Informatics in Medicine Unlocked 16 (2019) 100189

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Classification of malignant and benign tissue with logistic regression

Laila Khairunnahar[a,*], Mohammad Abdul Hasib[b], Razib Hasan Bin Rezanur[b], Mohammad Rakibul Islam[b], Md Kamal Hosain[a]

[a] Department of Electronics & Telecommunication Engineering, Rajshahi University of Engineering & Technology, Rajshahi, 6204, Bangladesh
[b] Department of Electrical and Electronic Engineering, Islamic University of Technology, Gazipur, 1704, Dhaka, Bangladesh

ARTICLE INFO

Keywords:
Breast cancer detection
Machine learning classifier
Logistic regression
Performance evaluation tests
Weighted sigmoid function

ABSTRACT

Detection of breast cancer is the preliminary phase in cancer diagnosis. So, classifiers with higher accuracy are always desired. A classifier with high accuracy offers very less chance to wrongly classify a patient of cancer. This research investigates the performance of a modified and improved version of the hypothesis used in the logistic regression. Both gradient descent and advanced optimization techniques are used for the minimization of the cost function. A weighting factor of β is assigned in the hypothesis which is a sigmoid function. The dependency of the weighting factor to the number of features, the size of the dataset and the type of optimization technique used are observed. The accuracy of breast cancer detection is improved significantly by appropriately choosing the value of β, which, is a function of both the number of features and the type of optimization techniques used. The obtained results are promising by providing a significant increment in accuracy, sensitivity, and specificity.

1. Introduction

Breast cancer refers to cancer from a malignant tumor in the cells of the breast tissue. A malignant tumor is a group of cancer cells that can grow into surrounding tissues or spread to distant areas of the body [1]. Breast cancer is uncontrolled multiplication of cells in breast tissue. A group of rapidly dividing cells may form a lump or architectural distortions. The second leading cause of death among women is breast cancer, as it comes directly after lung cancer. Breast cancer is a life taking disease and early detection can certainly reduce the rate of mortality. An analysis of the most recent data has shown that the survival rate is 88% after 5 years of diagnosis and 80% after 10 years of diagnosis [2].

Machine learning classifiers are very popular for detecting breast cancer. Several research works have been done in this area. Here a classifier algorithm named “Logistic Regression” has been modified to detect the malignancy or benignancy of the tumorous cell more accurately.

In 2015, an estimated 231,840 new cases of invasive breast cancer are diagnosed among women in the U.S. along with 60,290 new cases of non-invasive (in situ) breast cancer [3]. For women in the U.S. breast cancer death rates are higher than those for any other cancer, besides lung cancer. Worldwide breast cancer is the leading type of cancer in women, accounting for 25% of all cases. It is more common in developed countries and is more than 100 times more common in women than in men [4]. Outcomes for breast cancer vary depending on the cancer type, extent of disease, and person's age [5]. Survival rates in the developed world are high with between 80% and 90% of those in England and the United States alive for at least 5 years [6,7]. Lung cancer is the leading cancer site in males, comprising 17% of the total new cancer cases and 23% of the total cancer deaths. Breast cancer is now one of the leading cause of cancer death among females in economically developing countries [8]. Thus, a shift from the past decade during which the most common cause of cancer death was cervical cancer is visible. Furthermore, the mortality burden for lung cancer among females in developing countries is as high as the burden for cervical cancer, with each accounting for 11% of the total female cancer deaths. Although overall cancer incidence rates in the developing world are half those seen in the developed world in both sexes, the overall cancer mortality rates are generally similar. In Portugal, each year 4500 new cases of breast cancer are diagnosed and 1600 women are estimated to die from this disease [9].

In 2018, 1,735,350 new cancer cases and 609,640 cancer deaths are projected to occur in the United States. Over the past decade of data, the cancer incidence rate (2005–2014) was stable in women and declined by approximately 2% annually in men, while the cancer death

* Corresponding author.
E-mail addresses: laila.ete11ruet@gmail.com (L. Khairunnahar), tamimeee@iut-dhaka.edu (M.A. Hasib), razib1325@gmail.com (R.H.B. Rezanur), rakibul@iut-dhaka.edu (M.R. Islam), kamaleteruet@gmail.com (M.K. Hosain).

https://doi.org/10.1016/j.imu.2019.100189
Received 6 August 2018; Received in revised form 21 April 2019; Accepted 5 May 2019
Available online 14 May 2019

rate (2006–2015) declined by about 1.5% annually in both men and women. The combined cancer death rate dropped continuously from 1991 to 2015 by a total of 26%, translating to approximately 2,378,600 fewer cancer deaths than would have been expected if death rates had remained at their peak. Of the 10 leading causes of death, only cancer declined from 2014 to 2015. In 2015, the cancer death rate was 14% higher in non-Hispanic blacks (NHBs) than non-Hispanic whites (NHWs) overall (death rate ratio [DRR], 1.14; 95% confidence interval [95% CI], 1.13–1.15), but the racial disparity was much larger for individuals aged < 65 years (DRR, 1.07; 95% CI, 1.29–1.32) compared with those aged ≥ 65 years (DRR, 1.07; 95% CI, 1.06–1.09) and varied substantially by state. For example, the cancer death rate was lower in NHBs than NHWs in Massachusetts for all ages and in New York for individuals aged ≥65 years, whereas for those aged <65 years, it was 3 times higher in NHBs in the District of Columbia (DRR, 2.89; 95% CI, 2.16–3.91) and about 50% higher in Wisconsin (DRR, 1.78; 95% CI, 1.56–2.02), Kansas (DRR, 1.51; 95% CI, 1.25–1.81), Louisiana (DRR, 1.49; 95% CI, 1.38–1.60), Illinois (DRR, 1.48; 95% CI, 1.39–1.57), and California (DRR, 1.45; 95% CI, 1.38–1.54). Larger racial inequalities in young and middle-aged adults probably partly reflect less access to high-quality health care [10].

Breast cancer is also the most common cancers among Egyptian women as it represents 18.3% of the total general of cancer cases in Egypt and a percentage of 37.3% of breast cancer is considered treatable disease. The age of breast cancer affection in Egypt and Arab countries is ten years prior compared to other countries as the disease targets women in the age of 30 in Arab countries, while affecting women above 45 years in European countries. Breast cancer comes in the top of cancer list in Egypt by 42 cases per 100 thousand of the population. However 80% of the cases of breast cancer in Egypt are of the benign kind [11]. Worldwide this scenario is getting really hilarious day by day. Most of the cases women are the victim. But now-a-days it is also seen in men. So the breast cancer scenario worldwide is very severe. A woman's risk of breast cancer approximately doubles if she has a first-degree relative (mother, sister, and daughter) who has been diagnosed with breast cancer. About 15% of women who get breast cancer have a family member diagnosed with it [12].

