated using a statistical model of the cell appearance. Finally, they used dynamic programming to pick a set of nonoverlapping regions that were matched with the model. The results of cell detection in cervix images were then applied for the automation of cell-based experiments in cervical cancer domain. Based on the results, they achieved 86.99% of precision, 90.03% of recall, and 88.48% of the F-score value of the proposed model.

Fernandes *et al*.[21] proposed a framework to predict cross-modality individual risk and cross-expert subjective quality assessment of colposcopic images for different modalities by transferring knowledge gained from one expert/modality to another. Their research was by transfer learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem.[23] In the proposed framework, they transferred adjusted parameters of a source model which uses one of the three colposcopy modalities to a target model which uses another modality. Parameters include coefficients and weights of classification and regression models. According to their experiment, they achieved 68.30% of accuracy for their proposed algorithm. It is needed to point out that although Fernandes *et al*.[21] used three colposcopy modalities, their framework was based on a serial approach that examines one of the modalities in each step and applies the results on another modality in the next step. This approach is different from the approach of the proposed model in this study, which aims to use all the three colposcopy modalities simultaneously. Table 1 shows a summary of review on previous machine-learning and image-processing applications in cervical cancer domain.

According to Table 1, previous works did not investigate the simultaneous use of the three colposcopic modalities to predict the occurrence of cervical cancer based on colposcopy cervix images. Lack of research studies on this topic makes it unclear whether the artificial intelligence algorithms are capable of providing a model that utilizes the power of the simultaneous use of the three colposcopic modalities. Therefore, the present study is focused on the enhancement of cervical cancer prediction by merging features of different colposcopic images using ensemble classifier.

Dataset

The quality assessment of digital colposcopy dataset[17] is used as a dataset in this study. This dataset was acquired and annotated by professional physicians at Hospital Universitario de Caracas. The subjective judgments (target variables) were originally done in an ordinal manner (poor, fair, good, and excellent) and were discretized in two classes (bad and good). The images were randomly sampled from the original colposcopic sequences in the form of a video. The dataset has three colposcopy modality images of Greenlight, Hinselmann, and Schiller, as illustrated in Figure 1. It consists of approximately 100 cervigrams per modality and totally 287 cervigrams. For each cervigram, 69 features including 62 medial semantic features, six subjective judgments of physicians, and a consensus are considered. The semantic medical features[21] are mentioned as follows:

- Image area occupied by each anatomical body part (cervix, external os, and vaginal walls) and occluding objects (speculum and other artifacts)
- The area of each region occluded by artifacts or by specular reflections
- The maximum area difference between the four cervix quadrants
- Fitness goodness of the cervix to a given geometric model: Convex hull, bounding box, circle, and ellipse
- The distance between the image center and the cervix centroid/external os
- Mean and standard deviation of each RGB and HSV channel in the cervix area and in the entire image.

Proposed Model

The proposed model consists of three modules including merger module, classifier panel module, and classifier selector module. The initial dataset is first given to the merger module to apply aggregation strategies on the samples and produce newly merged datasets to feed in the next module. In the classifier panel module, different classification algorithms are applied on the merged datasets to produce input data for classifier selector module, in which different classifier selection methods are applied to generate final results of the model. A general flowchart of the proposed model is illustrated in Figure 2.

The following subsections describe all the three modules of the proposed model. Validation technique, evaluation measures, and single modality model, which have been applied to the data for the comparison of the results, are also described.

Merger module

This study considers the following scenario to create a merged dataset. First, all the cervigrams are separately grouped based on their colposcopy modalities, which include Greenlight, Hinselmann, and Schiller. Therefore, each patient has three cervigrams. It should be noted that

Table 1: Review of machine-learning and image-processing applications in cervical cancer domain

Reference	Research focus	Method used	One modality	Three modalities
[16]	Determination of a patient's cervical type	Development and implementation of a CNN-based algorithm for distinguishing between three cervical types	√	
[14]	Cell detection in cervix images for the automation of cell-based experiments in cervical cancer domain	Identifying a set of candidate cell-like regions in cervix image and evaluating each candidate region using a statistical model of the cell appearance. Dynamic programming is used to pick a set of nonoverlapping regions that match the model	√	
[13]	Classifying cervical cells as normal or cancer	Provide a large ensemble of segmentations which separate normal and cancer cases by using votes of different segments	√	
[15]	Improving the predictive performance of artificial intelligence-based system for screening of cervical cancer	Creating a hybrid ensemble which is, in fact, an ensemble of ensemble classifiers	√	
[21]	Predict cross-modality individual risk and cross-expert subjective quality assessment of colposcopic images for different modalities	Transfer knowledge gained from one modality to another		√
[5]	Investigating the performance of CNN features for cervical disease classification	Applying different classifiers to their data to find optimal parameters of each classifier	√	

CNN – Convolutional neural network √ – One modality or Three modality dataset is used

