Pareto-Optimal Multi-Objective Dimensionality Reduction Deep Auto-Encoder for Mammography Classification

Saeid Asgari Taghanaki*, Jeremy Kawahara, Brandon Miles, Ghassan Hamarneh

Medical Image Analysis Lab, Simon Fraser University, Canada

Abstract

Background and Objective: Feature reduction is an essential stage in computer aided breast cancer diagnosis systems. Multilayer neural networks can be trained to extract relevant features by encoding high-dimensional data into low-dimensional codes. Optimizing traditional auto-encoders works well only if the initial weights are close to a proper solution. They are also trained to only reduce the mean squared reconstruction error (MRE) between the encoder inputs and the decoder outputs, but do not address the classification error. The goal of the current work is to test the hypothesis that extending traditional auto-encoders (which only minimize reconstruction error) to multi-objective optimization for finding Pareto-optimal solutions provides more discriminative features that will improve classification performance when compared to single-objective and other multi-objective approaches (i.e. scalarized and sequential).

Methods: In this paper, we introduce a novel multi-objective optimization of deep auto-encoder networks, in which the auto-encoder optimizes two objectives: MRE and mean classification error (MCE) for Pareto-optimal solutions, rather than just MRE. These two objectives are optimized simultaneously by a non-dominated sorting genetic algorithm.

*Corresponding author

Email addresses: sasagarit@sfu.ca (Saeid Asgari Taghanaki*), jkawahar@sfu.ca (Jeremy Kawahara), bmiles@sfu.ca (Brandon Miles), hamarneh@sfu.ca (Ghassan Hamarneh)

Results: We tested our method on 949 X-ray mammograms categorized into 12 classes. The results show that the features identified by the proposed algorithm allow a classification accuracy of up to 98.45%, demonstrating favourable accuracy over the results of state-of-the-art methods reported in the literature.

Conclusions: We conclude that adding the classification objective to the traditional auto-encoder objective and optimizing for finding Pareto-optimal solutions, using evolutionary multi-objective optimization, results in producing more discriminative features.

Keywords: Breast cancer, computer aided diagnosis, feature reduction, auto-encoder, multi-objective optimization

1. Introduction

Although mammography is an effective modality for early breast cancer detection and diagnosis, on mammographic examinations, 10-30% of cancerous/noncancerous lesions may be misinterpreted [1]. To overcome this, computer aided diagnosis (CADx) systems have been developed. The accuracy of CADx for x-ray breast mammography still requires improvements to be useable as a flawless guide (an alert system flagging potential misclassification to the human operator) for radiologists or an independent clinical interpreter [2, 3]. Recently, CADx systems have been developed to help radiologists classify suspicious lesions, e.g., labeling the lesion according to the Breast Imaging Reporting And Data System (BI-RADS) assessment categories [4]. In order to build a classification system, a large number of features can be calculated from mammograms, but using high-dimensional features with relatively few training samples can lead to the classifier over-fitting to the training data. This can degrade the predictive model performance as well as having a high computational cost. Since features in mammograms can be noisy and/or highly correlated with each other, feature transformation and reduction is often used to extract relevant features with high discriminatory power from a large number of potential candidate features [5, 6].