It is estimated that each year 76,000 women die of breast cancer in South Asia (India, Bangladesh, Nepal, Myanmar, Pakistan, and Tibet). In Bangladesh, there is no national cancer registry. However, age-standardized incidence rates from Karachi, Pakistan (53.8/100,000) and Kolkata, India (25.1/100,000) (both with whom Bangladesh shares many cultural and historical similarities) suggest an annual incidence rate of 35–40/100,000. Therefore, in Bangladesh, we estimate an annual new breast cancer case burden of 30,000 women. It is projected that global breast cancer cases will grow from 1.4 million in 2008 to over 2.1 million cases in 2030. While countries with higher income celebrate significant progress toward curing women with breast cancer, low-income countries like Bangladesh are only beginning to recognize the extent and severity of the disease.

Early detection means using an approach that lets breast cancer get diagnosed earlier than the disease might have occurred aptly. Early detection of breast cancer can increase the rate of recovery to a great extent. Early detected breast cancer is easier to treat with fewer risks and reduces the mortality by 25% [13].

There are two major components of early detection of cancer. They are: (i) Education to promote early diagnosis & (ii) Screening. Recognizing possible warning signs of cancer and taking prompt action leads to early diagnosis. Increased awareness of possible warning signs of cancer among physicians, nurses and other health care providers as well as among the general public can have a great impact on the disease. Some early signs of cancer include lumps, sores that fail to heal, abnormal bleeding, persistent indigestion, and chronic hoarseness. Early diagnosis is particularly relevant for cancers of the breast, cervix, mouth, larynx, colon and rectum, and skin [14].

2. Preface and intuition on CAD

Computer-aided diagnosis (CAD) has become a part of the routine clinical work for detection of breast cancer on mammograms at many screening sites. This seems to indicate that CAD is beginning to be applied widely in the differential diagnosis and detection of many types of abnormalities in medical images obtained in various examinations by use of different imaging modalities. In fact, CAD has become one of the major research subjects in medical imaging and diagnostic radiology [15].

The purpose of CAD is to help Radiologists to make accurate diagnosis and provide a second opinion. For instance, CAD can minimize the operator dependent nature of Ultrasound Imaging [16]. It can obtain computational and statistical features that cannot be obtained visually. It also increases efficiency, saves time and effort.

A typical CAD application is the detection of tumors in a breast ultrasound image. Breast ultrasound CAD systems may help radiologists to evaluate ultrasound images and detect breast cancer. Such systems are used in addition to the human evaluation of the diagnosis. A breast ultrasound CAD system not only improves the ultrasound image quality, image contrast, lesion location, but also greatly reduces the human workload associated with the diagnosis, and improves the accuracy of detection and diagnosis [17]. Generally, ultrasound CAD systems for breast cancer detection involve four stages [18], as shown in Fig. 1.

3. Datasets and their features

When it comes to classification, there is a need of dataset to classify. Dataset is a statistical matrix which represents different features. It is a matrix where all the information about different features is given. Each column of the dataset represents the feature of the tumorous tissue and each row represents the number of instances. There are mainly three kinds of datasets which are mostly used in detecting the breast cancer. These are:

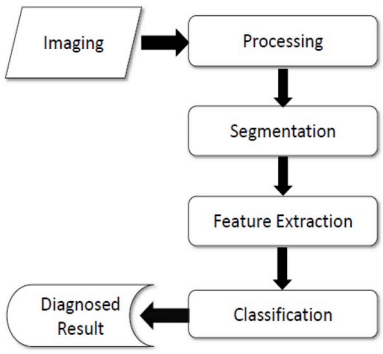

Fig. 1. A CAD system for breast cancer diagnosis.

i) Image preprocessing: The task of image preprocessing is to enhance the image and to reduce speckle without destroying the important features of BUS images for diagnosis.
ii) Image segmentation: Image segmentation divides the image into non overlapping regions and it separates the objects (lesions) from the background. The boundaries of the lesions are delineated for feature extraction.
iii) Feature extraction and selection: This step is to find a feature set of breast cancer lesions that can accurately distinguish between lesion and non-lesion or benign and malignant. The feature space could be very large and complex, so extracting and selecting the most effective features is very important.
iv) Classification: Based on the selected features, the suspicious regions will be classified into different categories, such as benign findings and malignancy. Many machine learning techniques such as linear discriminant analysis (LDA), support vector machine (SVM) and artificial neural network (ANN) have been studied for lesion classification.

Table 1
Datasets and their features.

Wisconsin diagnosis breast cancer (WDBC)	Wisconsin prognosis breast cancer (WPBC)	Wisconsin breast cancer (WBC)
The details of the attributes found in WDBC dataset [19]: ID number, Diagnosis (M = malignant, B = benign) and ten real-valued features are computed for each cell nucleus: Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry and Fractal dimension [20]. These features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image [21]. When the radius of an individual nucleus is measured by averaging the length of the radial line segments defined by the centroid of the snake and the individual snake points. The total distance between consecutive snake points constitutes the nuclear perimeter.	Details of the attributes found in WPBC dataset [14]: ID number, Outcome (R = recur, N = non-recur), Time (R = > recurrence time, N = > disease-free time), from 3 to 33 ten real-valued features are computed for each cell nucleus: Radius, Texture, Perimeter, Area, Smoothness, Compactness, and Concavity, Concave points, Symmetry and Fractal dimension.	
When the radius of an individual nucleus is measured by averaging the length of the radial line segments defined by the centroid of the snake and the individual snake points. The total distance between consecutive snake points constitutes the nuclear perimeter.	The thirty four is Tumor size and the thirty five is the Lymph node status.	In the clump thickness benign cells tend to be grouped in monolayers, while cancerous cells are often grouped in multilayer. While in the Uniformity of cell size/shape the cancer cells tend to vary in size and shape. That is why these parameters are valuable in determining whether the cells are cancerous or not.

A. Wisconsin diagnosis breast cancer (WDBC)
B. Wisconsin prognosis breast cancer (WPBC)
C. Wisconsin breast cancer (WBC)

These Datasets have some features of their own and are depicted in Table 1.

According to the attributes in WDBC and WPBC datasets, attributes have 3 values with 3 columns in the data set.

The following equations demonstrate these attributes:

The mean calculated as:

$$Mean = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

The standard error calculated as:

$$S_e = \nabla \frac{S}{n} \quad (2)$$

Where, ∇ refers to Standard error parameter, S refers to Standard deviation and n refers to sample size.

Worst mean or largest mean:

Feature selection is an important step in building a classification model. It is advantageous to limit the number of input attributes in a classifier in order to have good predictive and less computationally intensive models. Chi-square test and Principal Component Analysis (PCA) are the two feature selection techniques proposed in this paper.

Chi-square is a statistical test commonly used for testing independence and goodness of fit. Testing independence determines whether two or more observations across two populations are dependent on each other (that is, whether one variable helps to estimate the other). Testing for goodness of fit determines if an observed frequency distribution matches a theoretical frequency distribution. Principal Component Analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables.

4. Insights on machine learning classifier

Machine learning is employed in a range of computing tasks where designing and programming explicit, rule based algorithms is infeasible depicted in Fig. 2. Machine learning algorithms identifies to which of a set of categories a new observations belongs on the basis of a training set of data containing observation whose category membership is known. It explores the construction and study of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions. Using dataset in the classifier, the malignancy and benignancy of tumorous cell is identified [22].