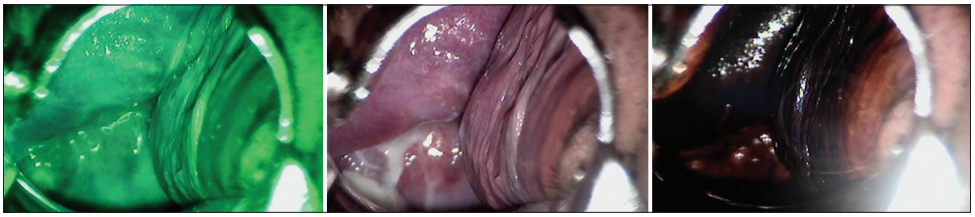

Figure 1: Greenlight, Hinselmann, and Schiller colposcopy modalities

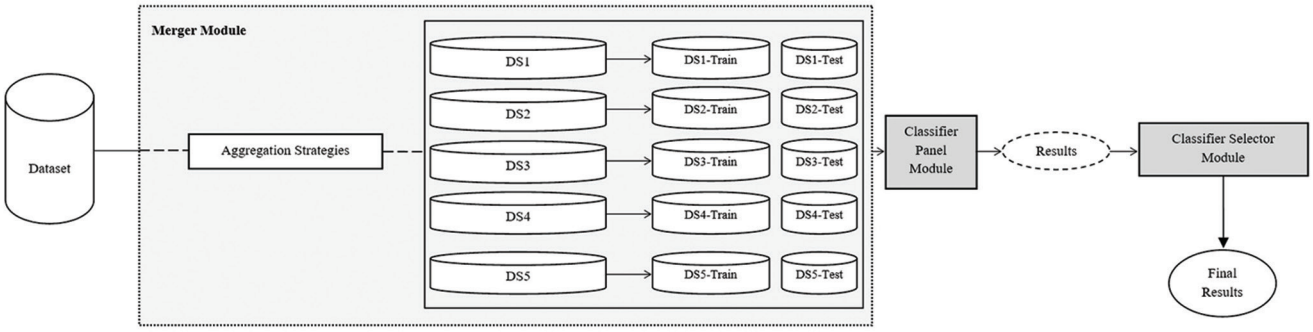

Figure 2: The general flowchart of the proposed model

for each cervigram, 62 semantic medical features were extracted from the cervix image and seven features were subjective judgments and consensus. In other words, 186 = 62 × 3 semantic medical features, 18 = 6 × 3 subjective judgments, and 3 = 1 × 3 consensuses are available for each patient. As this study applies the classification algorithms on the merged datasets, a class label for each patient needs to be considered in its merged feature vector. It is necessary to point out that the dataset of each single colposcopy modality has a class label and it is a consensus, which is available for each sample. There are three different modalities for a merged dataset. Therefore, three consensuses are available for each sample, and there is a need to generate a single class label for each sample in the merged dataset. The class label of samples in the merged dataset needs to be extracted from 18 subjective judgments and three consensuses by considering an aggregation function. Different possibilities

for aggregating subjective judgments and consensuses were obtained, which led to creating different merged datasets. Between all the possibilities, five meaningful aggregation strategies have been applied to the subjective judgments and consensuses. In each aggregation strategy, the result of its respective formula has been taken as the actual class label in the sample-merged feature vector. Generally, each patient-merged feature vector consists of 187 = 62 × 3 + 1 entry including 186 = 62 × 3 semantic medical features plus a class label resulting from the aggregation strategy. To formulate the calculation of each patient's actual class label, this study uses the definitions that are stated in Table 2.

For each image, the actual label of a sample is defined based on the following aggregation functions:

Majority vote between subjective judgments

$$F_{mvs}(x) = 1 \text{ if } (\Sigma S_{ij})/18 > 0.5 \text{ otherwise } 0 \quad (1)$$

Majority vote between consensuses

$$F_{mvc}(x) = 1 \text{ if } (\Sigma C_i)/3 > 0.5 \text{ otherwise } 0 \quad (2)$$

Product of all subjective judgments

$$F_{ps}(x) = 1 \text{ if } (\Pi S_{ij}) = 1 \text{ otherwise } 0 \quad (3)$$

Product of all consensuses

$$F_{pc}(x) = 1 \text{ if } (\Pi C_i) = 1 \text{ otherwise } 0 \quad (4)$$

At least one positive consensus

$$F_{opc}(x) = 1 \text{ if } (\Sigma C_i) >= 1 \text{ otherwise } 0 \quad (5)$$

This study does not use "at least one positive subjective judgment" as an aggregation formula because in such aggregation formula, the class label of all samples would be "1" because each sample has at least one positive subjective judgment between 18 subjective judgments and therefore, applying a classification algorithm is meaningless.

At the end of this phase, five different merged datasets (DS_{mvs}, DS_{mvc}, DS_{ps}, DS_{pc}, and DS_{opc}) are obtained, in which the class labels of samples are calculated based on respective aggregation function. According to the investigation of the final results of the model, it can be concluded that the proposed aggregation strategies provide better results regarding the desired evaluation measures. The results of the study imply two aspects. First, from the artificial intelligence aspect, the predictor model is more reliable for diagnosing cervical cancer. Second, from the clinical decision-making aspect, the results would be valuable for the decision makers in providing more reliable information. In addition, the results would advance the clinical decision-making procedure for cervical cancer detection or prediction. It might improve the procedure for taking subjective judgments for a colposcopic image or having consensus meetings about the prediction of cervical cancer based on colposcopic images in health centers.

