Computers and Electrical Engineering 84 (2020) 106624

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Combination of loss functions for robust breast cancer prediction☆

Hamideh Hajiabadi[a,*], Vahide Babaiyan[a], Davood Zabihzadeh[b], Moein Hajiabadi[a]

[a] Department of Computer Engineering, Birjand University of Technology, Birjand, Iran
[b] Computer Department, Engineering Faculty, Sabzevar University of New Technology, Sabzevar, Iran

ARTICLE INFO

Article history:
Received 11 June 2019
Revised 7 January 2020
Accepted 5 March 2020

Keywords:
Artificial neural network
Breast cancer detection
Margin enhancing loss function
Robust loss function

ABSTRACT

Cancer detection can be formulated as a binary classification in a machine learning paradigm. Loss functions are a critical part of almost every machine learning algorithm. While each loss function comes up with its own advantages and disadvantages, in this paper, inspired by ensemble methods, we propose a novel objective function that is a linear combination of single losses. We then integrate the proposed objective function into an Artificial Neural Network (ANN) to diagnose breast cancer. By doing so, both the coefficients of loss functions and weight parameters of the ANN are learned jointly via backpropagation. As the patients' data are sometimes very noisy, we evaluate our method by doing comprehensive experiments on Wisconsin Breast Cancer Diagnosis (WBCD) dataset at different noise levels. The experiments show its performance declines very slowly (from 0.97 to 0.96) compared to the peer methods with the increase of noise level.

1. Introduction

Nowadays, cancer is known as the most devastating disease in the world. It is reported that In the United States, the number of new cancer cases are around 1,700,000 each year while the number of cancer deaths accounted for around 580,000 [1]. Among women, breast cancer is the most commonly diagnosed, which accounts for 29% [1]. Nowadays, tumor Diagnosis is one of the trending and challenging issues in the medical informatics field of study.

Traditionally, radiologists predict breast cancer based on mammography. In 1994, a research was conducted to evaluate the accuracy of radiologists' predictions. In that research, 150 mammograms for tumour prediction were analyzed and diagnosed by 10 radiologists to detect breast cancers. In their study, in spite of the proved value of using mammograms for cancer detection, more than 3% of cancers were undiagnosed by 90% of radiologists. Today, there is an increasing attraction in employing more and more artificial intelligence technologies to collect and analyze medical data. considering the large volume of cancer cases, physicians find it difficult to learn every detailed cancer and therefore, having automatically data analysis tools as an assistant would help physicians to make an efficient decision.

☆ This paper is for CAEE special section SI-cih. Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. Hamed Vahdat-Nejad.

* Corresponding author.
E-mail addresses: hajiabadi@birjandut.ac.ir (H. Hajiabadi), babaiyan@birjandut.ac.ir (V. Babaiyan), d.zabihzadeh@mail.um.ac.ir (D. Zabihzadeh), m.hajiabadi@birjandut.ac.ir (M. Hajiabadi).

https://doi.org/10.1016/j.compeleceng.2020.106624
0045-7906/

Machine learning algorithms have been used in a wide variety of applications in recent years, and researchers have achieved valuable results in prediction and classification Even in some researches good comparisons have been conducted between different methods of Machine Learning (ML) [2]. Researchers have also achieved better results compared to the original algorithms by developing and improving existing methods [3,4]. To improve the quality of tumour detection, many researchers have been tending to use ML and data mining techniques to automatically and accurately detect cancers. It has been experimentally proved that these techniques can achieve good results in real-world problems. Breast cancer diagnosis can be transferred into a classification problem in the ML research area [5]. Being trained to recognize separate types of tumours, the classifier would later predict unseen tumour, based on the historical tumour data.

Since the early detection of cancer can increase the patient's chances of being cured, the accurate prediction is of importance for medical scientists. In the past recent years, researchers have continuously attempted to detect cancer in the early stage especially before appearing the symptoms [6]. During the several past years, large amounts of cancer data have been collected Using the new developed technologies, and researchers have easily used them to develop and evaluate their proposed tools.

Supervised learning is all about the ability to generalize knowledge. Specifically, the goal of supervised ML techniques is to train a classifier based on training data, in such a way that it would be able to correctly classify new unseen data, which is indicated by generalization capability. The generalization of a classifier is in correlation with its robustness to perturbations of the data. In other words, if a classifier can cope with data disturbance, it is expected to generalize well. In this paper, to improve robustness on one hand and generalization on the other hand, the use of a simple classifier extended with a new combined loss function is proposed.

The contributions of this paper are summarized as follows:-

1. inspired by ensemble techniques, we apply the ensemble approach to the objective function by incorporating an ensemble loss function to a multilayer neural network and it is then tested on Wisconsin Breast Cancer Diagnosis (WBCD) dataset.
2. incorporating the ensemble loss function as a tool, we advance the advantages of each individual loss function.
3. The advantage of our method compared to the current ensemble methods is that it simultaneously learns the parameter of the learner and the weights associated with every single loss.

This paper is structured as follows. We review the related works in Section 2. The supervised learning framework is discussed in Section 3. A review of several existing loss functions with their characteristics are provided in Section 4. The proposed method is explained in Section 5 for which the implementation and test results are provided in Section 6. Lastly, conclusion are included in Section 7 with a range of suggestions for future works.

2. Related works

Today, several machine learning algorithm-based breast cancer classification methods are available. In the following, a number of the most common algorithms will be presented using past studies carried out in this area.

In [7] a model utilized to diagnose and categorize breast cancer; it was offered to assist patients and doctors to accurately classify breast cancer. This model employs a Normalized Multi-Layer Perceptron Neural Network. The results obtained from this model display a high accuracy in comparison with the results acquired from past research with the use of Artificial Neural Networks (ANN). Data normalization was revealed to be very important in this study.