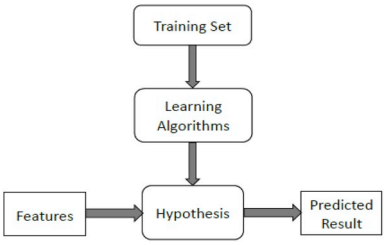

Fig. 2. Machine learning classifier process.

Several widely used classifiers are listed below:

- Naïve Bayes
- Logistic Regression
- Support Vector Machines (SVM)
- Decision Tree
- Reinforcement learning
- Neural Network
- *k*-Nearest Neighbors algorithm (KNN,IBK)

Here in this research work Logistic Regression classifier has been used.

5. Preferred use of logistic regression

Logistic regression classifier is used in this research work because it implies:

- *Probabilistic interpretation:*

The probabilistic interpretation is easy. But it is not available in Decision Tree and SVM classifier [23].

- *Feature correlation*:

Under this mathematical model, we do not need to worry about the features being correlated unlike Naïve Bayes [24].

- *Upgrade:*

The model can be easily updated to take in new data. This can be done using online gradient descent.

- *Adjustment:*

Logistic regression model helps to easily adjust Classification Thresholds when we are unsure about the confidence intervals.

- *Computational Considerations:*

Logistic regression models are usually computationally less complicated to build and require less computation time to train compared with ANNs.

6. Related works

Up to now, there have been many proposed techniques for classification of breast cancer patterns with high classification accuracies. A decision tree method (C 4.5) was used for breast cancer detection with 94.74% classification accuracy [25]. A rule induction algorithm based on the approximate classification method was applied to a breast cancer detection problem. The obtained accuracy was 94.99% [26]. Linear discriminant analysis (LDA) and neural networks (NN) methods were proposed to classify the breast cancer. The accuracy of the proposed LDA + NN was 96.8% [27]. A support vector machine classifier was used and the obtained classification accuracy was 97.2% [28].

The prevalence of germline *TP53* mutations has been discussed in Ref. [29], the risk of breast cancer and other cancers in mutation carriers and management implications for women with breast cancer and unaffected women. A classification scheme which was based on a feed forward neural network rule extraction algorithm was proposed. The reported accuracy was 98.10% [30]. A neuro-fuzzy technique was proposed by Nauck and Kruse [31]. The accuracy was 95.06%. An AR + NN method was proposed to be used in a breast cancer diagnosis problem. The obtained classification accuracy was 97.4% [32]. The LVQ, big LVQ, and AIRS methods were applied to breast cancer detection. 96.7%, 96.8%, and 97.2% correction classification rates were reported, respectively [33]. An increase in RR, sensitivity and CDR when adding CAD to SR has been observed in Ref. [34]. Kinetic features such as peak enhancement and early and delayed enhancement profiles were acquired using CAD while measuring the apparent diffusion coefficient (ADC), diffusion coefficient, pseudodiffusion coefficient, and perfusion fraction (f) using IVIM modeling [35]. With five-fold cross validation tests show that the proposed CAD system detects the mass location with an overall accuracy of 99.7%. The system also distinguishes between benign and malignant lesions with an overall accuracy of 97% [36].

A supervised fuzzy clustering technique was proposed for breast cancer detection. The accuracy of 95.57% was obtained [37]. The mixture experts (ME) network structure was proposed for breast cancer diagnosis. The authors reported 98.85% correct classification rate for this technique [38]. An improved Bayesian belief networks (BBNs) using the linear regression technique was applied to breast tumor gene classification [39]. The authors reported that their proposal could effectively recognize important genes. An isotonic separation approach for breast cancer detection was proposed in Ref. [40]. The proposed method yielded good achievements in breast cancer detection. Use of the concurrent-read CAD system for interpretation of screening ABUS studies of women with dense breast tissue who do not have symptoms is expected to make interpretation significantly faster and produce non-inferior diagnostic accuracy compared with interpretation without the CAD system [41]. A Bayesian classifiers performance was evaluated for the diagnosis of breast cancer using two different real datasets [42]. In Ref. [43], a new hybrid feature selection method named as AP has been formed to detect breast cancer, using association rules (Apriori algorithm) and Principal Component Analysis (PCA) together with artificial neural network classifier and as the results suggest, this system, which is performed through size reduction, is a feasible system for faster and more accurate diagnosis of diseases. A combination of statistical methods and particle swarm optimization (PSO) for mining breast cancer patterns was proposed in Ref. [44]. Detection and Diagnosis systems (CAD) are intended to assist the radiologist in making decisions. Hence, comparison study between the most recent CAD systems in order to clarify usefulness as well as to propose an emerging methodology for implementation that could certainly help clinicians during their diagnosis [45].

7. Working principles of logistic regression and its features

The binary logistic regression model is used to predict binary response based on one or more predictor variables (e.g. features). Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution [46]. The term regression comes from the fact that we are fitting a linear model to the feature space. Logistic Regression involves a more probabilistic view of classification [47]. The outcome of logistic regression should be discreet and not continuous however, the logistic regression can work on multidimensional feature space (features can be categorical or continuous) as:

7.1. Hypothesis and cost function

The model for logistic regression involves a vector β in d-dimensional feature space. For a point X in feature space, we project it onto α to convert it into a real number Z in the range -∞ to + ∞ as:

$$Z = \alpha + \beta.\ \chi = \alpha + \beta 1.\ X1 + \ldots \beta dXd \tag{3}$$

The value of Z thus obtained is then mapped to the range from 0 to 1 using the logistic function

$$p = \frac{1}{1 + e^{-z}} \tag{4}$$

Thus this function transforms a point X in the d dimensional feature space in to a range of 0–1.

Thus by applying a probability threshold corresponding to a particular class, we can identify a particular class for a corresponding feature subsets.

In order to fit in the optimum parameters β to the real number Z, we plug in the logistic function into the following cost function.

$$-\frac{1}{m}\left[\sum_{i=1}^{m}\left[y^{(i)}\log(hyp) + (1 - y^{(i)})\log(1 - hyp)\right]\right] \tag{5}$$

Here 'I' corresponds to the ith training example and hyp is our proposed hypothesis which is the sigmoid function.

The value of β was initialized to be zero, then the corresponding cost was calculated, then the value of β is adjusted, so that the cost function is minimized.

The particular cost function chosen for logistic regression is a bowl-shaped cost function having a global minima. The probability threshold for malignant and benign tumors was set to be 0.5. Thus any probability greater than 0.5 will result in the classified output to be malignant and any probability less than 0.5 will result in the classified output to be benign.