Table 2: Parameters in Aggregation function

Symbol	Description
S_{ij}, i=1 to N_m, j=1 to N_s	Subjective judgment from expert physician j for cervical image from modality i
C_i, i=1 to3	Consensus for cervical image from modality i
N_m	Total number of modalities in this study
N_s	Number of subjective judgments for each modality

Classifier panel module

In this phase, the data are split using a stratified training-test partition (80–20) for each of the five datasets resulting from the merger module. Different classification algorithms on each dataset are then applied to feed the results in the classifier selector module in the next phase. The classifiers that are considered as potential members of the ensemble include Naive Bayes, AdaBoost, Random Forest, Random Tree, support vector machine, Decision tree, and Logit Boost.

The motivation behind considering odd number of classifiers in the ensembling process is the pigeonhole principle,[24] which states that for natural numbers k and m, if $n = k_m + 1$ objects are distributed among m sets, at least one of the sets will contain at least $k + 1$ objects. For arbitrary n and m, it generalizes to $k + 1 = \lfloor (n-1)/m \rfloor + 1$, where $\lfloor \rfloor$ is the floor function. It means that in the two-class problem (healthy 0, unhealthy 1), in which each classifier has to give its vote for the class of a sample, there is a need to have an odd number of classifiers to avoid equal 0 and 1 predictions for a sample. This odd number is considered seven in this study. The increasing number of classifiers may obviously result in finding a more powerful model. However, this study limits the number of classifiers to seven.

Classifier selector module

After training and testing the seven classifiers that have been mentioned in the previous section, a new feature vector is built with eight features for each aggregation strategy. It includes predicted class labels by seven classifiers plus the actual class label tagged by respective aggregation strategy. Therefore, five datasets are built, and they are used to find an optimal ensemble for each aggregation strategy, such as the best subset or combination of classifiers. Several methods have been tested for selecting the best subset of classifiers. Best classifier and all classifiers are the most straightforward classifier selection methods that are applied to the classifiers. Forward search (FS) and backward search (BS) methods are also considered, the formal descriptions of which are taken from various studies.[20,25] The idea behind principal component analysis (PCA) is taken into account for our proposed "PCA on classifiers" classifier selection method. The investigated classifier selection methods are described as follows:

Best classifier

The ensemble contains only the best performing classifier.

All classifiers

All classifiers are members of the ensemble.

Forward search

FS is the most intuitive greedy algorithm. First, the best individual classifier is selected. In the next iterations, further classifiers are added if the performance of the ensemble increases. The process ends when no further performance increase is reached by adding more classifiers. In each iteration, ROC area is considered as the decisive evaluation measure, and it needs to be maximized. The evaluation measures will be discussed in classifier panel module subsection. Algorithm 1 gives a formal description of this search method.

Algorithm 1.

FSR = Classifier with Maximum RA

for all $cl_i \in CL$, $i = 1.7$

if RA (FSR U cl_i) > RA (FSR)

RA (FSR) ← RA (FSR U cl_i)

FSR ← FSR U cl_i

end if

end for

Where FSR stands for forwarding search result, RA stands for ROC area, CL stands for classifier list, and cl_i, $i = 1.7$ is a classifier.

Backward search

BS is symmetrical to FS. First, all classifiers are considered as members of the ensemble. Then, the classifiers are removed from the ensemble if the performance of the ensemble increases. The process stops when no further performance increase is reached by removing more classifiers. Similar to FS, in each iteration, ROC area is considered as the decisive evaluation measure, and it needs to be maximized. Algorithm 2 gives a formal description of this search method.

Algorithm 2.

BSR = Contains All 7 Classifiers

for all $cl_i \in CL$, $i = 1.7$

if RA (BSR-cl_i) > RA (BSR)

RA (BSR) ← RA (BSR-cl_i)

BSR ← BSR-cl_i

end if

end for

Where BSR stands for BS results, RA stands for ROC area, CL stands for classifier list, and cl_i, $i = 1.7$ is a classifier.

Principal component analysis on classifiers

Practically, a set of all the seven classifiers creates a vector, which is called as classifier vector in this study. In the experiment, a PCA algorithm is applied on the classifier vector to form a new classifier vector in which every component is a combination of classifiers. In each component of the new classifier vector, the selected classifiers are assigned a weight, which determines their level of effectiveness on classification. The number of new classifier vector components is fixed to three in order to prevent the test from immoderate runtime.

Model validation and evaluation

The 10-fold cross-validation is used to evaluate the classifiers presented in classifier panel module subsection and the classifier selection methods presented in classifier selector module subsection. Sensitivity, Specificity, F-score, AUC (ROC area), and mean ± standard deviation (STD) of estimates for 10-fold cross validation are considered as evaluation measures. Instead of considering precision and recall, which are more common in machine-learning tasks, this study prefers to assess how much sensitive and specific the proposed model is. In the clinical context, a more sensitive model is preferable as the cost of overlooking a positive sample is very high.[26] Also, a more specific model is required to eliminate unnecessary tests as to the cost of testing is very high.[26] It is needed to point out that Mann–Whitney statistic[27] is used to calculate ROC area.

To calculate mean ± STD of estimates for a method, in each run of 10-fold cross-validation, Eq. 6 is applied on the test samples:

$$\text{STD}(X) = \text{SQRT}([\text{SUM}([x - M] \text{ ^ } 2)]/N) \quad (6)$$

Where X is the set of estimates in which each estimate is 1 or 0, x is a test sample estimate in X, M is the mean of X and N is the number of test samples. Then, standard deviation of a method is calculated by taking the average of *STD (X)* among all the 10 runs of k-fold cross validation using Eq. 7:

$$\text{STD (Method)} = \text{STD}(\text{RUN}_i(\text{Method}))/I \quad (7)$$

Where I is the number of runs which is 10 in 10-fold cross validation, and i is 1–10. Finally, the mean ± STD for each method is calculated and reported.