Related mammography classification works. Mohanty et al. [7] designed an association rule mining based mammogram classification procedure to classify the extracted and hypothetically selected gray level co-occurrence matrix (GLCM) features. This method requires an accurate set of association rules between the features and labels to be defined. The high number of features required for breast cancer diagnosis makes defining these rules a difficult task, which may results in a large number of irrelevant associations. In one of the most recent works, Bria et al. [8] proposed a classification system based on a cascade boosting classifier. The authors defined 145 features to describe the micro-calcifications but only discriminated between 2 classes, which are insufficient to address the variety of BI-RADS classes. Oliver et al. [9] designed a CAD system using PCA and Bayesian combination of kNN and C4.5 classifiers. This was tested on 184 selected views (all with at least one mass) from a private digital mammographic dataset. They proved that considering density information influences the performance of CAD systems for the detection of breast masses. Without considering breast density information they obtained a 92% accuracy, while by taking density information into account, they achieved an accuracy of 94%. Subshini et al. [10] selected 43 mammograms from the MIAS database and preprocessed them to remove the pectoral muscle and radiopaque artifacts. Next, they extracted statistical features e.g., entropy, uniformity, standard deviation and others from the filtered images for breast characterization. Then, using a SVM classifier they classified the data into three classes of breast density achieving an accuracy of 95.44%. Verma et al. [11] selected 200 ROIs from the DDSM dataset and extracted several features like mass margin, density, patient age, mass shape, subtlety value, and abnormality assessment rank. They proposed to classify the ROIs into two classes of benign and malignant with a soft-clustered direct learning algorithm. An accuracy of 97% was obtained by their proposed method. CAD was applied to standard mammograms from 127 cases in Sadaf et al. [12]. The authors analyzed the CAD sensitivity under 10 classes based on mode of presentation, breast density, lesion size, lesion type, and histopathology. Their overall CAD sensitivity was 91% (115 of 127 cases).

Deserno et al. [13] used 2,796 patches and defined 12 classes based on BI-RADS assessment categories, BI-RADS tissue density classes, and type of lesion. For feature extraction they applied PCA, 2DPCA, and SVM. Finally, they tested a SVM with three different kernels as a classifier. The best result observed was 80% using 2DPCA feature extraction and a SVM with a Gaussian kernel.

1.1. From Shallow to Deep Learning Dimensionality Reduction

A significant amount of research has focused on shallow learning approaches such as support vector machines (SVM) [14], principal component analysis (PCA) [15], and linear discriminant analysis (LDA) [16]. Although SVMs are relatively easy to optimize and have good performance for feature transformation/reduction on continuous balanced data, even with advanced kernels, they do not perform well on imbalanced data which results in producing sub-optimal solutions [17]. Similar to PCA, the linearity and the underlying Gaussian assumption of LDA renders the LDA projections incapable of discriminating complex nonlinear data with non-Gaussian distributions. In 2006, the situation was changed by Hinton et al.'s revolutionary research on deep belief networks [18] along with work by Bengio et al. [19] and Poultney et al. [20]. This sparked a significant research effort into deep learning focused on solving the problems of training multiple layers in deep networks and improving initialization. To address this, several optimizations were proposed e.g., unsupervised greedy layer-wise pre-training of each layer [21], stochastic gradient descent methods, limited memory BFGS (L-BFGS) and conjugate gradient [22]. In recent years, deep learning strategies have been significantly improved. For a more detailed review on deep learning the reader is referred to [23–25].

1.1.1. Auto-encoders

Auto-encoders (AEs) encode high-dimensional input data into low-dimensional output codes and then recover the original data from the codes. Bengio et al., motivated the use of restricted Boltzmann machines (RBMs) as pre-training for AEs to build a deep structure [23]. To improve reconstruction fidelity, regularization of AEs was proposed. This can be divided into three models: sparse auto-encoders (SAEs), denoising auto-encoders (DAEs), and contractive auto-encoders (CAEs). SAEs were introduced by Ranzato et al. [26] and inspired by Bengio et al.s stacked AEs [27]. Sparsity of the representation could be obtained either by penalizing the hidden unit biases or by direct penalization of the hidden unit outputs. However, this penalty bias can potentially cause the weights to compensate for the bias, which weakens numerical optimization [24]. Vincent et al. [28] proposed DAE to modify the learning procedure from only reconstructing the raw data to reconstructing the corrupted (noisy) version of the data. These auto-encoders are optimized to, first, encode the noisy input data and second, recover the original input. A stacked denoising AE (SdAE) is constructed by stacking layers of DAEs. They utilize an additional layer to minimize the classification error, however, this is done sequentially not simultaneously [29]. CAEs [30] are an extension of DAEs, as they add a contractive penalty to the reconstruction error function, which penalizes attributes sensitivity to input variations. The fundamental weakness of the CAEs penalty is that it only considers the minuscule variations of input [24]. This was partially improved in [31], but not fully addressed.