In [8], to early detect breast cancer four machine learning algorithms were utilized. This research aims to analyze the results of blood tests via various machine learning methods. This is performed to specify the effectiveness of these methods in to diagnosis of breast cancer detection. ANN, support-vector machines (SVM), extreme learning machines (ELM), and the k-nearest neighbors (k-NN) algorithm were employed to carry out the study. In this research, some attributes of a UCI library dataset were used, which are listed in [8]. The results reveal that the standard ELM yielded the highest level of accuracy and the minimum training time.

Authors in [9] investigate computer-aided detection and diagnosis systems that work based on deep learning techniques that are utilized to carry out breast cancer screening. Thanks to deep neural networks, it is possible to identify learning algorithms almost as efficiently as it would be if it was carried out by a human. In some case studies, it was discovered that some remarkable progress has been attained concerning common machine learning methods. Deep learning has been improved in several areas.

A three-step hybrid structure utilized for breast cancer diagnosis was presented in the proposed method [10]. The data is normalized in the first step. Next, in the second step, the k-means clustering features are employed to weigh the normalized data. Ultimately, AdaBoostM1 is utilized for the classification of the weighted set. For breast cancer diagnosis, a hybrid model has been offered. This model consists of normalization, feature weighting and classification to combine breast cancer datasets. It is possible to use alternative and new clustering algorithms instead of k-means clustering. These alternative options include clustering based on optimization, mean-shift clustering, and fuzzy c-means. The results display that AdaBoostM1 classifiers in combination with the MAD normalization method constitute 75% of the accuracy level in the diagnosis of breast cancer. Whereas the combined method yielded a success level of 91.37%. The results indicate that the suggested system can be employed to detect breast cancer accurately.

In [11], it was tried to identify important factors involved with the occurrence of metastatic in breast cancers. The research was conducted via comparing logistic regression (LR) and ANN models. For detecting metastatic in breast cancers, the ANN is an appropriate alternative option compared to traditional LR models.

In [12], it was tried to examine machine learning techniques and their application for the diagnosis as well as the prognosis of breast cancer. This study [12] reviewed many machine learning techniques consisting of ANNs, SVMs, decision trees (DTs), and k-NNs. Next, these techniques are investigated through comparisons performed using the Wisconsin Breast Cancer Diagnosis (WBCD) dataset.

A new classification method for breast cancer was provided in [13]. In that method, first, data is clustered via Expectation Maximization (EM). In the next step, noise is removed. Next, the data is classified into two classes namely benign and malignant. In the classification step, Classification and Regression Trees (CART) are utilized to extract fuzzy rules. The experimental results from the WBCD and Mammographic Mass datasets reveal the fact that the method improves accuracy up to a significant level. They recommend that their method can be exploited as a clinical assistant system to help doctors with their healthcare activities.

A hybrid smart classification model is presented in [14]. This model comprises three stages. These stages include sample selection, feature selection, and classification. In the stage of sample selection, the fuzzy-rough feature selection method based on a weak gamma evaluator has been employed. This leads to the removal of unneeded or useless items. In the phase of feature selection, consistency-based feature selection method combined with an algorithm for re-ranking is utilized. These two steps of pre-processing help improve the training time. In the stage of model classification, fuzzy-rough nearest neighbor algorithm is employed. The WBCD dataset was used to test the effectiveness of the presented model.

A kind of Neural Network algorithm was proposed in [15] that is named GONN. Neural network is employed genetically to optimize architecture (consisting of weight and structure) for the purpose of classification. Moreover, some new crossover and mutation operators were presented that mitigates the harmful nature of common operators. In this research, the GONN algorithm is used to classify breast cancer tumors into the two different types of benign or malignant tumors. The WBCD database was employed to evaluate this algorithm. The results indicate that their proposed method works properly with the breast cancer database. Also, it was revealed that it could be a good alternative to popular machine learning methods.

In [16], a comparison was made on some machine learning algorithms on the WBCD dataset. These algorithms included, Naive Bayes (NB), k-NN, SVM, and Decision Tree (C4.5). The goal was to identify the highest level of classification accuracy. These algorithms performance was tested in terms of precision, specificity, accuracy, and sensitivity. The results suggested that SVM had the highest accuracy (97.13%). It had the lowest rate of error. It was effective and had a good accuracy level and a low rate of error for breast cancer detection. All the tests were conducted in similar simulation environments.

In [17], a neural-network based method named FFBPN was employed to classify breast cancer. Two types of breast cancer including malignant and benign are exist. In this research, some samples taken from the WBCD have been trained and also validated. They examined the effect of a different number of layers (1, 2, and 3 layers) on ANN in terms of output accuracy. Moreover, their research offers the results of the performance of ANN based on the number of different neurons in the hidden layer. Based on the results, it was discovered that the best design for the network is one that has three hidden layers, with 21 neurons present in its hidden layer and also TANSIG as its activation functions.

An approach to predict benign breast cancer based on neural networks was proposed in [18]. In that study, a feed-forward neural network was made. Next, a differential evolution propagation algorithm was employed for network training. The algorithm was tested using the Wisconsin Diagnostic and Prognostic Breast Cancer dataset. The purpose of the study was creating an effective tool to build neural models to help with the classification of different classes of breast cancer. In this study, in order to offer topologies with random-random policies and compare their results, two different types of migration topologies were offered. Based on the experiments, the time it takes for torus topology is higher in comparison with random topology. However, they offer similar quality levels. With this model, it becomes a possibility to automatically classify different kinds of breast cancers without expert opinion. The results show decreased time complexity and higher solution quality.

In [19], recent machine learning methods for cancer progression modeling were reviewed. Different input features and data samples were utilized to interpret predictive models. By analyzing the results acquired using these models, it was discovered that the integration of multidimensional heterogeneous data along with the utilization of various techniques for feature selection and classification offers an appropriate tool for the determination of cancer progression level.

ML algorithms have been introduced with several features, including effective performance in health care datasets. At the same time, noise in some applications can be a serious challenge [20]. Data quality issues such as noise, outliers, missing or duplicated data are exciting. In order to improve the quality of the resulting analysis, the quality of the data should also be improved [21].

