EAI Endorsed Transactions
on _____ Research Article EAI.EU

Discovering Optimal Algorithm to Predict Diabetic Retinopathy using Novel Assessment Methods

Shiva Shankar Reddy[1,*], Nilambar Sethi[2], R. Rajender[3]

[1] Research Scholar, Department of Computer Science and Engineering, Biju Patnaik University of Technology, Rourkela, Odisha
[2] Department of Computer Science and Engineering, GIET, Gunupur, Odisha, India
[3] Department of Computer Science and Engineering, LENDI Engineering College, Vizianagaram, India

Abstract

Diabetic retinopathy is a diabetes complication that effects eyes. It disrupts the vasculature of the sensitive tissue present at the back of the eye. If this complication is untreated it may lead to blindness. The aim of this work is to train a model that efficiently predicts diabetic retinopathy. Machine learning techniques like Decision tree, Random forest, Adaptive boosting and Bagging are used as primary algorithms to train predictive models. An algorithm namely 'Support Vector Machine using Gaussian kernel for retinopathy prediction' is proposed in this work. The proposed algorithm is compared with the primary algorithms based on five evaluation metrics namely accuracy, Youden's J index, concordance, Somers' D statistic and balanced accuracy. From the results obtained the proposed algorithm obtained better values for all considered evaluation metrics. Thus the use of SVM with Gaussian kernel is proposed to be used for prediction of diabetic retinopathy.

Keywords: Diabetic retinopathy, random forest, decision tree, adaptive boosting, bagging, support vector machine (SVM) using Gaussian kernel (GK), accuracy, Youden's J index, concordance, Somers' D statistic and balanced accuracy.

Received on 27 March 2020, accepted on 28 June 2020, published on 01 July 2020

doi: 10.4108/_____

*Corresponding author. Email: shiva.shankar591@gmail.com

1. Introduction

Diabetic retinopathy (DR) is one of the complications of diabetes that cause blindness and vision loss impairment. It is the side effect triggered by both type-1 and type-2 diabetes. The diabetic retinopathy cannot be detected in early stage but can be identified later based on symptoms like blurred vision, vision loss, fluctuating vision, colour vision impairment and spots floating in the vision. [1]

Diabetic retinopathy affects the retina in many ways which include abnormal growth of blood vessels and problems related to vision like blindness, retinal detachment, glaucoma and vitreous hemorrhage. The risk factors for DR include high blood pressure, long term diabetes, high cholesterol and smoking [1]. If any person is suffering from diabetes type-1 or type-2 the changes in their vision should be observed to avoid diabetic retinopathy in future.

In a survey it was stated that one out of 15 and one out of 45 people are having blindness and visually impairments respectively due to glaucoma. The survey concluded that 2.1 million and 4.2 million persons are suffering from blindness and vision loss impairments respectively in 2010 [2]. DR has some typical lesions that include micro aneurysms (MA), hemorrhages and hard exudates.

In some hospital based studies it is stated that over 34.6% of diabetic retinopathy patients are having diabetes [3]. In Nepal the presence of diabetes among the persons who are aged above 20 years was 40%, 40 years and above was 19%. The survey based analysis conducted worldwide from 1980 to 2008 includes 35 studies. This survey concluded that the presence of diabetic and proliferate retinopathy was 35.4% and 7.5% respectively. [4]