Related multi-objective (semi-supervised) autoencoders. In the following paragraphs we focus on detailing the most relevant works. AEs have traditionally been used to perform unsupervised (i.e. without considering the classification task at hand) dimensionality transformation and reduction [24]. This process requires a large amount of unlabeled data samples to produce good feature encodings for reconstruction. However, this process may fail to capture the relevant class information in the data [26]. To reduce the requirement for input data and to find a more meaningful link between the unlabeled data and a classification problem, semi-supervised variants of AE were proposed [29, 32, 33], i.e. techniques that minimize both reconstruction error and classification error (either in sequential steps or by combining the multi-objective function into a scalarizing function).

Socher et al. [32] introduced a semi-supervised greedy recursive AE, in which a scalar cost function, summing up both reconstruction error and cross-entropy-based classification error, was used. They applied the L-BFGS algorithm for optimization. However, the L-BFGS is highly dependent on the pre-conditioner to avoid degenerating to the steepest descent method [34]. Furthermore, their method required careful tuning of a user-defined parameter that weighs the contributions of the reconstruction and the cross-entropy error terms. Similarly, other researchers [26, 35–37] translated the multi-objective problem into a single-objective scalar function. More specifically, their scalarization (weighted-sum) method minimizes a positively weighted convex sum of the objectives (reconstruction error and discriminative error). However, as mentioned by Goldberg and Holland, [38]: "there are times when several criteria are present simultaneously and it is not possible (or wise) to combine these into a single number". Moreover, it is difficult to define a set of appropriate weights to control the scalarized function to produce *Pareto-optimal solutions* (a solution is considered as a *pareto optimal* if it is not dominated by other solutions) [39]. Although several weight optimizations have been previously introduced e.g. [40], it is a complicated task to find relevance between the weights. Moreover, conducting several weight optimizations is computationally expensive. Additionally, the scalarization method suffers from two technical drawbacks. First, the relation between the *Pareto curve* and the objective function weights is a monotone spread of weight parameters, however, it does not generally produce uniformly distributed points on the *Pareto curve*. Second, minimizing the convex combinations of the objective functions does not necessarily result in reaching the non-convex portions of the *Pareto set* [41]. More recently, Almousli et al. [33] changed the cost function of the DAE in order to produce more accurate results for a supervised task. They included a matrix in the cost function instead of the Euclidean-norm to reflect the dependency between the input vector x and the label y. We note that there is no unique definition for the mentioned matrix and thus a careful examination is required to define a proper matrix for different problems. In this paper, we test the hypothesis that extending traditional auto-encoders (which only minimize reconstruction error) to multi-objective optimization for finding Pareto-optimal solutions outperforms traditional single objective and other types of multi-objective (i.e. scalarization and sequential) autoencoders. The SdAE method [29] is designed to reduce both the misclassification rate (via its last layer) and reconstruction error. However, SdAE does so in two distinct sequential steps and suffers from two significant drawbacks: its high computational cost and failure at high-dimensional mapping [42].

1.2. Contributions

In addition to being the first work that adopts a deep AE for multi-class imbalanced mammographic data, our first technical contribution of this work is introducing a multi-objective (MO) AE that transforms and reduces the dimension of features. Our method leverages the mean classification error (MCE) criterion to influence the learning process and in the same time (i.e. not sequentially as in [29]) considers the MRE to find the best Pareto solutions. The result is a deep network trained to produce the most informative features using both MRE and MCE objectives.