The cost function chosen will penalize the algorithm with high cost, if the predicted output was malignant (1) instead of (0), alternatively if

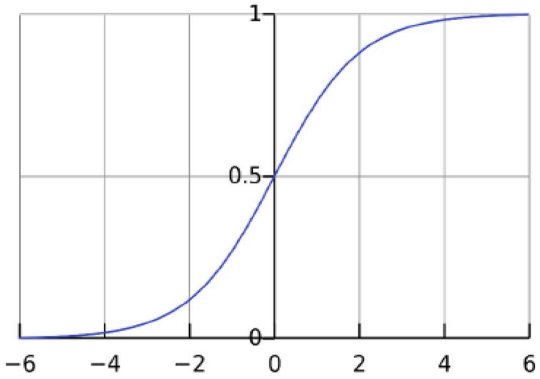

Fig. 3. Sigmoid function.

the predicted output was benign (0) instead of malignant (1) the cost function will also penalize with a high cost.

7.2. Sigmoid function

Sigmoid function is the most commonly known function used in feed forward neural networks because of its nonlinearity and the computational simplicity of its derivative [31]. Sigmoid function is the logistic function of the form

$$p = \frac{1}{1 + e^{-z}} \tag{6}$$

This sigmoid function maps the value of Z to a range from 0 to 1. The sigmoid function has the form shown in Fig. 3.

The threshold chosen was selected to be 0.5. The hypothesis predicts an output 1 for a probability greater than 0.5 and an outcome of 0 for a probability less than 0.5.

8. Proposed weighted function

We proposed a weighted hypothesis, where we add a weight factor θ to the existing logistic function. We use brute force search algorithm to find an optimum value for the weight factor what will yield a higher accuracy, sensitivity, specificity. The added weight factor will make an aggressive or submissive approach towards classification. The following expression shows the modified hypothesis with a dynamic weight added to it.

$$p = \frac{1}{1 + e^{-\theta z}} \tag{7}$$

Using a weighted sigmoid function, we have observed a significant improvement in the classifier performance compared with the one which used the classical techniques of logistic regression.

In this work, an enhanced classifier based on sparse representation-based classifier (SRC) is represented that provides us near-to-ideal operating points in the ROC curve with an AUC of 0.9754.

9. Insights on machine learning classifier

9.1. Optimization techniques

Optimization techniques are used in order to find suitable values of β that will minimize the cost function, so that with this optimized value of β classification can be performed on the test dataset. In our work, we used two optimization techniques: stochastic gradient descent & advanced optimization technique.

9.1.1. Gradient descent algorithm

Gradient descent is a first order optimization algorithm to find the global minima. In order to find the minima, one takes steps proportional to the negative of the gradient (or of the approximate gradient) of the function at the current point.

The gradient descent algorithm has the following format

$$\textbf{Repeat}\left\{\beta_{new}: =\beta_{old} - \mu \sum_{i=1}^{m} (hyp - y^{(i)}) * x^{(i)}\right\} \tag{8}$$

In this optimization technique, we simultaneously update the value of β and iterate the whole process, until a fairly global minima is reached. Thus, we have observed that with an iteration of 100 we fairly reached the global minima. Here μ is called the learning rate which is a measure of the steps that the algorithm will take in order to converge. A low value of μ will mean that the algorithm will require more steps to reach the global minima where as a high value of μ will mean that the cost function will converge at a faster rate [48]. However, a very high value of the learning rate although leads to a faster convergence does not necessarily imply that the function is guaranteed to reach the global minima. Hence in practice an optimum value of the learning rate is chosen which makes a tradeoff between the convergence rate and the global minima reached. In our work, we have chosen the learning rate to be 0.8 and have fairly reached towards the global minima.

One of the limitations of such an algorithm is that it only works well in finding the global minima of a convex function. If the function is not convex in practice, this algorithm will only help to reach the local minima of the cost function instead of proceeding towards the global minima.

9.1.2. Advance optimization technique

In the advanced optimization technique, we do not have to manually choose the value of the learning rate μ, which is chosen automatically. In the advanced optimization technique, we have used the BFGS (Broydon Fletcher Goldfarb Shanno) algorithm. BFGS is an iterative method for solving unconstrained nonlinear optimization problems. The BFGS method approximates Newton's method, a class of hill-climbing optimization techniques that seeks a stationary point of a (preferably twice continuously differentiable) function. For such problems, a necessary condition for optimality is that the gradient be zero. Newton's method and the BFGS methods are not guaranteed to converge unless the function has a quadratic Taylor expansion near an optimum. These methods use both the first and second derivatives of the function [49].

Such an algorithm has been proven to perform fairly well even for non-convex function where there may be multiple locally optimal points and it can take a lot of time to identify whether the problem has no solution or if the solution is global.

9.2. Brute force technique for finding optimum weights

In this section we have used our modified sigmoid function and have placed a dynamic weight of θ as discussed before. We use a brute force or exhaustive search algorithm and varied the value of θ from 0.01 to 1 with a step size of 0.01 and from 1 to 100 with a step size of 0.5. Our search algorithm takes into account the performance parameters of the classifier outputs for each value of θ and finds the optimum value of the weight θ that gives the greatest accuracy, and an improved sensitivity and specificity. The search algorithm compares the performance parameters each time with the previously obtained performance parameters of a particular value of θ. Since the performance parameters somehow changes arbitrarily with the value of θ we therefore compare the parameters for the adjacent values of θ and put the higher performance parameters on a secondary array. The search algorithm then chooses the value of θ from the secondary array which will lead towards obtaining the maximum accuracy, sensitivity and specificity.

9.3. Relationship of the weighted hypothesis

We dynamically varied our weight factor θ and obtained the value for θ that gives the maximum classifier performance. The proposed system of adding weights to the sigmoid function was tested with different datasets where we have observed that the addition of the weight factor always leads to an improvement in the classification performance. Moreover our proposed system gave further insights about the magnitude of the weight factor being added to the sigmoid function.

The proposed system was tested with Wisconsin breast cancer dataset consisting of 32 features. Keeping the optimization technique constant to advance optimization technique, when classification was performed with logistic regression on the larger dataset comprising of 32 features compared to 12 in Wisconsin diagnostic breast cancer dataset, an important observation came into role. The magnitude of the weight factor θ increases as the number of features in the dataset increases.

The proposed system was again tested by keeping the size of the data fixed and varying the optimization techniques used to reach the global minima of the cost function. The value of θ seems to change depending on whether stochastic gradient descent algorithm or advanced optimization technique was used.

Summarizing the above observation we came to a conclusion that

$$\theta = f(\textit{optimisation technique},\ \textit{feature's number in the dataset})$$

where, θ is a function of optimization technique and the number of features in the dataset.

Such an important observation leads us to improve our search algorithm where we can initialize our value of θ to a higher in order to reach the optimum value of θ and save a lot of computation time.

10. Result analysis

10.1. Results from MATLAB simulation for small features

The proposed system comprising of a weighted sigmoid function is tested on WDBC (Wisconsin diagnostic breast cancer) dataset. A two-fold cross validation is performed on the dataset and each time half of the dataset is held as the test set. The next section will give a performance comparison when logistic regression was applied with and without the weighted sigmoid function.