The quality of classification systems is based on the training dataset on which the classification models are built. When class labels are incorrectly assigned to samples, the training process changes and can therefore have a negative impact on classification performance. This problem is formally known as class label noise.

One of the most effective ways to reduce the harmful consequences of inaccurate objects is noise of three classification algorithms (C4.5, SVM, and NN), with and without filtering that removes noisy samples from training data. Two approaches have been introduced to control the class noise and its negative affect on the learning process [22]: Algorithm-level approaches that reconcile existing algorithms to properly manage the noise. For instance, they used C4.5 which uses pruning strategies to build the tree in such way that is less influenced by noisy data. Data-level techniques which rely on eliminating

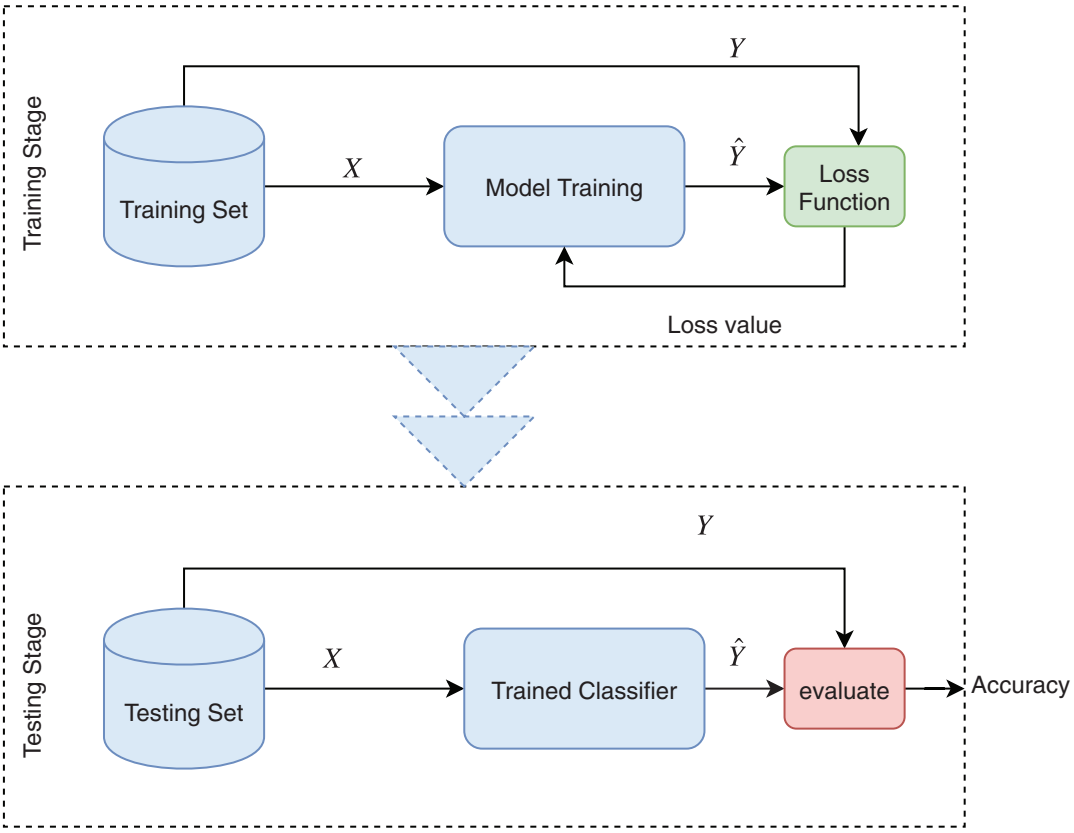

Fig. 1. The supervised learning framework.

noisy samples to decrease the effect of noise before building classifiers. In [22] the effectiveness of various noise filters on real-world medical datasets which are affected by class noise were investigated. They analyzed the performance of noise filtering by considering six different noise filters and 12 medical datasets with different noise levels. After reviewing the results, they did not come up with a general solution but they recommended to test all the available classifiers and noise filters on similar issues to obtain the best result.

3. The supervised learning framework

Fig 1 shows the supervised learning framework which generally consists of three major components:

Data. Let (X, Y) represents all samples where inputs denotes by $X = \{\vec{x_1}, \vec{x_2}, \ldots, \vec{x_N}\}$ and $Y = \{y_1, y_2, \ldots, y_N\}$ as the true labels. $\vec{x_i} \in R^d$ is a vector with d elements, each for one feature and $y_i \in \{+1, -1\}$ for binary classification.

Hypothesis class. In the learning process, candidate hypotheses are considered as the class F which consists of functions from X to Y. It reflects some kind of prior knowledge about the problem at hand. Here, is an example about the class of half-spaces,

$$H_{halfspace} = \{f_w(\vec{x}) = sign(w^T\vec{x}) | f_w \in F : R^d \rightarrow \{+1, 1\}, w \in R^d\}$$

Loss function. a loss function is to quantify how much the estimated label calculated by the classifier $f \in F$ is close to the true label. The loss function is defined as a function $l(\vec{x}, y; f)$. The most intuitive loss function in the binary case is the $0 - 1$ loss function which is defined by

$$l(\vec{x_i}, y_i; f) = 1_{[f(\vec{x_i}) \neq y_i]}$$

A learning algorithm is aimed to find a classifier $f \in F$ which is optimal [23]. It means that the expected loss function over all samples (seen and unseen) which is defined as follows is minimized

$$R_{l,P}(f) = \int_{X \times Y} l(y, f(\vec{x}))dP(X, Y)$$

where $P(X, Y)$ is the joint probability distribution over X and Y. Generally $R_{l,P}(f)$ is not generally computable as the distribution $P(X, Y)$ is unknown. Instead, an approximation of $R(f)$ which is known as empirical risk is calculated. It is the average