Another issue with AEs is the weight learning process and convergence. Despite their ability to represent highly nonlinear mappings and perform dimensionality reduction, it is difficult to optimize the weights in multilayer AEs. An AE often uses a back-propagation method for learning the network. However, using back-propagation may prove problematic for training networks; as errors get back-propagated all the way to the initial layers, they become insignificant and inconsequential [18]. The result is that the network tends to reconstruct the average of all the observed data. Further, they may be entrapped in local minima. Adopting alternative optimization algorithms, such as evolutionary algorithms (EA) and RBM pre-training, improve the AEs but still suffers from slow convergence and entrapment in local minima [18, 43].

In our second contribution, we handle the *simultaneous* (to find Pareto-optimal solutions) multi-objective optimization (instead of scalarization or optimization in sequential steps). For the weight learning, we utilize the non-dominated sorting genetic algorithm NSGA-II, an improved version of NSGA

[44]. This avoids local minima which commonly arise from the back-propagation method. Compared to other multi-objective evolutionary algorithms, NSGA-II's improved performance can be attributed to its clever ”elitist” strategy for selecting and combining parents and offspring, reducing entrapment in local minima, and its lower time and space (memory) complexity. Moreover, elitism drives the convergence towards a Pareto-optimal set. Diversity and spread of solutions is ensured without the need to apply additional modifiers e.g., sharing parameters [44]. The NSGA-II method considers both the MCE and MRE objectives simultaneously for finding Pareto solutions.

The remainder of this paper is organized as follows. Section 2 introduces our novel multi-objective autoencoder and discusses its structural details. Section 3 presents class definition and implementation details for our method and other related methods from the literature. Section 4 empirically evaluates the methods and discusses the obtained results. Finally, a brief conclusion is provided in Section 7.

2. Proposed Deep Multi-objective Autoencoder

We present a new nonlinear multi-objective AE that transforms and reduces the number of features such that both classification and feature reconstruction errors are minimized. The input is a set of image features X_i extracted from image i (detailed in Section 3) and the output is the set of reconstructed image features $\hat{X}_i$, i.e. equal number of nodes in the input and output layers (Figure 1). The number of hidden (middle) layer nodes is equal to the dimensionality d of the low dimensional representation Y_i. Training the network on X_i leads to a network in which Y_i retains as much information from X_i as possible, i.e. produces $\hat{X}_i$ that is optimally close to X_i from only d variables.

Traditionally, optimality corresponds to minimizing the mean square reconstruction error (MRE):

$$\text{MRE} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}(X_{ij} - \hat{X}_{ij})^2 \tag{1}$$

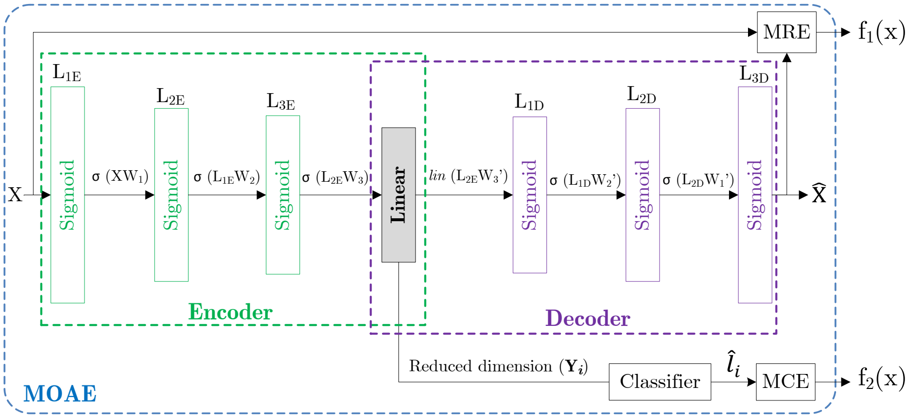


Figure 1: Multi-objective, multi-layered auto-encoder (MOAE) for feature reduction and transformation.

where N is the number of samples, M is the number of features, X_{ij} is j^{th} feature of the i^{th} input sample and $\hat{X}_{ij}$ is the corresponding reconstructed output feature.