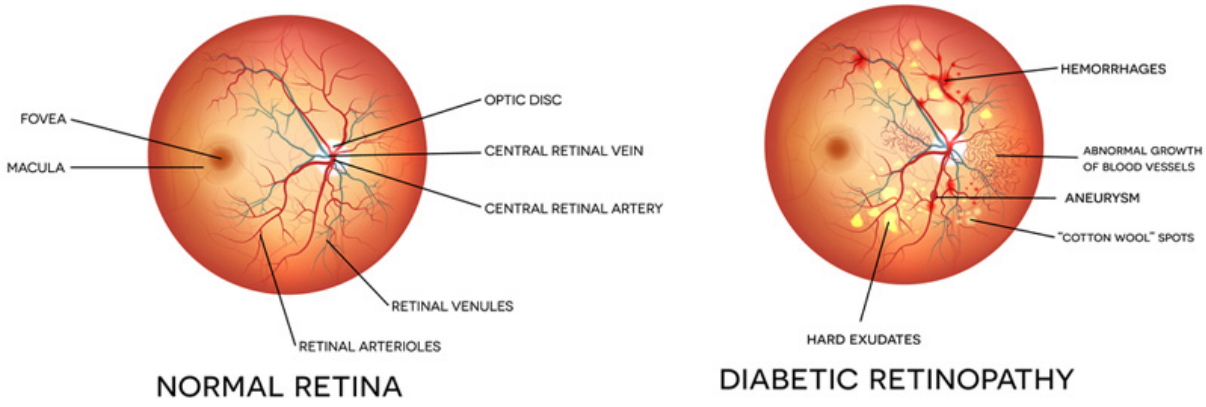

Figure 1. Normal retina and Diabetic retina

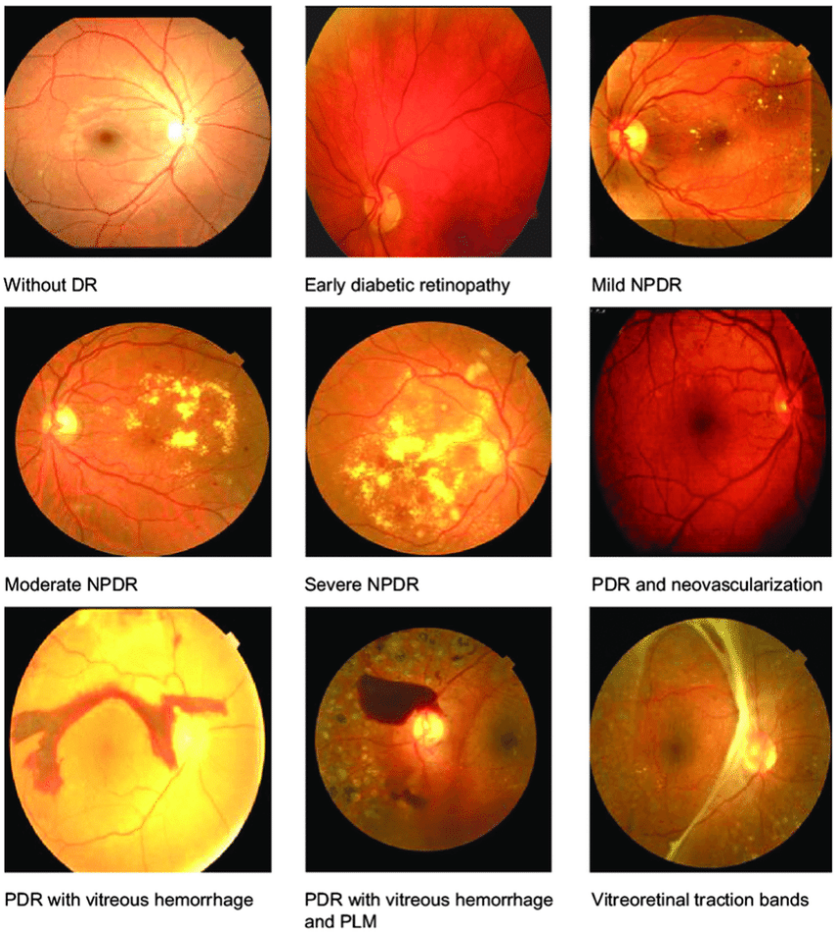

Figure 2. Stages of diabetic retinopathy

DR leads to the damage of neurons and minute blood vessels in the retina. It will cause the loss of blood from the eye that will cause swelling of eyes as shown in Figure 1. There are some physical diagnosis methods namely fluorescein angiography and optical coherence tomography to detect DR. The treatment is done only after observing retinal image by applying fluids or dyes on the patient eye.

Diabetic Retinopathy have five stages namely 0, 1, 2, 3 and 4 where a doctor can determine diabetic retinopathy by observing the presence of lesions which were only related to the abnormalities that occurs in vascular. Different stages are illustrated in Figure 2. As the diabetic patients are increasing day by day the infrastructure needed for detection of DR should also be high. Some former efforts have made a good significance by using pattern recognition, image classification and machine learning (ML) techniques to predict the disease.

In this work machine learning algorithms namely random forest, adaptive boosting, bagging and decision tree are implemented for prediction of diabetic retinopathy. A proposed algorithm namely SVM with Gaussian kernel is also implemented to predict diabetic retinopathy more accurately using R programming.