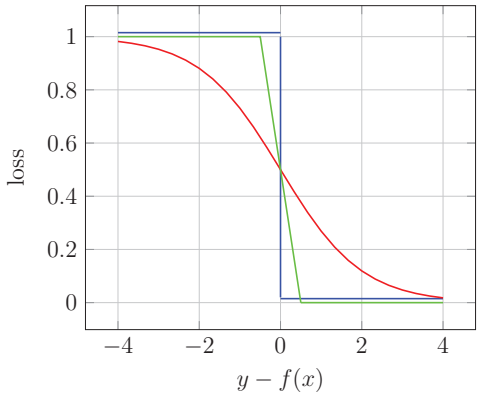

Fig. 2. Ramp loss function (green), the $0-1$ loss (blue), Sigmoid loss (red)

value of loss function over all training set containing n samples [24],

$$R_{\text{emp}}(f) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, f(\vec{x_i})), \text{ and}$$

$$\hat{f} = \arg\min_{f\in\mathcal{F}} R_{\text{emp}}(f)$$

As it is shown in the above equation, loss functions have an important role in the optimization problem and the obtained classifier. In the following section, we explore different types of loss functions and their useful characteristics.

4. Loss functions

The central components of machine learning algorithms are loss functions indicating which sample and how much it affects the model training. In other words, loss functions attribute to each sample a value, which indicates how much that sample is involved in the optimization problem. For example, If the employed loss function assigns a very large value to an outlier sample, the outlier may negatively impact the model's parameters [25]. To better explain our purpose, three loss functions with almost similar shapes are explained.

As the $0-1$ loss function penalizes all misclassified samples (even outliers) with value 1, it is known as robust. An outlier does not influence much the obtained classifier, which leads to a robust learner. This characteristic is desired for a robust classifier but it does not penalize correct samples with small margins and thus, it cannot be an appropriate choice for applications with margin importance [26].

The Sigmoid function as another type of a loss function is, in appearance, similar to the $0-1$ function. The only difference is that Sigmoid is differentiable while $0-1$ is not. Still, Sigmoid assigns a value (although very small) to correct samples, which means that those samples also participate in the optimization problem, and thus, it leads to a less sparse optimization compared to $0-1$ loss.

Ramp loss function whose shape is almost similar to $0-1$ loss with a minor difference that correct samples with small margins is also penalized by Ramp. This difference makes Ramp loss function more suitable for applications with margin importance [27] compared to $0-1$ loss. On the other hand, since Ramp loss is not differentiable, the optimization process under Ramp loss function is not easy. These three loss functions are shown in Fig. 2.

The ultimate goal of a learning method is the ability of classifying unseen data. Thus, the classifier should be robust to data perturbance. Sometimes training and testing data may be sampled from different processes, which are similar to some extent but are not identical. A more difficult scenario can happen in noisy environments where outliers corrupt the training data, testing data or both. In order to cope with very noisy environments, an effective approach is the use of a robust loss function.

Robust loss function. a loss function is known as robust if a constant k exist that samples with $e_i > k$ do not be given large values by the loss function where e_i denotes the error of ith sample [26].

Correntropy loss function which is defined as follows was first introduced by Liu in 2006. It is a kind of robust loss function and has been applied to many machine learning algorithms including Covoloutional Neural Networks (CNNs), Least Mean Square (LMS) Filters, etc.

$$L_{corr}(\hat{y}_i, y_i) = 1 - \exp\left(-\frac{||y_i - \hat{y}_i||^2}{\sigma^2}\right) \tag{1}$$

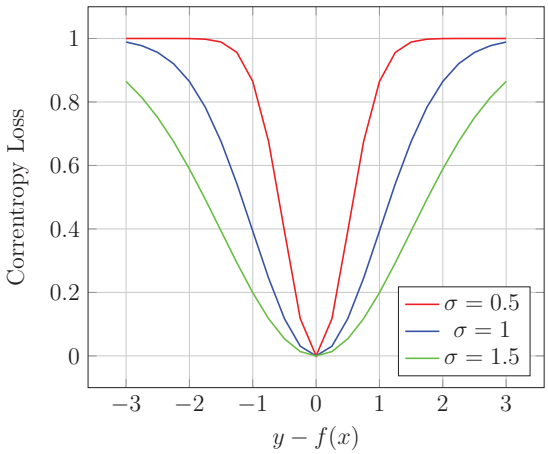

Fig. 3. Correntropy loss function with different values of σ.

where y_i is the label of ith sample, $\hat{y}_i$ is the corresponding estimated label, and σ is the kernel bandwidth. Correntropy loss function with different values of σ is depicted in Fig. 3 and as is illustrated by Fig. 3 the higher bandwidth, the sooner it levels off.

Sometimes, despite the fact that the learned classifier is able to classify the training data well, it fails to estimate unseen data (test), and it leads a high generalization error although the training error is low. A reason for this failure is often known as the overfitting problem, which means that the classifier fits on the training data and lose its ability of generalization. A solution for better generalization is to use a loss function resulting in a more generalized classifier. in the following characteristics of such a loss function are discussed.

Loss functions with better generalization. A loss function results in a more generalized classifier if minimization of expected loss value leads to a enhanced-margin classifier. If a classifier has an enhanced margin, it will work better with the unseen data and would be better generalized. Penalizing correct samples close to the classification line would result in an enhanced classifier [26] and the loss function is called margin enhancing. For example, Zero-One loss function does not penalize correct samples at all and therefore, it is not margin enhancing.

Hinge loss function which is used in SVM algorithm penalizes correct data near the classification line (this is why it is so useful to determine margins), and hence, it is a margin enhancing loss function. The other advantage of Hinge loss function over others is that it does not penalize the correct samples and so, it would lead to dual sparsity (not guaranteed). However, it does not help at probability estimation which is a big advantage of the Logistic loss function [28,29].

To conclude this section, as each loss function comes with its own merits and demerits and there is not yet a comprehensive loss function working well with all applications, we propose the use of an ensemble of losses in our classification model.