Our key contribution here is that we define a second optimality criteria, namely the mean classification error (MCE) objective so the AE considers both the classification and reconstruction error:

$$\text{MCE} = \frac{1}{N}\sum_{i=1}^{N} e_i, \text{ where } e_i = \begin{cases} 0, & l_i = \hat{l}_i \\ 1, & l_i \neq \hat{l}_i \end{cases} \tag{2}$$

l_i is the desired class output of image i, and $\hat{l}_i$ is the classifier's output.

Our auto-encoder employs the sigmoid function for each node in all layers except the last layer of the encoder (first layer of the decoder), which uses a linear transform function. In the training stage, X_i is passed to the network, the encoder module reduces data from X_i to Y_i, and the decoder then reverses the process, increasing the dimensionality of the data with weight transposing. For an optimal mapping, network weights W_i are sought that minimize the MRE and MCE; a faithful reconstruction $\hat{X}_i$ of X_i and, simultaneously, a minimization of the classification error. The classification error is calculated using Y_i as features, the known class labels, and a predetermined classifier (several classifiers are tested in Section 3). The optimal weights are sought using the NSGA-II algorithm (detailed in Figure 2) that minimizes the two objectives $f_1(x) = \text{MRE}, f_2(x) = \text{MCE}$ (Figure 1).

The NSGA-II method works by initializing the population and sorting the population to generate a series of non-dominated Pareto fronts. The first front is the non-dominated set in the current population and the second front is dominated by the first front and continuing for all fronts in the series. For each individual in the population, rank (based on the MCE and MRE) and crowding distance are computed. Crowding distance measures the adjacency of each individual to its neighbors. The larger the average crowding distance, the more diverse the population. Individuals in the first front take the best fitness value (one) and individuals of second front take the second best (two), etc. Once the non-domination ranks and crowding distances are assigned, the evaluation process is started and parents are selected to generate offspring. This is done using binary tournament selection, which works based on the crowded-distance-operator. An individual is picked if its rank is smaller or its crowding distance is greater than all others. Next, the selected population produces offspring using crossover (simulated binary crossover) and mutation (polynomial mutation) operators. The new generation's population is derived from a combined group of offspring and the current population. Therefore, elitism is assured because the best individuals from offspring and parents are selected. Each front is filled in ascending order until reaching the population size. The whole process is iterated until the final generation. For more details refer to [44].

3. Implementation Details

Class Definitions. We adopt the BI-RADS terminology (Table 1) noting both the different tissue breast density classes and assessment classes (Table 1). We note that most of the previous works [9–11, 45–47] have tried to address either the tissue or the assessment classification problem, while similar to [9, 12, 13], we examine a more challenging multi-class problem shown in Table 2,

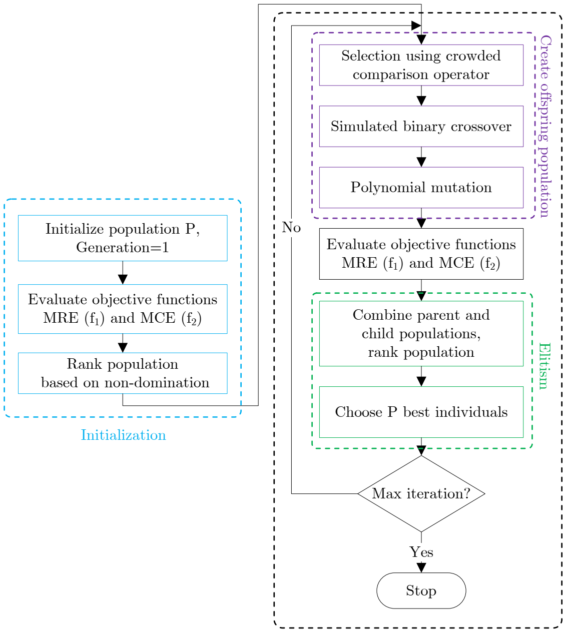

Initialize population P, Generation=1
Evaluate objective functions MRE (f_1) and MCE (f_2)
Rank population based on non-domination
Initialization
Selection using crowded comparison operator
Simulated binary crossover
Polynomial mutation
Create offspring population
No
Evaluate objective functions MRE (f_1) and MCE (f_2)
Combine parent and child populations, rank population
Choose P best individuals
Elitism
Max iteration?
Yes
Stop