- No DR (0) indicates absence of diabetic retinopathy.
- Mild (1) indicates the presence of DR but it is mild and non-proliferative DR.
- Moderate (2) specifies the presence of DR where the complication is moderate and it is non proliferative DR.
- Severe (3) specifies the existence of DR where the complication is severe and it is non proliferative too.
- Proliferative (4) indicates the existence of DR where the complication is very high and it is a proliferative DR. [5]

A doctor can determine diabetic retinopathy by observing the presence of lesions which were only related to the abnormalities that occurs in vascular. As the diabetic patients are increasing day by day the infrastructure needed for detection of DR should also be high. Some former efforts have made a good significance by using pattern recognition, image classification and machine learning techniques for prediction of disease.

In this work machine learning algorithms namely random forest, adaptive boosting, bagging and decision tree are implemented for prediction of diabetic retinopathy. A proposed algorithm namely SVM with Gaussian kernel is also implemented to predict diabetic

2. Literature Survey

Wang et al. [6] focused on diagnosing diabetic retinopathy. Diabetic retinopathy (DR) analysis approaches in the literature are regularly criticized as being limit in detecting DR-related features or being absence of interpretability. They evaluated the excellence of annotations in DR grading by measuring inter-grader inconsistency.

Dai [7] highlighted his work on detecting micro aneurysm (MA) to prevent the vision-loss impairments occur due to diabetic retinopathy. The existing methods fail to face the large and small intra class variations to detect the funds image. The clinical report engaged to fill the gaps between low level and high level visual features and MA detect from high level image features. He used performance measures like precision and recall. And it is easy to detect multiple lesions in funds images.

Leeza and Farooq [8] used bag of features model to detect diabetic retinopathy. They used support vector machine with radial basis kernel and neural network techniques to categorize the pictures into five modules they are normal, mild, moderate, severe non-proliferative diabetic retinopathy and proliferative diabetic retinopathy. They considered segmentation for data pre-processing and removal, collection, and cataloguing of features as post pre-processing stages.

Costa et al. [9] focused on detection of diabetic retinopathy based on retinal images. They trained a model that correctly identifies DR depending on the occurrence of various retinal lesions. They proposed a procedure constructed on the multiple instance learning to overcome the requirement by exploiting the implicit evidence in the annotations at the level of image.

Dinesh Pandey et al. [10] mainly focused on segmenting thick and thin blood vessels in retina. Four databases namely DRIVE, STARE, CHASE_ DB1 and HRF were used to assertaining the performance of proposed method. Local phase-preserving denoising, line detection, local normalization and maximum entropy thresholding techniques were used for detecting thin blood vessels. The technique used for detecting thick blood vessels is maximum entropy thresholding. It was concluded that the proposed technique performed efficiently based on specificity, accuracy, sensitivity, Matthews Correlation Coefficient (MCC) and AUC.

Pires et al. [11] the members of IEEE in their research beyond lesion based diabetic retinopathy which was done in 2017 have stated that DR leads to blindness. They also stated that when the DR is identified in correct time the loss occurred by DR is negotiable and can be cured by taking certain measures. They have used Bossanova and Fisher vector for lesion based detection. In their research they concluded that the automatic detection of DR can be useful for that which patients can be referred to doctor and reduce the classification error by 40%. There a current method in automatic detection but it is too much dependent on individual lesion detections. They proposed the process of detection through three steps. They are detection of individual DR lesions, fusion of the lesion responses and referability decision.

Rafiqul Islam et al. [12] mined data from social networks to detect depression. The data is collected from facebook users. Some ML techniques namely KNN, SVM, ensemble and decision tree were used in their work. These techniques are implemented based on emotion, linguistic style, temporal process and all three features combined

(emotional, linguistic and temporal). By comparing the results obtained in these four procedures the decision tree algorithm has performed better than remaining techniques.

Hui Zheng et al. [13] proposed a fuzzy association rule mining based on dynamic optimization (DOFARM) technique. They used this method for obtaining sentiment strength as positive or negative (i.e. emotional and sentimental computing). A dual compromise scheme is developed which comprises of first trade-off and second trade-off. In the first trade-off different metrics of fuzzy association rules are balanced in order to improve performance. In the second trade-off the performance of DOFARM method is improved by balancing accuracy and effectiveness. They compared DOFARM and the other fuzzy association rule mining techniques in terms of accuracy and effectiveness and concluded that DOFARM has better performance.