5. The proposed method

Let (x, y) be a sample where $x \in \mathcal{R}^d$ is the input and $y \in \{0, 1\}^C$ is one-hot encoding of the label (C is the number of classes). Let θ be the parameters of neural network classifier with a top softmax layer so that the probability estimates are $\hat{y} = softmax(f(x; \theta))$. We make use of a novel objective function which is the combination of Correntropy loss function, Hinge loss function and Cross-Entropy. The weights of the neural network are updated through backpropagation in the training stage. Fig. 4 shows the ANN and the way it is updated through backpropagation.

Our purposed objective function is a combination of Correntropy, Hinge and Cross-Entropy loss functions which are leading to a more robust, generalized and improved probability estimator respectively. Using these three loss function we assume that the classifier performance will increase significantly. Let $\left\{L_j(y, \hat{y})\right\}_{j=1}^{K}$ denote K single loss functions. In addition to finding the optimal parameter of the neural network, the goal is to find the best weights, $\{\lambda_1, \lambda_2, \ldots, \lambda_K\}$, to combine K base loss functions in order to generate a better application-tailored loss function. We need to add a further constraint to avoid yielding near zero values for all λ_i weights. The proposed objective function is defined as below.

$$L = \sum_{j=1}^{K} \lambda_j L_j(y, \hat{y}), \qquad \sum_{j=1}^{K} \lambda_j = 1 \tag{2}$$

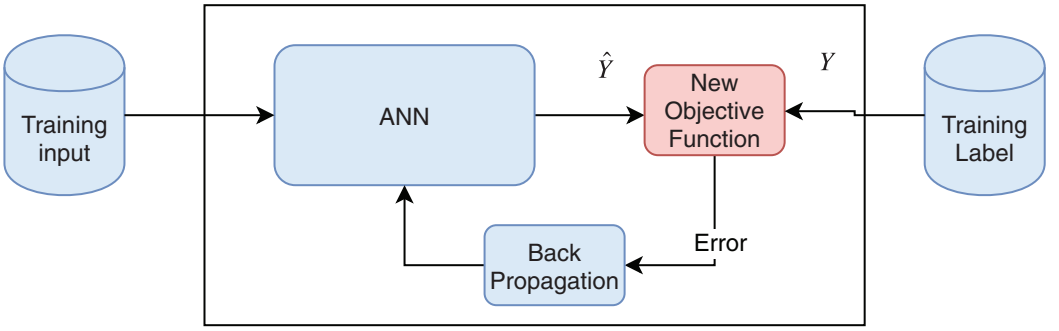

Fig. 4. The proposed ANN's structure.

Table 1
Several well-known loss functions, $z = y' \cdot \hat{y} \in \mathcal{R}$. where y is the true label and $\hat{y}$ is the estimated one.

Name of loss function	$L(y, \hat{y})$
Hinge Loss	$L_H = \begin{cases} 0 & z \geq 1 \\ \max(0, 1-z) & z < 1 \end{cases}$
Cross-Entropy Loss	$L_{C-E} = \log(1 + \exp(-z))$
Correntropy Loss	$L_C = \exp \frac{\|y-\hat{y}\|_2^2}{\sigma^2}$

Eq. (3) shows the optimization, given N training samples

$$\begin{aligned} &\underset{w,\lambda}{\text{minimize}} \sum_{i=1}^{N}\sum_{j=1}^{K} \lambda_j L_j(y_i, \hat{y}_i) \\ &\text{s.t.} \sum_{j=1}^{K} \lambda_j = 1, \qquad \lambda_j \geq 0 \end{aligned} \tag{3}$$

where y_i are the true labels and $\hat{y}_i$ are the estimated label. To omit $\lambda_j \geq 0$, we use λ_j^2 instead of $lambda_j$ The resulting optimization problem is

$$\begin{aligned} &\underset{w,\lambda}{\text{minimize}} \sum_{i=1}^{N}\sum_{j=1}^{K} \lambda_j^2 L_j(y_i, \hat{y}_i) \\ &\text{s.t.} \sum_{j=1}^{K} \lambda_j^2 = 1 \end{aligned} \tag{4}$$

We then incorporate the constraint, $\sum_{j=1}^{K} \lambda_j^2 = 1$ as a regularization term based on the concept of Augmented Lagrangian. The modified objective function using Augmented Lagrangian is presented in Eq. (5) [25].

$$\begin{aligned} &\underset{w,\lambda}{\text{minimize}} \sum_{i=1}^{N}\sum_{j=1}^{K} \lambda_j^2 L_j(y_i, \hat{y}_i) + \\ &\eta_1 \left(\sum_{j=1}^{K} \lambda_j^2 - 1 \right) + \eta_2 \left(\sum_{j=1}^{K} \lambda_j^2 - 1 \right)^2 \end{aligned} \tag{5}$$

Note that η_2 must be significantly larger that η_1. $lambda_j$ and the weights of network are learned through backpropagation using the gradient descent method. In the new section, we provide more details on the new objective function.

5.1. The objective function

In particular, the objective function is a linear combination of three loss functions: 1) cross-entropy, 2) correntropy, and 3) hinge. The formula of each loss is shown in Table 1. The ensemble loss function inherits the advantages of each.

1. **Cross-Entropy** is a very popular loss function in neural networks. The minimization of the average value of the cross-entropy loss function over data means maximization of the conditional log-likelihood of the data.
2. **Correntropy** is less sensitive to outliers since it is a bounded function. The Correntropy loss, therefore, leads to robust classifiers. Section 4 contains more explanation on robust loss functions.

Table 2
Some Statistics on WBCD dataset.