Single-modality model

To compare the results of our proposed model with the situation in which the only dataset of one of the three colposcopy modalities is used, such as only one type of cervigram is investigated, this study separately applied all the seven classifiers as well as all classifier selection methods on each modality dataset. In this case, the subjective judgments have been added to the feature vector of a sample and consensus has been considered as the actual class label for that sample.

A general flowchart of applying classifiers as well as classifier selection methods on each single colposcopy modality dataset is illustrated in Figure 3. The same training-test partition as the proposed model (80–20) and 10-fold cross-validation have been used for each single-modality model.

By comparing the general flowchart of the proposed model [Figure 2] and the flowchart of the single-modality model [Figure 3], it is clear that the main difference between two models is the merger module, which is absent in the single-modality model. Therefore, a comparison that investigates the effect of merger module would be informative. This comparison is presented in comparison with single modality model subsection.

Results

Weka, which is a collection of machine-learning algorithms for data-mining tasks,[28] is used to train, test, and evaluate the proposed model as it has two important characteristics; it is a free software system and it uses ARFF files that can be easily used and modified without data format problems. The results of applying the proposed model as well as the comparison between single-modality model are introduced in single-modality model subsection, and the proposed model will be discussed in model selection and comparison with single-modality model subsections. In addition, a comparison between the proposed model and other cervical cancer detection and prediction systems will be discussed in comparison with other cervical cancer detection and prediction systems subsection.

Model selection

According to Figure 2, five aggregation functions were applied on the initial dataset, which led to the creation of five merged datasets. Then, all the seven classifiers were applied to each merged dataset. Tables 3-7 contain sensitivity, specificity, F-score, and ROC area corresponding to different aggregation strategies for different classifiers. According to the results, it can be concluded that Random Tree and Logit Boost outperform other classifiers. The best performing classifier for each aggregation strategy is depicted by bold format in each table. The average of evaluation measures is taken to choose the best performing classifier.

The test samples that are labeled by different classifiers are then fed in the classifier panel module to choose the best classifier selection method for each aggregation strategy. Table 8 contains sensitivity, specificity, F-score, and ROC area corresponding to different aggregation strategies for different classifier selection methods. The average of evaluation measures is taken to choose the best performing classifier selection method. As shown in Table 8, the best results of applying "All classifiers" method were obtained by considering F_{ps} and F_{pc} as aggregation functions, and the best results of applying "Forward Selection" method were obtained by considering F_{mvs} as the aggregation function. In addition, the best results of applying "PCA on classifiers" method were obtained in combination with both F_{pc} and F_{mvc} aggregation functions, nothing that the combination of F_{mvc} aggregation function with PCA on classifiers leads to the best results that have been achieved in this study.

Based on the results, it can be shown that considering all classifiers as members of the ensemble does not necessarily improve the performance of the model. Therefore, regarding the role of seven classifiers results, which are named as classifier vector, for classifier selection methods, it can be concluded that considering the numbers of classifiers for the ensemble part of proposed model does not guarantee an improvement in the results of the model. Instead, it is the method of weighing the elements of the classifier vector that determines the advantageous of a classifier selection method.

As shown in Figure 4, the results of the best classifier selection method for each aggregation strategy can be observed. They indicate that applying PCA on classifiers as a classifier selection method after majority voting

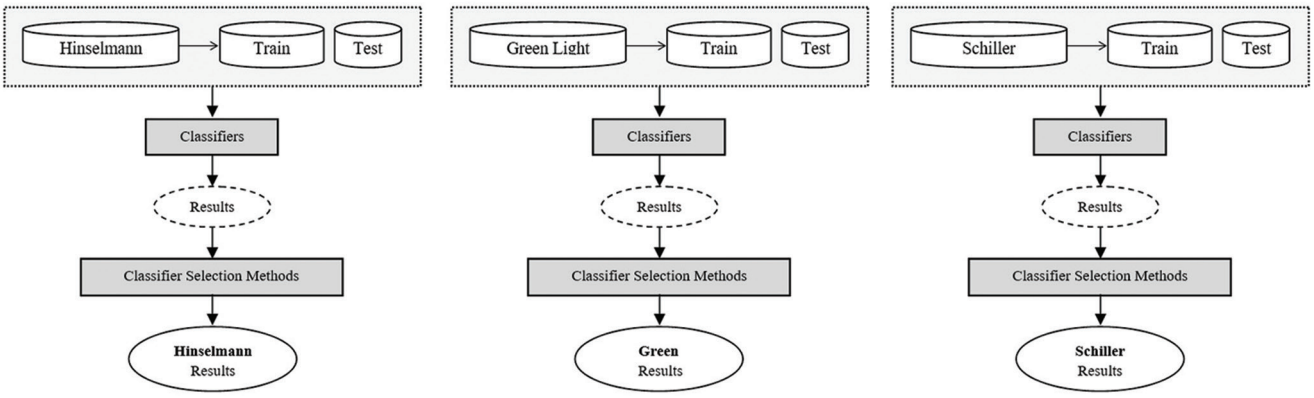

Figure 3: A general flowchart of applying the algorithm on single colposcopy modality datasets