10.1.1. Results obtained from the classical system

The classical logistic regression without incorporating the weighted sigmoid function, when applied to the dataset consisting of 12 features lead to the following performance parameters. The parameters remained same of each of the 2 cross validated subsets aside for test purposes. Fig. 4 depicts the following results:

- Accuracy: 95.7082
- Error Rate: 0.0429
- Sensitivity: 0.9944
- Specificity: 0.8333
- Confusion matrix: 178 1 9 45

10.1.2. Results obtained from the proposed system

The proposed system has been implemented on the same dataset and using the brute force search algorithm, an optimum weight factor of 0.5 in the sigmoid function lead to the following maximum achievable performance parameters. Here Fig. 5 shows the following experimental quantities:

- Accuracy: 97.4249
- Error Rate: 0.0258
- Sensitivity: 0.9944
- Specificity: 0.9074

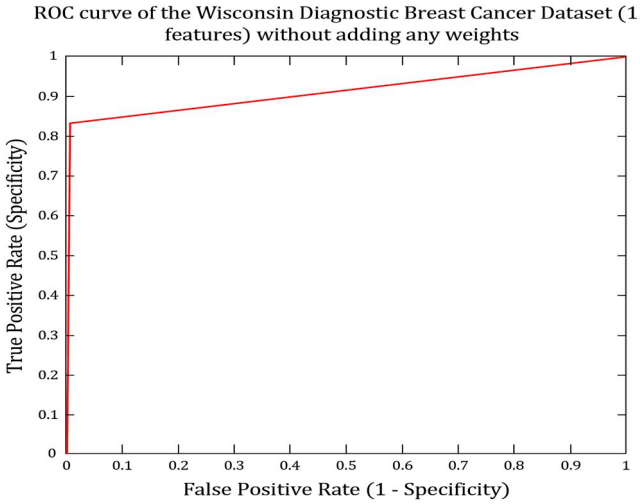

Fig. 4. Result of classical system for smaller dataset.

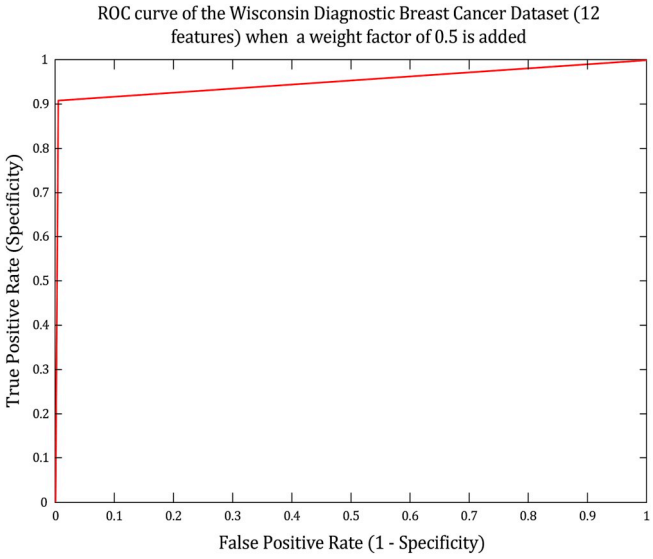

Fig. 5. Result of proposed system for smaller dataset.

Table 2
Performance parameters of existing and proposed method for smaller dataset.

Performance Parameters	Existing Method	Proposed Method
Accuracy	95.7082	97.4249
Sensitivity	0.9944	0.9944
Specificity	0.8333	0.9074
Error rate	0.0429	0.0252
Confusion matrix	178 1	178 1
	9 45	5 49

- Confusion matrix: 178 1 5 49

Comparing the two figures and the performance parameters above (Table 2), it is fairly easy to visualize the significant change and improvements in the accuracy, sensitivity and specificity when the weighted sigmoid function has been applied. The area under the receiver operating characteristic curve has been increased thus reflecting an improvement in the performance parameters. Confusion matrix are calculated for accuracy, sensitivity, specificity to visualize predicted outcomes and actual results. The calculated matrix are shown in Fig. 6.

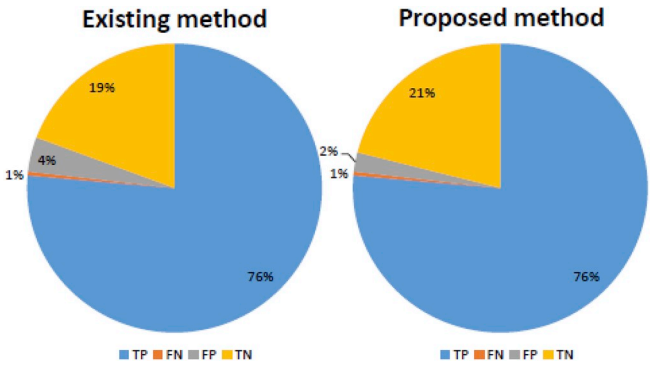

Fig. 6. Entries of confusion matrix from existing and proposed method for smaller dataset.

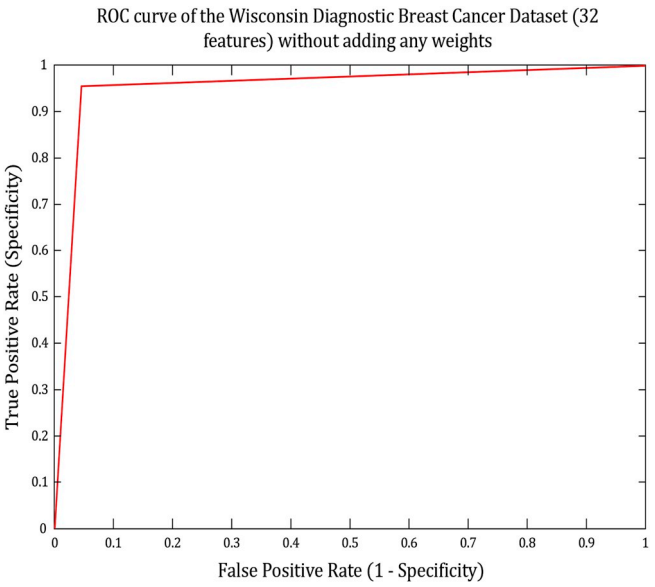

Fig. 7. Result of existing system for larger dataset.

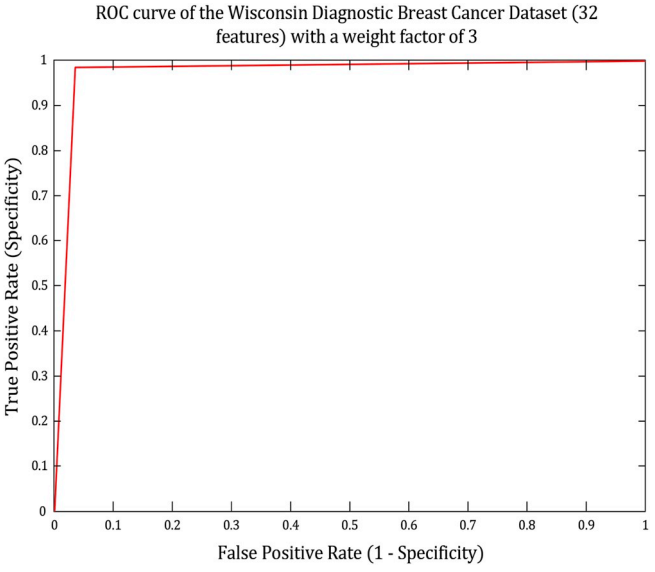

Fig. 8. Result of proposed system for larger dataset.