Attributes	Mean	Standard error	Maximum
Radius	6.98-28.11	0.112-2.873	7.93-36.04
Texture	9.71-39.28	0.36-4.89	12.02–49.54
Symmetry	0.106-0.304	0.008-0.079	0.157-0.664
Perimeter	43.79–188.50	0.76-21.98	50.41–251.20
Compactness	0.019-0.345	0.002-0.135	0.027-1.058
Smoothness	0.053-0.163	0.002-0.031	0.071-0.223
Concavity	0.000-0.427	0.000-0.396	0.000–1.252
Fractal dimension	0.050-0.097	0.001-0.030	0.055-0.208
Area	143.50–2501.00	6.80–542.20	185.20–4254.00
Concave points	0.000-0.201	0.000-0.053	0.000-0.291

Table 3
Network's configuration.

Description	Values
Hidden Layers	16, 16, 16
Activation function	ReLu
Dropout rate	0.1
Mini-Batch size	128
Regularization weight	0.1
number of epochs	150
Correntropy's kernel width	0.1

3. **Hinge** is margin enhancing which leads to a better generalization. The character of a loss function causing better generalization have been fully explained in Section 4.

In the following section, some experiments on benchmark datasets are provided.

6. Experiments

We evaluate our method on a benchmark dataset of Wisconsin Breast Cancer Diagnostic (WBCD) from the Unversity of California- Irvine repository. The WBCD data set contains 32 features in 10 categories for each cell nucleus, which are texture (standard deviation of gray-scale values), radius (mean of distances from center to points on the perimeter), smoothness (local variation in radius lengths), perimeter, area, concave points (number of concave portions of the contour), c symmetry, compactness ($\text{perimeter}^2/\text{area} - 1.0$), concavity (severity of concave portions of the contour) and fractal dimension ("coastline approximation"- 1). We have calculated maximum value, standard error and mean value for every category reported in Table 2. Those different measurements are treated as different features in the dataset. As scales of different features are not the same, the error function is affected by large-scale values. Therefore, we first normalize the data. The dataset contains 569 instances and true labels. Fig 5 and Fig 6 provides some statistics about on dataset. As is illustrated by Fig. 6 the dataset is quite balanced. We first normalize all the data so that they are between $[-1, 1]$.

While the proposed cost function can be incorporated into other classifiers, we have applied it into a neural network classifier to detect cancer. it works good on WBCD dataset but its speed may reduce regarding high dimensional data as it does not use any feature reduction technique.

The proposed objective function is incorporated into a neural network with three hidden layers contains 16 neurons followed by a dropout layer. We also use 5 fold cross-validation to tune the network's parameters. Table 3 shows the hyper parameters' values.

We use precision, recall, f1-Score and accuracy to compare our models with the peers one. The diagnosis accuracy is defined as follows

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

which FP denotes False Positive, TN is True Negative, TP: True Positive and FN denotes False Negative.

Table 4 reports experimental results in a normal situation. RBF was used for the kernel SVM with the gamma and C parameters selected by grid search from [1,0.1,0.01,0.001] and [0.1,1,10,100] respectively. 5 neighbours has been considered for kNN algorithm. At normal situation (without adding any noise), as the data are not very complicated. all the classifiers achieve acceptable results with Kernel SVM and our proposed method obtaining the best result (0.97). Kernel SVM uses a kernel to cope with the nonlinearity of data and also the Hinge loss function which leads to an enhanced margin classifier, so it improves the performance in the face of unseen data (test data). The proposed approach which uses 3 loss function, which each of them is to handle one difficulty (margin enhancing, robustness) also achieves the best result as well.