Table 3: Results of applying seven single classifiers on DS_{mvs} dataset

Method	Sensitivity (%)	Specificity (%)	*F*-score (%)	ROC area	Mean±STD
NavieBayes	72	43	66	0.43	0.8±0.16
AdaBoost	68	32	56	0.32	0.8±0.22
Random Forest	67	35	56	0.32	0.8±0.10
Random tree	**79**	**64**	**75**	**0.68**	**0.8±0.07**
SVM	68	32	56	0.32	0.8±0.08
Decision tree	63	29	53	0.49	0.8±0.28
Logit boost	63	46	53	0.29	0.8±0.14

Random tree is the best performing classifier on the dataset acquired by applying F_{mvs} aggregation function. STD – Standard deviation; SVM – Support vector machine; ROC – Receiver operating characteristic

Table 4: Results of applying seven single classifiers on DS_{mvc} dataset

Method	Sensitivity (%)	Specificity (%)	*F*-score (%)	ROC area	Mean±STD
NavieBayes	74	52	73	0.59	0.83±0.15
AdaBoost	72	26	63	0.36	0.83±0.10
Random Forest	74	26	61	0.64	0.83±0.23
Random tree	**77**	**52**	**73**	**0.63**	**0.83±0.18**
SVM	68	24	60	0.46	0.83±0.22
Decision tree	58	21	54	0.43	0.83±0.08
Logit boost	68	24	60	0.47	0.83±0.09

Random tree is the best performing classifier on the dataset acquired by applying F_{mvs} aggregation function. STD – Standard deviation; SVM – Support vector machine; ROC – Receiver operating characteristic

Table 5: Results of applying seven single classifiers on DSps dataset

Method	Sensitivity (%)	Specificity (%)	*F*-score (%)	ROC area	Mean±STD
NavieBayes	67	43	66	0.43	0.09±0.23
AdaBoost	68	32	56	0.32	0.09±0.16
Random Forest	68	32	56	0.32	0.09±0.10
Random tree	**74**	**52**	**73**	**0.63**	**0.09±0.07**
SVM	68	24	60	0.46	0.09±0.10
Decision tree	58	21	54	0.43	0.09±0.27
Logit boost	68	24	60	0.49	0.09±0.14

Random tree is the best performing classifier on the dataset acquired by applying F_{ps} aggregation function. STD – Standard deviation; ROC – Receiver operating characteristic; SVM – Support vector machine

Table 6: Results of applying seven single classifiers on DS_{pc} dataset

Method	Sensitivity (%)	Specificity (%)	*F*-score (%)	ROC area	Mean±STD
NavieBayes	68	76	69	0.77	0.48±0.10
AdaBoost	58	52	58	0.57	0.48±0.20
Random forest	63	73	63	0.71	0.48±0.16
Random tree	63	61	64	0.62	0.48±0.22
SVM	58	70	57	0.64	0.48±0.24
Decision tree	47	46	48	0.38	0.48±0.10
Logit boost	**79**	**82**	**79**	**0.78**	**0.48±0.12**

Logit Boost is the best performing classifier on the dataset acquired by applying F_{pc} aggregation function. STD – Standard deviation; ROC – Receiver operating characteristic; SVM – Support vector machine

between consensus for building merged dataset leads to the construction of the best performing model. However, PCA on classifiers in combination with product consensuses aggregation strategy as well as FS in combination with majority voting between subjective judgments are the next possible choices for building the model.

Comparison with single-modality model

In Table 9, the results of applying different classification algorithms as well as classifier selection methods on each single-modality dataset can be observed. In single-modality model subsection, it was shown that the main difference between the proposed model and single-modality model is the

Table 7: Results of applying seven single classifiers on DS_{opc} dataset

Method	Sensitivity (%)	Specificity (%)	*F*-score (%)	ROC area	Mean±STD
NavieBayes	68	52	69	0.69	0.96±0.28
AdaBoost	58	52	58	0.57	0.96±0.15
Random forest	68	32	56	0.32	0.96±0.22
Random tree	**78**	**56**	**72**	**0.77**	**0.96±0.11**
SVM	68	24	60	0.46	0.96±0.14
Decision tree	59	21	54	0.43	0.96±0.23
Logit boost	63	29	53	0.29	0.96±0.18

Random tree is the best performing classifier on the dataset acquired by applying F_{opc} aggregation function. STD – Standard deviation; SVM – Support vector machine; ROC – Receiver operating characteristic

Table 8: Results of applying different classifier selection methods on each merged dataset (corresponding to noted aggregation function)

Method	F_{mvs}	F_{mvc}	F_{ps}	F_{pc}	F_{opc}
Best classifier (%)	79	77	74	79	77
	64	52	52	82	56
	75	73	73	79	72
	0.68	0.68	0.63	0.78	0.77
All classifiers (%)	80	73	**84**	88	78
	76	63	**73**	82	56
	75	62	**77**	80	72
	0.78	0.65	**0.70**	0.83	0.77
PCA on classifiers (%)	83	**96**	83	**93**	79
	75	**94**	67	**87**	63
	70	**91**	76	**77**	66
	0.80	**0.94**	0.70	**0.89**	0.74
Forward selection (%)	**83**	77	80	86	73
	87	63	71	67	71
	86	74	76	82	54
	0.90	0.81	0.70	0.83	0.79
Backward selection (%)	84	76	81	84	**80**
	70	63	71	66	**73**
	72	64	76	80	**64**
	0.79	0.66	0.72	0.86	**0.79**