Table 3
Performance parameters of existing and proposed method for larger dataset.

Performance Parameters	Existing Method	Proposed Method
Accuracy	95.4225	96.8310
Sensitivity	0.9539	0.9631
Specificity	0.9552	0.9851
Error rate	0.0458	0.0317
Confusion matrix	207 10	209 8
	3 64	1 66

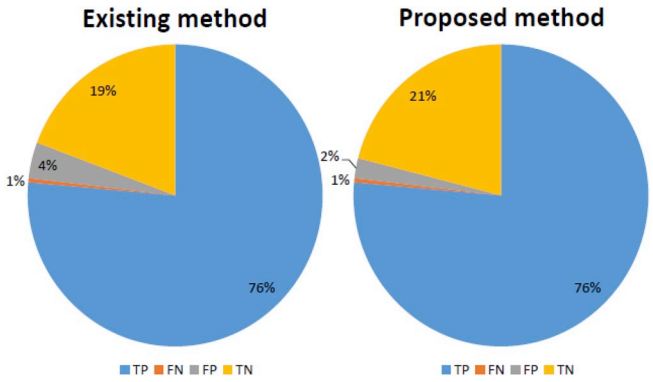

Fig. 9. Entries of confusion matrix from existing and proposed method for larger dataset.

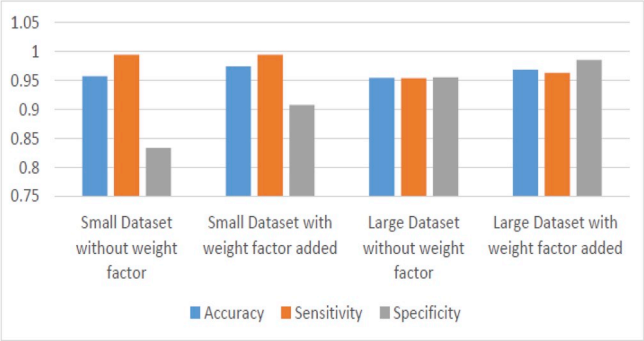

Fig. 10. Overall performance comparison.

10.2. Results from MATLAB simulation for large features

In order to ensure the effectiveness of the proposed system, the modified sigmoid function is applied to another dataset. The WBC (Wisconsin breast cancer) dataset is consisted of 32 features. The optimization technique was kept constant and hence an advance optimization technique with BFGS algorithm is deployed. The classical method without adding the weight factor in the hypothesis and with adding the weight factor is analyzed and the following performance parameters were obtained.

10.2.1. Results obtained from the classical system

The following performance parameters are obtained when no dynamic weights are added in the sigmoid function. The two fold cross validation technique has been applied, and the same performance parameters are obtained for each subsets, set aside for testing purposes. The performance parameters are shown below as Fig. 7:

- Accuracy: 95.4225
- Error Rate: 0.0458
- Sensitivity: 0.9539
- Specificity: 0.9552
- Confusion matrix: 207 10 3 64

10.2.2. Results obtained from the proposed system

The proposed system has been applied to the larger dataset and an addition of a dynamic weights to the sigmoid function using a brute force search algorithm led to an improvement in the ROC curve. The increase in area under the ROC curve also reflects the significant improvement that the proposed hypothesis brings.

The search algorithm found the value of the dynamic weight to be 3 for the larger dataset. The both the subsets, held aside for testing purposes in a 2 fold cross validation gave the following results which have been presented in Fig. 8:

- Accuracy: 96.8310
- Error Rate: 0.0317
- Sensitivity: 0.9631
- Specificity: 0.9851
- Confusion matrix: 209 8 1 66

Comparing the two figures and the performance parameters of the proposed system with that of the classical system, the improvements in the performance of the proposed system is evident. The following Table 3 gives further insight of an overall comparison for both the datasets.

Here is the comparison of different entries from confusion matrix shown in Fig. 9:

11. Conclusion and future work

The proposed modified sigmoid function successfully improved the performance upon the addition of a dynamic weight for both the datasets. Keeping the optimization technique constant the trend of the magnitude of the weighted factor is observed. The magnitude of the weight being added grows with the number of features in the dataset. Similar simulations are performed keeping the size of the data same but changing the optimization technique. Under such circumstances the value of the dynamic weights are also found to be different, thereby concluding the fact that this factor is a function of both the data size and the optimization technique. The following Fig. 10 summarizes the comparison between the two methods for both the datasets.

In addition to it, certain improvements can be brought by plotting in different features with each other and observe to the extent to which they are correlated. Another important insight towards improving the results obtained by logistic regression, specially our proposed method might include addition of online learning algorithm so as to discover the attributes associated with malignant and benign tumors. We can incorporate methods of automatic feature selection, in case of our data set which has large number of features we can automatically select features that will most likely to contribute towards the classification of benign or malignant tumor. In order to reduce the features or enhance it if there are less number of features, a method of regularization could also be undertaken so as to improve the accuracy of our proposed system.

Here a modified sigmoid function has been proposed which has a dynamic weight added to it. Future work regarding the searching criterion of the brute force search algorithm can also be done that would reduce the algorithm complexity. Furthermore, in this work, it was only managed to find the dependency of the weight factor with the size of the data set and the optimization technique. However, the exact relationship is yet to be worked upon. If the exact relationship with the data size can be found then the complexity of the search algorithm can be reduced significantly by initializing the weight with the desired value so that the algorithm does not have to search unnecessarily through the weights that does not seem to improve the performance. In order to reduce the features or enhance it if there are less number of features, a method of regularization could also be undertaken so as to improve the accuracy of our proposed system.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imu.2019.100189.