Figure 2: Flowchart of NSGA-II for optimizing the two objectives: MRE and MCE.

which enables a classifier to predict both the tissue type and the assessment category at the same time. This approach is important since, as recent research has shown, CAD systems perform poorly when analyzing dense breasts [48, 49]. Note the assessment categories are typically reduced to classes 1, 2 and 5 as they are the most clinically relevant [13].

Table 1: Tissue density classes and assessment categories of BI-RADS.

Class	**Tissue density classes BI-RADS**	**Class**	**Assessment categories BI-RADS**
I	almost entirely fatty	0	need additional imaging evaluation and/or prior mammograms for comparison
II	scattered fibroglandular	1	no findings (negative)
III	heterogeneously dense	2	benign
IV	extremely dense	3	probably benign
		4	suspicious abnormal (biopsy should be considered)
		5	highly suggestive malignant
		6	known biopsy-proven malignant

Class Distribution. In real world learning problems like CADx, samples

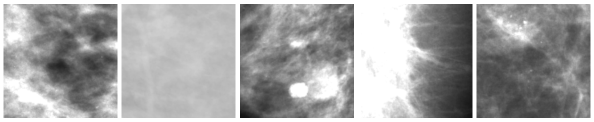


Figure 3: Sample image patches used for experiments

are typically not distributed equally over all classes [50]. This is true for our classes as well (Table 2). It has been shown that a neural network trained from imbalanced data can be biased towards the classes with more samples [51]. To cope with this weakness, dimension (feature) reduction can be very helpful [52, 53]. This motivates our goal of finding a lower dimensionality representation that works well on imbalanced data.

Table 2: Class definitions used for classification and (number of samples).

		Assessment categories		
		1	2	3
Tissue density classes	I	C_1 (107)	C_2 (41)	C_3 (42)
	II	C_4 (179)	C_5 (46)	C_6 (59)
	III	C_7 (48)	C_8 (142)	C_9 (29)
	IV	C_{10} (51)	C_{11} (144)	C_{12} (61)

Feature extraction. In order to reduce computational costs and to have a more invariant and discriminative image representation, we propose to run our experiments on features extracted from 128x128 image patches (Figure 3) instead of evaluating raw pixels. We computed 70 features capturing the textural properties of the mammograms: 48 produced by the segmentation-based fractal analysis (SFTA), which computes the fractal dimensions of decomposed binary images from gray-level mammograms [54], and 22 features based on the Gray Level Co-Occurrence Matrix (GLCM) (Table 3).

Table 3: The 22 extracted GLCM features.

Feature names
Contrast, correlation, cluster prominence, cluster shade, dissimilarity, entropy, homogeneity, maximum probability, maximal correlation coefficient, angular, second moment, sum of squares, variance, sum average, sum entropy, sum, variance, difference variance, difference entropy, information measure of,correlation 1 and 2, inverse difference, inverse difference normalized, and inverse difference moment normalized.