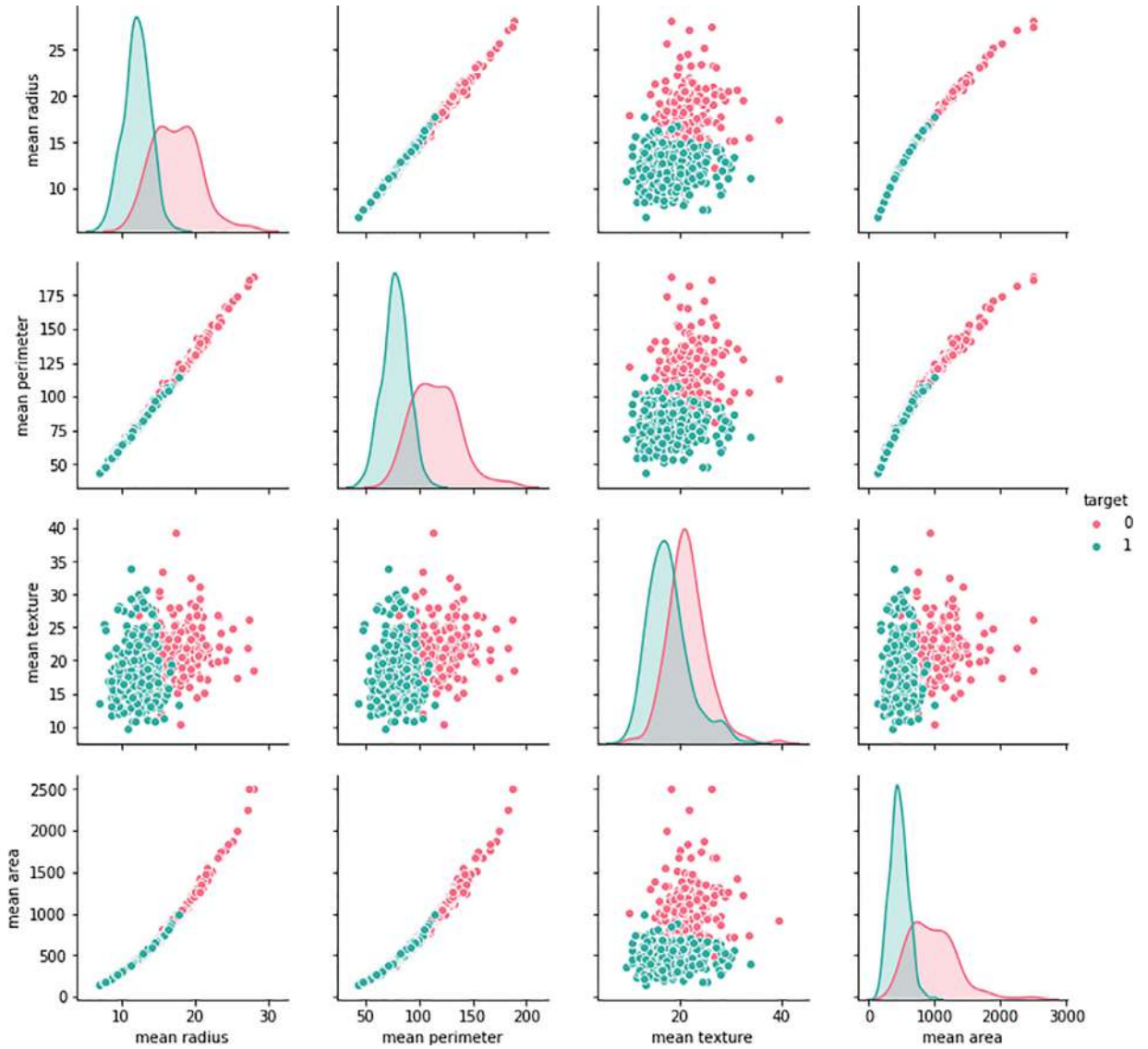

Fig. 5. Visualizing selected features by target.

As is explained in Section 5.1, Correntropy, Cross-Entropy and Hinge loss functions are selected for the individual losses and used to form the proposed ensemble loss function. Table 5 reports the results obtained by the ensemble loss compared to each individual loss function. The ensemble loss achieves the best result.

We also explore the contribution of each loss function in forming the ensemble loss function. As the constraints included in Eq. (4) are not hard ones, $\sum_{j=1}^{3} \lambda_j^2 \neq 1$. So we re-scaled the λ_j to satisfy $\sum_{j=1}^{3} \lambda_j^2 = 1$ and the weights are reported in Table 6.

To explore the robustness of the proposed method, we add label noise to the training samples. We investigate the performance of our method on data corrupted with label noise. To add label noise, we randomly select 10% and 30% of samples and change their labels. The accuracy obtained in the face of label noise is reported in Table 7. The best results are presented with boldface. As is shown in Table 7, all the classifiers experience decrease respect to label noise while our proposed objective function performed well. since we use an ensemble of loss function as the objective function. Correntropy loss as a part of our proposed objective function works well in the face of outliers as it does not assign high value to the samples with large error. Hence, those samples do not pull the classification hyperplane towards themselves. In contrary kNN which is very prone to the noise achieves the worst result as it discovers the label of each sample based on its k nearest samples. So noisy samples affect their neighbours and the errors are somehow propagated.

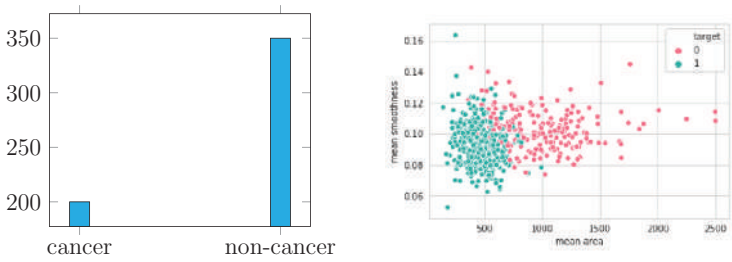

Fig. 6. Class distribution.

Table 4
Results on WDBC dataset.

Method	Label	Precision	Recall	f1-Score	Accuracy
Ours	0	1.00	0.94	0.97	**0.97**
	1	0.96	1.00	0.98	
kNN	0	1.00	0.85	0.92	0.94
	1	0.90	1.00	0.95	
Kernel-SVM [30]	0	1.00	0.94	0.97	**0.97**
	1	0.96	1.00	0.98	
Random Forest	0	1.00	0.83	0.91	0.93
	1	0.89	1.00	0.94	
SVM	0	1.00	0.90	0.95	0.96
	1	0.93	1.00	0.96	

Table 5
Test accuracy with 5-fold cross-validation on WDBC.

Loss Function	Accuracy
Correntropy	0.95
Cross-Entropy	0.96
Hinge	0.93
Ensemble	0.97

Table 6
Weights of each loss function in the ensemble loss.

Loss Function	Weights
Correntropy	0.29
Cross-Entropy	0.36
Hinge	0.34

Table 7
Accuracy for data with label noise.

Method	Accuracy (10% label noise)	Accuracy (30% label noise)
Ours	0.96	0.96
kNN	0.91	0.87
Kernel-SVM	0.94	0.93
Random forest	0.92	0.91
SVM	0.95	0.94

7. Conclusion

This paper proposed a new objective function which is a combination of three loss functions (Correntropy, Hinge and Cross-Entropy) for cancer prediction. This work focused on breast cancer classification task. Correntropy leads to a robust classifier while hinge loss function improves the generalization. We have incorporated this objective function to a very simple neural architecture, but in principle, the proposed method can be used in any other neural network.

In the task of breast cancer prediction, the proposed method shows improvements compared to other machine learning methods listed in Table 4 since our approach encourages generalization and robustness via the proposed objective function. Furthermore, the approach is more robust against the presence of label noise compared to others. Moreover, according to our experiments, the gradient descent method quickly converged.

The weights of each individual loss and also the network's parameters are simultaneously learned through error back-propagation using the gradient descent technique. In future works, we will explore other frameworks in order to learn the weights of the loss functions and the network's parameter separately. The proposed approach does not make the loss functions' weights sparse, by doing so the generalization will improve as well as the time complexity.

Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Hamideh Hajiabadi: Conceptualization, Methodology, Software. **Vahide Babaiyan:** Writing - original draft. **Davood Zabihzadeh:** Writing - review & editing. **Moein Hajiabadi:** Supervision.