Each cell contains sensitivity, specificity, *F*-score, and ROC area for corresponding setup. PCA – Principal component analysis; ROC – Receiver operating characteristic

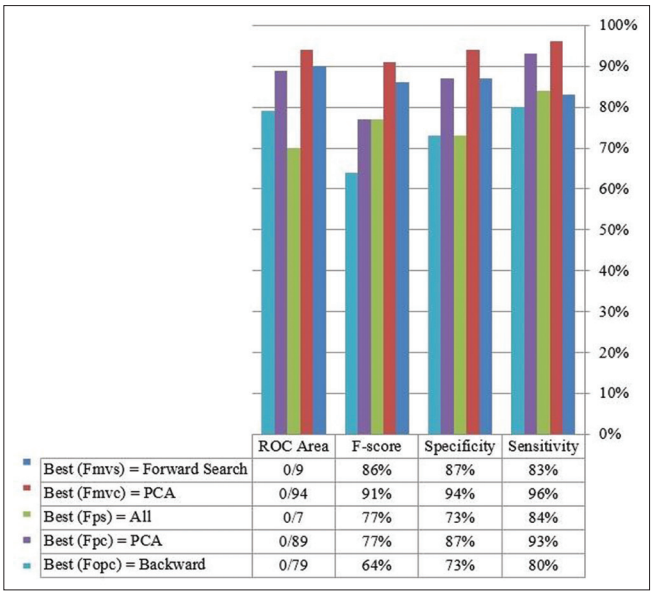

Figure 4: Results of the best classifier selection method for each aggregation strategy (corresponding to the noted aggregation function)

merger module. It is clear that the classifier selection methods enhance the results of applying single classifiers on single colposcopy modality datasets, but there is still a considerable difference between the best performing classifier selection methods and the results of the proposed model as indicated in the last line chart table of Figure 4. Therefore, the proposed model has its strength from both merger module and ensemble classifiers. In fact, the idea of applying classifier selection method for building up an effective ensemble classifier, in line with the idea of aggregating subjective judgments and consensuses, leads to acceptable results of the proposed model.

Comparison with other cervical cancer detection and prediction systems

Although the experiment has achieved the acceptable results in building a model, another important challenge is to compare the current work with other previous methods. A related work that is similar, which was using artificial intelligence to build the model and reporting the same evaluation measures, to the present study, has been reviewed. However, the majority of the previous studies applied their private datasets and reported the results in different forms as there is no standard for this process. Therefore, as shown in Table 10, the proposed approach has significantly provided better performance than the other techniques regarding the clinically important measures.

According to the literature review, previous studies that did not report the evaluation measures have been used in this study. Therefore, as a comparison, this study chose the study that at least reported one of the evaluation measures that has been used in the current study. Another issue is that several studies apply their model to private or nonimage datasets, and this made comparison much more challenging. By considering all these limitations, this study tries to find the closest works in literature.[5,13,14] The methods used in these studies are previously introduced in literature review section.

As shown in Table 10, the results of a study performed by Xu *et al*.[5] are near to the results of the best single-modality model that has been applied on the dataset. It is reasonable because Xu *et al*. used one modality of the cervix images. However, the proposed model, which uses three colposcopy modalities, outperforms the results of Xu *et al*.[5] It confirms the effectiveness of using aggregation strategies in combination with an ensemble classifier. Phoulady *et al*.[13] tackled cervical cancer classification problem by classifying cells as cancer and normal. This perspective is different to the problem of the current study, however, ROC area value that they achieved is still less than the value that has been achieved in this study. Arteta *et al*.[14] proposed a model that is applicable to different modalities by considering one-modality framework. They achieved 89% of F-score, 86.99% of precision, and 90.03% of recall. Their F-score is near to that of the current study, however, it is still lower than the currently obtained score. Their lower results confirmed the importance of considering all modalities together in achieving better results for a cervical cancer detection and prediction model. The general results showed that the proposed model is exceedingly competitive in this field, and it is effective to be used in CAD systems for cervical cancer prediction.

Table 9: Results of applying different classification algorithms on a single dataset of each modality

Method	Green (%)	Hinselmann (%)	Schiller (%)
NavieBayes	**63/56**	58/15	68/67
	64/0.69	58/0.55	68/0.67
AdaBoost	58/36	68/18	63/56
	59/0.59	64/0.76	61/0.72
Random forest	63/38	79/21	68/57
	59/0.59	70/0.75	62/0.69
Random tree	68/50	79/39	58/56
	66/0.59	76/0.59	58/0.57
SVM	63/47	90/61	58/42
	62/0.55	88/0.75	43/0.50
Decision tree	63/38	**84/78**	**74/67**
	59/0.51	**85/0.76**	**72/0.67**
Logit boost	53/33	74/20	63/60
	51/0.47	67/0.71	63/0.69
All	66/54	74/39	59/60
	63/0.65	75/0.59	58/0.57
PCA on classifiers	**83/76**	79/39	**76/67**
	78/0.78	72/0.66	**72/0.74**
Forward	63/66	**86/78**	74/67
	59/0.73	**80/0.83**	72/0.67
Backward	70/62	74/50	63/60
	63/0.69	67/0.65	60/0.62
Proposed (F_{mvc}, PCA)		**96%/94%/91%/0.94**	