Parameters. The parameters for our 'MOAE-2c' were set as follows. We choose to set the NSGA-II parameters according to [44]. Binary tournament selection using a crowded-comparison operator was applied. Simulated binary crossover (SBX) was used with a probability of 0.9 and polynomial mutation with a probability of 1/n, where n corresponds to the number of weights in the auto-encoder layers. Both distribution indices of crossover and mutation operators were set to 20. The number of nodes in each network layer are computed as, $layer_1 = \lceil(|X_i| \times 1.2)\rceil + 15$, $layer_2 = \lceil(|X_i| \div 2)\rceil + 10$, and $layer_3 = \lceil(|X_i| \div 4)\rceil + 5$, where $|X_i|$ is the number of initial features in a single training example. Training the 'AERBM-1c' was performed using the contrastive divergence algorithm, and a binary stochastic activation function was used. The learning rate was set to 0.1. For 'AEEA-1c/2c', the mutation probability was set to 0.4, the variance of the mutation was set to 1. For the mentioned methods the population size used in the evolutionary optimization was set to 200 and the number of generations was set to 500. We also compared our results to combining the MCE with the MRE in a one-objective AE and optimized with back-propagation ('AEBP-2c') and an evolutionary algorithm ('AEEA-2c') where a softmax classifier layer was added as a final network layer. Accordingly, we have one-objective two criteria (2c) 'AEEA-2c' and 'AEBP-2c' where the error in their cost functions is computed as,

$$Error = \alpha\text{MRE} + (1 - \alpha)\text{MCE} \tag{3}$$

where α is a weight for balancing the two criteria of MRE and MCE which was set to 0.5. Note that for different classifiers e.g., KNN, Naive Bayes, Bayes net, J48 the classification layer is changed based on the classifier used. For all methods mentioned in Table 4, d: the size of our lower dimensional representation (i.e. layer 4 in Figure 1) was varied from 70 to 1 for each method, and the d that obtained the best classification result was selected.

4. Evaluation

Dataset. Our data-set comprises 949 images: 319 from INbreast [55] and 630 from the IRMA (Image Retrieval in Medical Applications) version of DDSM [13]. Adopting Deserno et al.'s class definitions [13], all radiographs were classified by radiologists into 12 classes based on the BI-RADS tissue and assessment categories [4].

Methods evaluated. We compared our proposed multi-objective auto-encoder (MOAE-2c), which optimizes the two criteria (2c) of MRE and MCE, to seven other dimension reduction methods (six AEs and PCA). Table 4 shows the AEs' abbreviations and descriptions. To demonstrate whether the multi-criteria multi-objective optimization improved classification, we included additional one-criterion one-objective methods: 'AEBP-1c', 'AEEA-1c', and 'AERBM-1c'. We also tested combining the two different criteria into a single error by giving equal weight to the MRE and MCE then summing them together: 'AE-BP2c', 'AEEA-2c'. We note that 'SdAE-2c' [29] optimizes MRE and MCE in two separate stages and include it to demonstrate how the proposed multi-objective optimization of MRE and MCE for Pareto-optimal solutions improves our classification results. We also compared with PCA because an auto-encoder will act like PCA when a linear activation function has been used in all neurons (and when a quadratic reconstruction error has been used).

Table 4: The seven implemented methods describing what criteria each method was optimized for and the optimization process