References

[1] Siegel R, Naishadham D, Jemal A. Cancer statistics for hispanics/latinos, 2012. CA Cancer J Clin 2012;62(5):283–98.
[2] Milukow HA, Binns AD, Adamowski J, Bonakdari H, Gharabaghi B. Estimation of the Darcy–Weisbach friction factor for ungauged streams using gene expression programming and extreme learning machines. J Hydrol (Amst) 2019;568:311–21.
[3] Ebtehaj I, Sattar AM, Bonakdari H, Zaji AH. Prediction of scour depth around bridge piers using self-adaptive extreme learning machine. J Hydroinf 2016;19(2):207–24.
[4] Karami H, Karimi S, Bonakdari H, Shamshirband S. Predicting discharge coefficient of triangular labyrinth weir using extreme learning machine, artificial neural network and genetic programming. Neural Comput Appl 2018;29(11):983–9.
[5] Hebli A, Gupta S. Brain tumor prediction and classification using support vector machine. In: Proceedings of the international conference on advances in computing, communication and control (ICAC3). IEEE; 2017. p. 1–6.
[6] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144(5):646–74.
[7] Alickovic E, Subasi A. Normalized neural networks for breast cancer classification. In: Proceedings of the international conference on medical and biological engineering. Springer; 2019. p. 519–24.
[8] Aslan MF, Celik Y, Sabanci K, Durdu A. Breast cancer diagnosis by different machine learning methods using blood analysis data. Int J Intell Syst Appl Eng 2018;6(4):289–93.
[9] Burt JR, Torosdagli N, Khosravan N, RaviPrakash H, Mortazi A, Tissavirasingham F, et al. Deep learning beyond cats and dogs: recent advances in diagnosing breast cancer with deep neural networks. Br J Radiol 2018;91(1089):20170545.
[10] Polat K, Sentürk U. A novel ML approach to prediction of breast cancer: combining of mad normalization, KMC based feature weighting and adaboostm1 classifier. In: Proceedings of the 2nd international symposium on multidisciplinary studies and innovative technologies (ISMSIT). IEEE; 2018. p. 1–4.
[11] Qiu C, Jiang L, Cao Y, Hu C, Yu Y, Zhang H. Factors associated with de novo metastatic disease in invasive breast cancer: comparison of artificial neural network and logistic regression models. Transl Cancer Res 2019;8(1):77–86.
[12] Yue W, Wang Z, Chen H, Payne A, Liu X. Machine learning with applications in breast cancer diagnosis and prognosis. Designs 2018;2(2):13.
[13] Nilashi M, Ibrahim O, Ahmadi H, Shahmoradi L. A knowledge-based system for breast cancer classification using fuzzy logic method. Telemat Informat 2017;34(4):133–44.
[14] Onan A. A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. Expert Syst Appl 2015;42(20):6844–52.
[15] Bhardwaj A, Tiwari A. Breast cancer diagnosis using genetically optimized neural network model. Expert Syst Appl 2015;42(10):4611–20.
[16] Asri H, Mousannif H, Al Moatassime H, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput Sci 2016;83:1064–9.
[17] Abdel-Ilah L, Šahinbegović H. Using machine learning tool in classification of breast cancer. In: Proceedings of the CMBEBIH. Springer; 2017. p. 3–8.
[18] Thein HTT, Tun KMM. An approach for breast cancer diagnosis classification using neural network. Adv Comput 2015;6(1):1.
[19] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17.
[20] Tahmooresi M, Afshar A, Rad BB, Nowshath K, Bamiah M. Early detection of breast cancer using machine learning techniques. J Telecommun Electr Comput Eng (JTEC) 2018;10(3-2):21–7.
[21] Omondiagbe DA, Veeramani S, Sidhu AS. Machine learning classification techniques for breast cancer diagnosis. In: Proceedings of the IOP conference series: materials science and engineering, 495. IOP Publishing; 2019. p. 012033.
[22] Sáez JA, Krawczyk B, Woźniak M. On the influence of class noise in medical data classification: treatment using noise filtering methods. Appl Artif Intell 2016;30(6):590–609.
[23] Hajiabadi H, Godaz R, Ghasemi M, Monsefi R. Layered geometric learning. In: Proceedings of the international conference on artificial intelligence and soft computing. Springer; 2019. p. 571–82.
[24] Holland MJ, Ikeda K. Minimum proper loss estimators for parametric models. IEEE Trans Signal Process 2016;64(3):704–13.
[25] Hajiabadi H, Molla-Aliod D, Monsefi R. On extending neural networks with loss ensembles for text classification. In: Proceedings of the Australasian language technology association workshop; 2017.
[26] Hajiabadi H., Monsefi R., Yazdi H.S.. Relf: robust regression extended with ensemble loss function. arXiv:1810110712018;.
[27] Tang L, Tian Y, Yang C, Pardalos PM. Ramp-loss nonparallel support vector regression: robust, sparse and scalable approximation. Knowl Based Syst 2018.
[28] Moore R, DeNero J. L1 and l2 regularization for multiclass hinge loss models. In: Proceedings of the symposium on machine learning in speech and language processing; 2011.
[29] Yan K, Li Z, Zhang C. A new multi-instance multi-label learning approach for image and text classification. Multimed Tools Appl 2016;75(13):7875–90.
[30] Zheng B, Yoon SW, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. Expert Syst Appl 2014;41(4):1476–82.

Hamideh Hajiabadi received her degree in Software Engineering from Ferdowsi University of Mashhad and her current fields of study is in Machine Learning, Image Processing and Natural Language processing.

Vahide Babaiyan received her degree in computer networking from Shiraz University of Technology in 2013. Her research interests include IoT, Data mining and WSN.

Davood Zabihzadeh is a faculty member at Sabzevar University of New Technology and received his degree in Software Engineering from Ferdowsi University of Mashhad. His current research area is in Machine Learning.

Moein Hajiabadi received his degree in Electrical Engineering from Birjand University and his recent works are on applying machine learning techniques to problems.