Each cell contains sensitivity, specificity, *F*-score, and ROC area for the corresponding setup. PCA – Principal component analysis; SVM – Support vector machine; ROC – Receiver operating characteristic

Discussion

The ability of an artificial intelligence model in predicting the possibility of cervical cancer is imperative for decreasing the number of cervical cancer cases and mortality. The ability in this study is expressed in terms of evaluation measures including sensitivity, specificity, F-score, and ROC area that in our best configuration, the experimental results, respectively, show the values of 96%, 94%, 91%, and 0.94 for these evaluation measures. This study highlights two important aspects: first, the effectiveness of using three colposcopic images of a patient instead of using only one image on the model performance and second, the effectiveness of using an ensemble classifier instead of a single classifier for building cervical cancer prediction model. The effectiveness is maximizing the artificial intelligence model performance and is another expression of ability, which had been discussed earlier in this section. For the first aspect, the use of one colposcopy modality in the best configuration has reached to 86%, 78%, 80%, and 0.83 for sensitivity, specificity, F-score, and ROC area, which were obtained by applying forward selection in classifiers on the Hinselmann dataset. These values are lesser than the results of the best configuration of the proposed model. This fact proves that the proposed model that simultaneously uses three colposcopy modalities is more effective than the model that only uses one modality. For the second aspect, based on Table 8, it can be seen that none of the single classifiers [denoted by the best classifier as the best performing single classifier for each aggregation function in the first row of Table 8] could achieve better results than any of the ensemble classifiers, such as PCA on classifiers, forward selection, and backward selection. The best results in the situation in which no ensemble classifier is used are obtained by applying Logit Boost on DSpc dataset producing 79%, 82%, 79%, and 0.78 of sensitivity, specificity, F-score, and ROC area, which are lesser than the proposed model results. The proposed

Table 10: Comparison of cervical cancer prediction systems

Method	Sensitivity (%)	Specificity (%)	*F*-score (%)	ROC area
[5]	88%	83%	NA	0.87
[13]	NA	NA	NA	0.91
[14]	NA	NA	89	NA
Best single-modality model (%)	86	78	80	0.83
Proposed model (%)	96	94	91	0.94

NA – Not available; ROC – Receiver operating characteristic

model has been generally proven to provide at least 10%, 12%, 11%, and 0.11 progresses in sensitivity, specificity, F-score, and ROC area, which are considerably better than the results of configurations that do not use merged dataset and ensemble classifier. These results show that the idea of merging feature vectors of three colposcopy modality images and using an ensemble classifier instead of a single classifier may improve the ability and effectiveness of the proposed artificial intelligence model.

Conclusions

Colposcopy screening enables the physician to detect and diagnose cervical cancer by investigating and analyzing colposcopic images. The artificial intelligence models that use images acquired by colposcopic imaging have been underscored in the previous studies. However, there is a limited number of works that underscore the simultaneous use of different colposcopic images of a patient. Therefore, this study introduces a new approach that merges features of different cervigrams of a patient and produces an effectively merged dataset. It culminates with the formulation of a new model, which is considered as the novelty of the present study.

The experimental results of this study have shown that merging extracted features of three different colposcopy modality images and aggregating the subjective judgments and consensus of the experts have improved the quality of the cervical cancer prediction system. To build a reliable model, this study investigated different aggregation strategies and different classifier selection methods. By comparing the results obtained from different configurations of the model with a single-modality model, it can be seen that the proposed model is a more reliable system that can support clinical decision makers by providing more reliable information. It is necessary to highlight that merging feature vectors alone do not give a significant improvement in the model performance. On the contrary, the application of an aggregation strategy along with the use of an ensemble classifier has significantly improved the results. The proposed model is a robust artificial intelligence model for predicting cervical cancer, especially in terms of sensitivity and specificity that are clinically valuable evaluation measures. This improvement would increase the performance of the cervical cancer CAD system in the clinical environments. As a conclusion, this study confirms that combining features of different colposcopic images and using ensemble classifier would be advantageous for clinical decision makers.

Our study raises opportunities for future analyses on cervical cancer prediction models. As mentioned in classifier panel module subsection, this study uses seven classifiers in a classifier panel module. This limitation is due to a tradeoff between model simplicity and maximum possible values of evaluation measures. This study outlines the model simplicity and chooses seven classifiers. However, it is obvious that adding more classifiers may lead to better results. Future investigations may tackle the proposed model by adding more classifiers and applying more classifier selection methods. Furthermore, another opportunity for future researches would be extending the proposed model for other types of cancer in which patient-related data from different sources are available.

Conflicts of interest

There are no conflicts of interest.