References

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015: cancer statistics, 2015. CA. Cancer J. Clin. Jan. 2015;65(1):5–29.
[2] Heymach J, et al. Clinical cancer advances 2018: annual report on progress against cancer from the American society of clinical oncology. J Clin Oncol Jan. 2018;36(10):1020–44.
[3] "Breastcancer.org - Breast Cancer Information and Support," Breastcancer.org. [Online]. Available: https://www.breastcancer.org/. [Accessed: 11-Feb-2019].
[4] Chen W, et al. Cancer statistics in China, 2015. CA. Cancer J. Clin. Mar. 2016;66(2):115–32.
[5] PDQ Adult Treatment Editorial Board. Breast cancer treatment (PDQ®): health professional version. PDQ cancer information summaries. Bethesda (MD): National Cancer Institute (US); 2002.
[6] Mustafa M, Nornazirah A, Salih FM, Illzam E, Suleiman M, Sharifa A. Breast cancer: detection markers, prognosis, and prevention. 2016.
[7] B.P, L.B. World cancer report. 2008.
[8] Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA. Cancer J. Clin. Apr. 2011;61(2):69–90.
[9] N. P. Pérez, "Improving variable selection and mammography-based machine learning classifiers for breast cancer CADx", p. 185.
[10] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA A Cancer J Clin 2018;68(1):7–30.
[11] G. I. Salama, M. B. Abdelhalim, and M. A. Zeid, "Breast cancer diagnosis on three different datasets using multi-classifiers," vol. vol. 01, no. 01, p. 8.
[12] Story HL, Love RR, Salim R, Roberto AJ, Krieger JL, Ginsburg OM. Improving outcomes from breast cancer in a low-income country: lessons from Bangladesh. *International Journal of Breast Cancer*; 2012 [Online]. Available:. https://www.hindawi.com/journals/ijbc/2012/423562/, Accessed date: 11 February 2019.
[13] Gvamichava R, Lomtadze N, Alibegashvili T, Charkviani T, Beruchashvili T, Jugeli L. Cancer screening program in Georgia (results of 2011). Georgian Med News Jul. 2012(208–209):7–15.
[14] "UCI Machine Learning Repository." [Online]. Available: http://archive.ics.uci.edu/ml/index.php. [Accessed: 11-Feb-2019].
[15] "Computer-Aided Detection and Diagnosis in Medical Imaging," CRC Press. [Online]. Available: https://www.crcpress.com/Computer-Aided-Detection-and-Diagnosis-in-Medical-Imaging/Li-Nishikawa/p/book/9781439871768. [Accessed: 11-Feb-2019].
[16] Moon WK, Shen Y-W, Huang C-S, Chiang L-R, Chang R-F. Computer-aided diagnosis for the classification of breast masses in automated whole breast ultrasound images. Ultrasound Med Biol Apr. 2011;37(4):539–48.
[17] Tang J, Rangayyan RM, Xu J, El Naqa I, Yang Y. Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. IEEE Trans. Inf. Technol. Biomed. Publ. IEEE Eng. Med. Biol. Soc. Mar. 2009;13(2):236–51.
[18] J. Shan, "A fully automatic segmentation method for breast ultrasound images," p. 111.
[19] Salama GI, Abdelhalim MB, Zeid MA. Experimental comparison of classifiers for breast cancer diagnosis. 2012 seventh international conference on computer engineering & systems (ICCES), cairo, Egypt. 2012. p. 180–5.
[20] Street WN, Wolberg WH, Mangasarian OL. Nuclear feature extraction for breast tumor diagnosis. Presented at the IS&T/SPIE's symposium on electronic imaging: science and technology, san jose, CA. 1993. p. 861–70.
[21] Wolberg WH, Street WN, Heisey DM, Mangasarian OL. Computerized breast cancer diagnosis and prognosis from fine-needle aspirates. Arch. Surg. Chic. Ill May 1995;130(5):511–6. 1960.
[22] Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer Inf Feb. 2007;2:59–77.
[23] Xue JH, Titterington DM. "Comment on: 'On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. Neural Process Lett 2008;28:169–87.
[24] Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology Sep. 1983;148(3):839–43.
[25] Quinlan JR. Improved use of continuous attributes in C4.5. Feb. 1996. arXiv:cs/9603103.
[26] H. J. Hamilton, N. Shan, and N. Cercone, "RIAC: a rule induction algorithm based on approximate classi cation," p. 22.
[27] Polat K, Güneş S. Breast cancer diagnosis using least square support vector machine. Digit Signal Process Jul. 2007;17(4):694–701.
[28] Bennett KP, Blue JA. "A support vector machine approach to decision trees. Anchorage AK, editor. 1998 IEEE international joint conference on neural networks proceedings. IEEE world congress on computational intelligence (cat. No.98CH36227), vol. 3. 1998. p. 2396–401. USA.
[29] Schon K, Tischkowitz M. Clinical implications of germline mutations in breast cancer: TP53. Breast Canc Res Treat Jan. 2018;167(2):417–23.
[30] R. Setiono, Generating concise and accurate classification rules for.
[31] Nauck D, Kruse R. Obtaining interpretable fuzzy classification rules from medical data. Artif Intell Med Jun. 1999;16(2):149–69.
[32] Karabatak M, Ince MC. An expert system for detection of breast cancer based on

association rules and neural network. Expert Syst Appl Mar. 2009;36(2):3465–9.
[33] Goodman DE, Boggess LC, Watkins andrew B. Artificial immune system classification of multiple-class problems. In proc. Of intelligent engineering systems. 2002. p. 179–84.
[34] Henriksen EL, Carlsen JF, Vejborg IM, Nielsen MB, Lauridsen CA. The efficacy of using computer-aided detection (CAD) for detection of breast cancer in mammography screening: a systematic review. Acta Radiol Jan. 2019;60(1):13–8.
[35] Song SE, et al. Intravoxel incoherent motion diffusion-weighted MRI of invasive breast cancer: correlation with prognostic factors and kinetic features acquired with computer-aided diagnosis. J Magn Reson Imaging Jan. 2019;49(1):118–30.
[36] Al-masni MA, et al. Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. Comput Methods Progr Biomed Apr. 2018;157:85–94.
[37] Abonyi J, Szeifert F. Supervised fuzzy clustering for the identification of fuzzy classifiers. 2003.
[38] Zhao X, Huang D, Dellandrea E, Chen L. Automatic 3D facial expression recognition based on a Bayesian belief net and a statistical facial feature model. 2010 20th international conference on pattern recognition, istanbul, Turkey. 2010. p. 3724–7.
[39] Ryu YU, Chandrasekaran R, Jacob VS. Breast cancer prediction using the isotonic separation technique. Eur J Oper Res Sep. 2007;181(2):842–54.
[40] Cruz-Ramírez N, Acosta-Mesa HG, Carrillo-Calvet H, Nava-Fernández LA, Barrientos-Martínez RE. Diagnosis of breast cancer using Bayesian networks: a case study. Comput Biol Med 2007;37(11):1553–64.
[41] Jiang Y, Inciardi MF, Edwards AV, Papaioannou J. Interpretation time using a concurrent-read computer-aided detection system for automated breast ultrasound in breast cancer screening of women with dense breast tissue. Am J Roentgenol May 2018;211(2):452–61.
[42] Maskery SM, Hu H, Hooke J, Shriver CD, Liebman MN. A Bayesian derived network of breast pathology co-occurrence. J Biomed Inform Apr. 2008;41(2):242–50.
[43] O. Inan, M. S. Uzer, and N. Yılmaz, "A New hybrid feature selection method based on association rules and pca for detection of breast cancer," p. 13.
[44] Yeh W-C, Chang W-W, Chung YY. A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. Expert Syst. Appl. May 2009;36(4):8204–11.
[45] Afef L, Rania T, Hanen C, Lamia S, Ahmed BH. Comparison study for computer assisted detection and diagnosis 'CAD' systems dedicated to prostate cancer detection using MRImp modalities. 2018 4th international conference on advanced technologies for signal and image processing (ATSIP). 2018. p. 1–6.
[46] Hilbe JM. Logistic regression models. CRC Press; 2009.
[47] Han J, Moraga C. The influence of the sigmoid function parameters on the speed of Backpropagation learning. Proceedings of the international workshop on artificial neural networks: from natural to artificial neural computation, london, UK, UK. 1995. p. 195–201.
[48] "Stochastic gradient descent," Wikipedia. 26-Jan-2019.
[49] Dai Y-H. Convergence properties of the BFGS algoritm. SIAM J Optim Jan. 2002;13(3):693–701.