Jiahua Du et al. [14] considered data from twitter to detect hay fever. The dataset is a text dataset. They proposed a deep learning architecture namely neural networks technique with character embedding and attention mechanism. The neural network (NN) considered was bidirectional Long Short-Term Memory (LSTM). They considered two models. The first model is the combination of bidirectional LSTM and attention mechanism and the second one is the combination of bidirectional LSTM, character embedding and attention mechanism. The accuracies obtained for first and second techniques are 77.72% and 79.51% respectively.

Jinyuan He et al. [15] mainly focused on classifying heartbeat using ECG records. Two databases namely MIT-BIH-AR and INCART 12-leads arrhythmia are considered for their work. To improve the performance they proposed a pyramid like model for classifying heartbeat classification. The performance metrics considered for evaluation are accuracy, sensitivity and positive predicted value. They concluded that the proposed technique had better performed than the rivals.

Iftikhar Naseer et al. [16] proposed Mamdani fuzzy inference expert system for diagnosing heart disease. They provided input fields for the features like age, pain in chest, electrocardiography, cholesterol, high blood pressure and diabetes. Fuzzy rules are formulated for these features. Based on these fuzzy rules the input data is classified as negative or border line or positive or strongly positive. The proposed technique has performed better with an accuracy of 94%.

Dinesh Pandey et al. [17] considered problem of segmenting region of interest i.e. breast. They also considered breast density in MRI more accurately. They proposed a methodology with three steps. In first step they used adaptive wiener filtering and *k*-means clustering for reducing noise, maintain edges and eliminate undesirable artefacts. In step 2 they used a contour based level sets for excluding the heart area. Here they determined the initial points by using convolution method and maximum entropy thresholding. In third step they used morphological operations and local adaptive thresholding in order to remove pectoral muscle. The evaluation metrics considered are accuracy, sensitivity, precision, specificity, AUC, misclassification rate, jaccard coefficient and dice similarity coefficient.

Iqbal Sarker et al. [18] developed a model for providing eHealth services for diabetes patients. The optimal k-nearest neighbor technique is used for diabetes mellitus prediction and analysis. The data from 500 patients is considered. The performance metrics considered are precision, recall, f-measure and area under ROC curve. They compared the traditional KNN with optimal KNN technique and stated that optimal KNN has obtained better performance metrics. Then optimal KNN technique is compared with some existing algorithms namely AdaBoost, logistic regression, naive bayes, decision tree and SVM. They concluded that optimal KNN has performed better than existing algorithms.
Mansour [19] focused on complication of diabetes which is diabetic retinopathy. The majority of the existing models perform the analysis of diabetic retinopathy CAD systems. In his work he revealed that evolutionary computing plays a vital for optimising DR-CAD and pre-processing the image and dimensional reduction & classification.

Zhu et al. [20] proposed their work related to diabetes type-2 and diabetic retinopathy. They have conducted a survey on the patients who are suffering with diabetes in China. In this study they performed detection of DR on the basis of retinal photographs. They have named R0, R1, R2 and R3 based on the severity of the DR. They used a quantitative method for determining the global tortuosity of retinal arterioles. In this work they have used linear and logistic regressions and they have compared those retinal images with those who don't have DR.

Kar and maity [21] highlighted their work on detection of DR using retinal lesions. They have stated that the DR is a microvascular side effect of diabetes at first it is asymptomatic and later this tips to mild blindness and vision blurred and sometimes it may lead to death in some cases. In this work they used Gaussian and Matched Filtering. Matched filtering with gaussian kernel yields high response to lesion detection of the candidate.

Zeng et al. [22] had proposed that DR is the most effective and it is not detected in early stages of Diabetes and time consuming process for Diagnostic procedure. They trained a conventional NN with architecture similar to Siamese with a transfer learning procedure. They concluded that accept the input and to learn the correlation for prediction of DR.

Sun and Zhang [23] used electronic health records for diagnosing and analysing DR. They have taken the data from the Medical Big Data center which was taken from the 301 hospitals during 4 years period of time. They have replaced the missing values of the demographics of the data taken and ID mapping and classification of the data. In this work they have used some algorithms namely Support Vector module, Logistic regression, Decision tree, Random forest and naive bayes to analyze DR. In this work they have concluded that the machine learning technique random forest has obtained the highest accuracy with 92%. They have also stated by their model that cost is low and the accuracy is higher than normal DR detection technique as

the modern world people are more conscious about their convenience.