References

1. Papillomavirus H. Related Cancers. WORLD. Summary Report Update; 2010.
2. Fernandes K, Cardoso JS, Fernandes J. Temporal segmentation of digital colposcopies. In: Iberian Conference on Pattern Recognition and Image Analysis. IbPRIA 2015, Spain: Springer; 2015.
3. Vassilakos P, Petignat P, Boulvain M, Campana A. Primary screening for cervical cancer precursors by the combined use of liquid-based cytology, computer-assisted cytology and HPV DNA testing. Br J Cancer 2002;86:382-8.
4. Lorincz AT. Cancer diagnostic classifiers based on quantitative DNA methylation. Expert Rev Molec Diagn 2014;14:293-305.
5. Xu T, Zhang H, Xin C, Kim E, Long LR, Xue Z, *et al*. Multi-feature based Benchmark for cervical dysplasia classification evaluation. Pattern Recognit 2017;63:468-75.
6. Akbar S, Hayat M, Iqbal M, Jan MA. IACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space. Artif Intell Med 2017;79:62-70.
7. Soares F, Becker K, Anzanello MJ. A hierarchical classifier based on human blood plasma fluorescence for non-invasive colorectal cancer screening. Artif Intell Med 2017;82:1-0.
8. Wei L, Wan S, Guo J, Wong KK. A novel hierarchical selective ensemble classifier with bioinformatics application. Artif Intell Med 2017;83:82-90.
9. Duncan ID. Cervical screening. Obstet Gynaecol 2004;6:93-7.
10. Fusco E, Padula F, Mancini E, Cavaliere A, Grubisic G. History of colposcopy: A brief biography of Hinselmann. J Prenatal Med 2008;2:19.
11. Ashfaq R, Bartholomew D, Flowers L, Garcia F, Padilla L, Solomon D, O'Connor D. The CERVIX: Colposcopy of the Uterine Cervix.
12. Tumer K, Ramanujam N, Ghosh J, Richards-Kortum R. Ensembles of radial basis function networks for spectroscopic detection of cervical precancer. IEEE Trans Biomed Eng 1998;45:953-61.
13. Phoulady HA, Chaudhury B, Goldgof D, Hall LO, Mouton PR, Hakam A, *et al*. Experiments with Large Ensembles for Segmentation and Classification of Cervical Cancer Biopsy Images. Systems, man and Cybernetics (SMC). 2014 IEEE International Conference on, IEEE; 2014.
14. Arteta C, Lempitsky V, Noble JA, Zisserman A. Learning to Detect cells Using Non-Overlapping External Regions. International Conference on Medical Image Computing and Computer-Assisted Intervention. MICCAI 2012, France,: Springer; 2012.
15. Sarwar A, Sharma V, Gupta R. Hybrid ensemble learning technique for screening of cervical cancer using Papanicolaou smear image analysis. Personal Med Univ 2015;4:54-62.

16. Pfohl S, Triebe O, Marafino B. Guiding the Management of Cervical Cancer with Convolutional Neural Networks; 2017.
17. Fernandes K, Cardoso JS, Fernandes J. Quality Assessment of Digital Colposcopies Data Set. UCI Machine Learning Repository; 2017.
18. Catto JW, Linkens DA, Abbod MF, Chen M, Burton JL, Feeley KM, *et al*. Artificial intelligence in predicting bladder cancer outcome: A comparison of neuro-fuzzy modeling and artificial neural networks. Clin Cancer Res 2003;9:4172-7.
19. Lisboa PJ, Taktak AF. The use of artificial neural networks in decision support in cancer: A systematic review. Neural Networks 2006;19:408-15.
20. Antal B, Hajdu A. An ensemble-based system for automatic screening of diabetic retinopathy. Knowled Based Syst 2014;60:20-7.
21. Fernandes K, Cardoso JS, Fernandes J. Transfer Learning with Partial Observability Applied to Cervical Cancer Screening. Iberian Conference on Pattern Recognition and Image Analysis. IbPRIA 2017, Spain: Springer; 2017.
22. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2011;2:27.
23. West J, Ventura D, Warnick S. Spring research presentation: A theoretical foundation for inductive transfer. Brigham Young University, College of Physical and Mathematical Sciences. 2007;1.
24. Trybulec WA. Pigeon hole principle. Formalized Mathem 1990;1:575-9.
25. Ruta D, Gabrys B. Classifier selection for majority voting. Inform Fusion 2005;6:63-81.
26. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ Br Med J 1994;308:1552.
27. Mason SJ, Graham NE. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Q J Royal Meteorol Soc 2002;128:2145-66.
28. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: An update. ACM SIGKDD Explorat Newslett 2009;11:10-8.

BIOGRAPHIES

Elham Nikookar has received her B.Sc. degree in computer engineering from Shahid Chamran University, Ahvaz, Iran in 2008, and her M.Sc. degree in computer engineering from University of Tehran, Tehran, Iran, in 2012. She is currently an instructor and a faculty member in Computer Engineering Department of Shahid Chamran University, Ahvaz, Iran. Her research interests are Bioinformatics, medical data mining and pattern recognitions.

Email: e.nikookar@scu.ac.ir

Ebrahim Naderi has received his B.Sc. degree in computer engineering from Shahid Chamran University, Ahvaz, Iran in 2008, and his M.Sc. degree in computer engineering from Shahid Chamran University, Ahvaz, Iran in 2011. He is currently an instructor in Computer Engineering Department of University of Applied science and technology. His research interests are mainly image processing, data mining and deep learning.

Email: naderi.cse@gmail.com

Ali Rahnavard is a computational biologist and an assistant professor of Biostatistics and Bioinformatics at the George Washington University. He earned a Ph.D. in computer science, applied statistics, and bioinformatics at New Mexico State University in 2014. Rahnavard completed postdoctoral work in the biostatistics department at Harvard T.H. Chan School of Public Health and the Infectious Disease and Microbiome Program at the Broad Institute in 2018. He holds a master's degree in computer engineering/software systems from Shiraz University in Iran and a bachelor's degree in computer engineering from Razi University of Kermanshah in Iran.

Email: rahnavard@gwu.edu