Update

Informatics in Medicine Unlocked

Volume 20, Issue , 2020, Page

DOI: https://doi.org/10.1016/j.imu.2020.100435

Informatics in Medicine Unlocked 20 (2020) 100435

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: http://www.elsevier.com/locate/imu

Erratum regarding missing Declaration of Competing Interest statements in previously published articles

Owing to a Publisher error Declaration/Conflict of Interest statements were not included in the published versions of the following articles, that appeared in previous issues of Informatics in Medicine Unlocked.

The appropriate Declaration/Conflict of Interest statements, provided by the Authors, are included below.

1. The relationship between retinal vessel geometrical changes to incidence and progression of Diabetic Macular Edema (Informatics in Medicine Unlocked; 2019; Vol 16C; Article number: 100,248) https://doi.org/10.1016/j.imu.2019.100248

   Declaration of Competing Intere: The Authors have no interests to declare.

2. A machine learning algorithm to improve patient-centric pediatric cardiopulmonary resuscitation (Informatics in Medicine Unlocked; 2020; Vol 19C; Article number: 100,339) https://doi.org/10.1016/j.imu.2020.100339

   Declaration of interest: The Authors have no interests to declare.

3. Decision support system for diagnosing Rheumatic-Musculoskeletal Disease using fuzzy cognitive map technique (Informatics in Medicine Unlocked; 2019; Vol 18C; Article number: 100,279) https://doi.org/10.1016/j.imu.2019.100279

   Declaration of interest: The Authors have no interests to declare.

4. Channel binary pattern based global-local spatial information fusion for motor imagery tasks (Informatics in Medicine Unlocked; 2020; Vol 20C; Article number: 100,352) https://doi.org/10.1016/j.imu.2020.100352

   Declaration of interest: The Authors have no interests to declare.

5. Classification of malignant and benign tissue with logistic regression (Informatics in Medicine Unlocked; 2019; Vol 16C; Article number: 100,189) https://doi.org/10.1016/j.imu.2019.100189

   Declaration of interest: The Authors have no interests to declare.

6. On parameter interpretability of phenomenological-based semi-physical models in biology (Informatics in Medicine Unlocked; 2019; Vol 15C; Article number: 100,158) https://doi.org/10.1016/j.imu.2019.02.002

   Declaration of interest: The Authors have no interests to declare.

7. Multistage Classifier-Based Approach for Alzheimer's Disease Prediction and Retrieval (Informatics in Medicine Unlocked; 2018 vol 14C; Pages 34–42) https://doi.org/10.1016/j.imu.2018.12.003

   Declaration of interest: The Authors have no interests to declare.

8. A novel somatic cancer gene based biomedical document feature ranking and clustering model (Informatics in Medicine Unlocked; 2019; Vol 16C; Article number: 100,188) https://doi.org/10.1016/j.imu.2019.100188

   Declaration of interest: The Authors have no interests to declare.

9. Developing an ultra-efficient microsatellite discoverer to find structural differences between SARS-CoV-1 and Covid-19 (Informatics in Medicine Unlocked; 2020 vol 19C; Article number: 100,356) https://doi.org/10.1016/j.imu.2020.100356

   Declaration of interest: The Authors have no interests to declare.

10. Patient-specific optimization of mechanical ventilation for patients with acute respiratory distress syndrome using quasi-static pulmonary P–V data (Informatics in Medicine Unlocked; 2018;

DOIs of original article: https://doi.org/10.1016/j.imu.2020.100352, https://doi.org/10.1016/j.imu.2020.100356, https://doi.org/10.1016/j.imu.2019.100248, https://doi.org/10.1016/j.imu.2018.12.003, https://doi.org/10.1016/j.imu.2019.100177, https://doi.org/10.1016/j.imu.2019.100189, https://doi.org/10.1016/j.imu.2019.100256, https://doi.org/10.1016/j.imu.2020.100339, https://doi.org/10.1016/j.imu.2019.100279, https://doi.org/10.1016/j.imu.2019.100188, https://doi.org/10.1016/j.imu.2019.100171, https://doi.org/10.1016/j.imu.2018.06.003, https://doi.org/10.1016/j.imu.2018.12.001, https://doi.org/10.1016/j.imu.2020.100308, https://doi.org/10.1016/j.imu.2019.100184, https://doi.org/10.1016/j.imu.2019.02.002, https://doi.org/10.1016/j.imu.2020.100304.

https://doi.org/10.1016/j.imu.2020.100435

Available online 28 September 2020

Vol 12C; Pages 44–55) https://doi.org/10.1016/j.imu.2018.06.003

Declaration of interest: The Authors have no interests to declare.

11. A technology framework for remote patient care in dermatology for early diagnosis (Informatics in Medicine Unlocked; 2019; Vol 15C; Article number: 100,171) https://doi.org/10.1016/j.imu.2019.100171

Declaration of interest: The Authors have no interests to declare.

12. CBIR System Using Capsule Networks and 3D CNN for Alzheimer's disease Diagnosis (Informatics in Medicine Unlocked; 2018; Vol 14C; Pages 59–68) https://doi.org/10.1016/j.imu.2018.12.001

Declaration of interest: The Authors have no interests to declare.

13. The effect of computerized physician order entry on mortality rates in pediatric and neonatal care setting: Meta-analysis (Informatics in Medicine Unlocked; 2020; Vol 19C; Article number: 100,308) https://doi.org/10.1016/j.imu.2020.100308

Declaration of interest: The Authors have no interests to declare.

14. Identification of the core ontologies and signature genes of polycystic ovary syndrome (PCOS): A bioinformatics analysis. (Informatics in Medicine Unlocked; 2020; Vol 18C; Article number: 100,304) https://doi.org/10.1016/j.imu.2020.100304

Declaration of interest: The Authors have no interests to declare.

15. Balanites aegyptiaca (L.) Del. for dermatophytoses: Ascertaining the efficacy and mode of action through experimental and computational approaches (Informatics in Medicine Unlocked; 2019; Vol 15C; Article number: 100,177) https://doi.org/10.1016/j.imu.2019.100177

Declaration of interest: The Authors have no interests to declare.

16. Barriers and technologies of maternal and neonatal referral system in developing countries: A narrative review (Informatics in Medicine Unlocked; 2019; Vol 15C; Article number: 100,184) https://doi.org/10.1016/j.imu.2019.100184

Declaration of interest: The Authors have no interests to declare.

17. Automated Grading of Prostate Cancer using Convolutional Neural Network and Ordinal Class Classifier (Informatics in Medicine Unlocked; 2019; Vol 17C; Article number: 100,256) https://doi.org/10.1016/j.imu.2019.100256

Declaration of interest: The Authors have no interests to declare.