Deeksha et al. [24] compared various classification techniques to predict different stages of diabetic retinopathy. They used Messidor dataset extracted from UCI repository. The data is classified as two classes based on fundus image of eye. They used binary particle swarm optimization (BPSO) for feature selection. The algorithms used for prediction are decision trees with bagging and boosting techniques, weighted k-nearest neighbor, subspace discriminant analysis and support vector machine. From result analysis they observed that subspace discriminant analysis and boosting has obtained highest accuracy.

Karan et al. [25] used some classification algorithms to diagnose diabetic retinopathy. They used Messidor dataset for implementing algorithms. The dataset contains attributes related to optic disc diameter, lesion especially like micro aneurysms and exudates, image level like pre-screening, AM/FM and quality assessment. The algorithms used are KNN, pattern classifier, decision tree, adaptive boosting, naive bayes, random forest and SVM. They used ensemble technique of these algorithms. They used forward search and backward search to obtain best ensemble technique.

Wen cao et al. [26] used principal component analysis and machine learning to detect micro aneurysms. They used DIARETDB1 dataset. It contains 25 X 25 pixel patches which are obtained from the fundus pictures. They used Principal Component Analysis (PCA) to reduce dimensionality of input data. The algorithms used are random forest, NN and SVM. They implemented all the algorithms using leave-out cross validation technique. They stated that compared to a deep learning technique the implemented algorithms has obtained better AUC and F-Measure values.

This chapter i.e. Chapter 2 includes the work done related to diabetic retinopathy. In chapter 3 the methodologies used in this work are provided. It includes explanation regarding dataset, data pre-processing, system architecture of the developed model and data visualization. In chapter 4 the proposed work is provided which includes the brief explanation of four algorithms used and description of the proposed algorithm. In chapter 5 the results obtained are provided for all the algorithms which also include the analysis of results. Chapter 6 contains conclusion based on the result analysis.

3. Methodology

3.1. Objectives of the work

Most of the diabetic patients are affected by Diabetic retinopathy as a side effect. The main problem of retinopathy is that this disease may lead to blindness if not diagnosed early. The machine learning techniques have been effectively used in the medical field to detect or predict diseases including retinopathy. But the main problem lies in selecting of optimal algorithm for prediction of diabetic retinopathy early. This problem is handled in this work by proposing a machine learning kernel method namely SVM with Gaussian kernel to predict diabetic retinopathy early. The objectives of this work to accomplish the proposal are

- To implement machine learning methods to predict diabetic retinopathy. SVM with Gaussian kernel is proposed for predicting diabetic retinopathy. Random forest, Decision tree, Adaptive boosting and Bagging are algorithms considered for comparison with proposed algorithm. These algorithms are implemented using R programming.
- To obtain best performing algorithm among all the considered algorithms. The proposed algorithm SVM with Gaussian kernel performed more accurately when compared to other algorithms. The comparison is done using metrics like accuracy, youden's j index, concordance, somers d statistic and balanced accuracy.

3.2. Dataset

The dataset was created based upon different types of retinal images and then extracted into attributes. Also some attributes related to diabetes and hypertension were added to the data extracted from UCI repository. The entire data contains 24 attributes with 1151 instances based on which it is predicted whether the patient is having diabetic retinopathy or not. There are 23 predictor attributes and one target attribute. The training dataset consists of 80% of the dataset and test dataset contains remaining 20% of the dataset. The attributes in the dataset are named as q, ps, nma.a-nma.f, nex.a-nex.h, dd, dm, am/fm, fa.glucose, po.pra.glucose, SBP, DBP and class. The attribute class is the target attribute having values 1 and 0 for positive and negative. All the 24 attributes are described in the table 1.

Table 1. Attributes in the dataset

Attribute	Description
q	This attribute represents Quality assessment. It has binary values 1 and 0. The value 0 indicates poor quality and 1 indicates sufficient quality.
ps	This attribute called Pre Screening represents abnormality of retina. It has binary values 1 and 0. The value 0 indicates lack of abnormality and1 indicates severe abnormality.
nma.a-nma.f	These six attributes represents values obtained by detection of micro aneurysms or micro aneurism (MA) detection. Each feature value tells the number of micro aneurism found at confidence levels after MA detection.
nex.a-nex.h	These eight attributes represent values of exudates. They contain same information as nma.a–nma.f for exudates.These options are normalized by dividing the

[10] Pandey D, Yin X, Wang H, Zhang Y. Accurate vessel segmentation using maximum entropy incorporating line detection and phase-preserving denoising. Computer Vision and Image Understanding. 2017 Feb 1; 155:162-72.
[11] Pires R, Avila S, Jelinek HF, Wainer J, Valle E, Rocha A. Beyond lesion-based diabetic retinopathy: a direct approach for referral. IEEE Journal of Biomedical and Health Informatics. 2017; 21(1):193-200.
[12] Islam MR, Kabir MA, Ahmed A, Kamal AR, Wang H, Ulhaq A. Depression detection from social network data using machine learning techniques. Health information science and systems. 2018 Dec 1; 6(1):8.
[13] Zheng H, He J, Huang G, Zhang Y, Wang H. Dynamic optimisation based fuzzy association rule mining method. International Journal of Machine Learning and Cybernetics. 2019 Aug 1; 10(8):2187-98.
[14] Du J, Michalska S, Subramani S, Wang H, Zhang Y. Neural attention with character embeddings for hay fever detection from twitter. Health information science and systems. 2019 Dec 1; 7(1):21.
[15] He J, Le Sun JR, Wang H, Zhang Y. A pyramid-like model for heartbeat classification from ECG recordings. PloS one. 2018; 13(11).
[16] Naseer I, Khan BS, Saqib S, Tahir SN, Tariq S, Akhter MS. Diagnosis Heart Disease Using Mamdani Fuzzy Inference Expert System. EAI Endorsed Transactions on Scalable Information Systems.2020; 7(26).
[17] Pandey D, Yin X, Wang H, Su MY, Chen JH, Wu J, Zhang Y. Automatic and fast segmentation of breast region-of-interest (ROI) and density in MRIs. Heliyon. 2018 Dec 1; 4(12):e01042.
[18] Sarker IH, Faruque MF, Alqahtani H, Kalim A. K-Nearest Neighbor Learning based Diabetes Mellitus Prediction and Analysis for eHealth Services. EAI Endorsed Transactions on Scalable Information Systems. 2020; 7(26):e4
[19] Mansour RF. Evolutionary computing enriched computer-aided diagnosis system for diabetic retinopathy: a survey. IEEE reviews in biomedical engineering. 2017 May 17; 10:334-49.
[20] Zhu S, Liu H, Du R, Annick DS, Chen S, Qian W. Tortuosity of Retinal Main and Branching Arterioles, Venules in Patients With Type 2 Diabetes and Diabetic Retinopathy in China. IEEE Access. 2020 Jan 3; 8:6201-8.
[21] Kar SS, Maity SP. Automatic detection of retinal lesions for screening of diabetic retinopathy. IEEE Transactions on Biomedical Engineering. 2017 May 24; 65(3):608-18.
[22] Zeng X, Chen H, Luo Y, Ye W. Automated diabetic retinopathy detection based on binocular Siamese-like convolutional neural network. IEEE Access. 2019 Mar 5; 7:30744-53.
[23] Sun Y, Zhang D. Diagnosis and Analysis of Diabetic Retinopathy based on Electronic Health Records. IEEE Access. 2019 May 23; 7:86115-20.
[24] Karanjgaokar D, Prasanna BK, Chaurasiya RK. Comparison of classification methodologies for predicting the stages of diabetic retinopathy. Proceedings of the 2017 International Conference on Intelligent Sustainable Systems (ICISS); 2017 Dec 7; IEEE; 2018 Jun. p. 700-705.
[25] Bhatia K, Arora S, Tomar R. Diagnosis of diabetic retinopathy using machine learning classification algorithm. Proceedings of the 2nd International Conference on Next Generation Computing Technologies (NGCT); 2016 Oct 14; IEEE; 2017 Mar. p. 347-351.
[26] Cao W, Czarnek N, Shan J, Li L. Microaneurysm detection using principal component analysis and machine learning methods. IEEE transactions on nanobioscience. 2018 May 24; 17(3):191-8.
[27] Shah SA, Aziz W, Arif M, Nadeem MS. Decision trees based classification of cardiotocograms using bagging approach. Proceedings of the 13th International Conference on Frontiers of Information Technology (FIT); 2015 Dec 14; IEEE; 2016 Feb. p. 12-17.
[28] Bjurgert J, Valenzuela PE, Rojas CR. On Adaptive Boosting for System Identification. IEEE transactions on neural networks and learning systems. 2017 Oct 12; 29(9):4510-4.