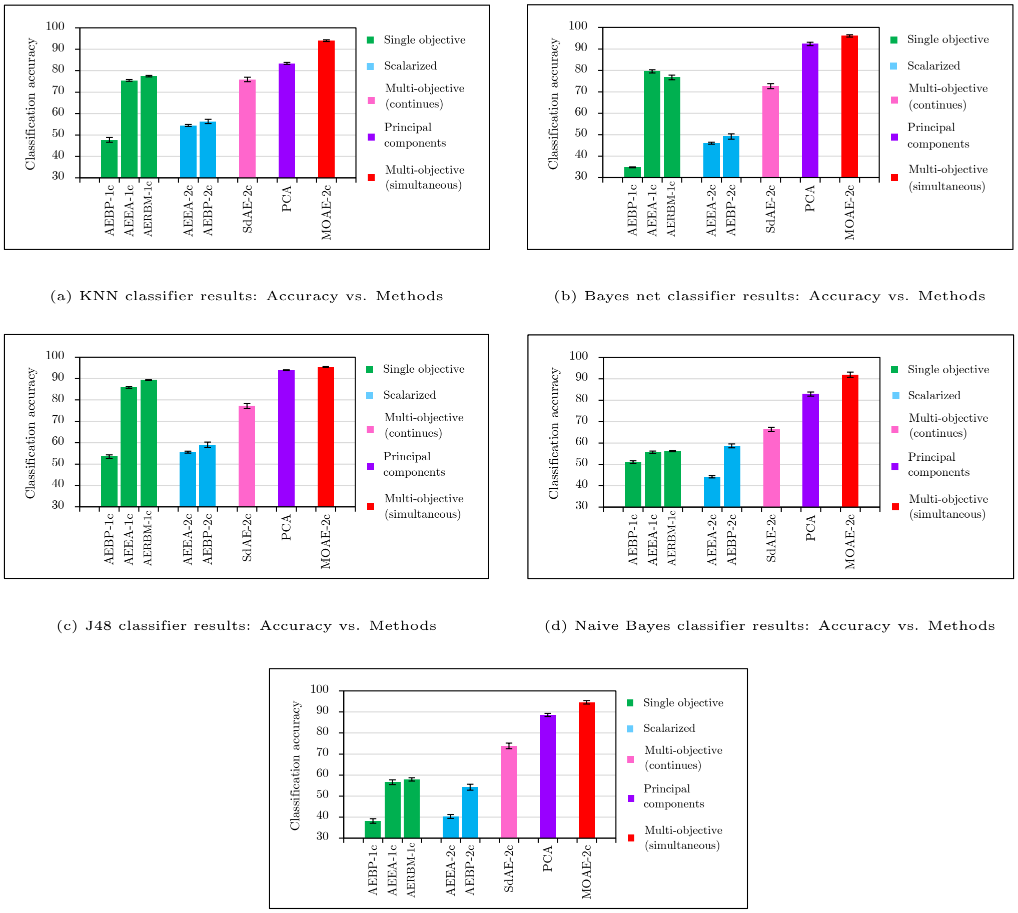
Abbreviation	Description	MRE	MCE	Single-objective	Multi-objective (sequential)	Multi-objective (scalarization)	Multi-objective (Pareto-optimal solutions)
AEBP-1c	AE with back-probagation learning	✓		✓			
AEBP-2c	AE with back-propagation learning	✓	✓			✓	
AEEA-1c	AE with evolutionary optimization	✓		✓			
AEEA-2c	AE with evolutionary optimization	✓	✓			✓	
AERBM-1c	AE pre-trained using RBM	✓		✓			
SdAE-2c	Stacked denoising AE	✓	✓		✓		
MOAE-2c	Proposed AE with multi-objective NSGA-II algorithm optimization	✓	✓				✓

Classifiers tested. To demonstrate how the proposed method is not sensitive to the choice of classifier, five different classifiers were used to test the effectiveness of the transformed features in diagnosing new samples: (i) k-nearest neighbor (a distance-based classifier); (ii) Naive Bayes and (iii) Bayes net (Bayesian classifiers); (iv) J48 (a decision tree classifier) and (v) Softmax (a Softmax regression model). Ten-fold cross validation was adopted.

5. Results

Classification performance I: Accuracy with iterative feature reduction. We evaluated all feature reduction methods over all 70 extracted features. For the auto-encoders, we reduced the number of nodes in the fourth hidden layer (Figure 1) iteratively starting from 70 nodes down to a single node. Applying $MRE + MCE$ adds regularization that results in a model that generalizes better to unseen data by reducing over-fitting. However, in order to make sure that all the models described in Table 4 avoid over-fitting and have a fair comparison, we applied 10-fold cross-validation. Figure 4 shows the accuracy using 10-fold cross-validation. This figure also reports comparisons of the MOAE-2c method versus alternative methods.

Classification performance II: Area under ROC, area under PRC, F-measure and overall Accuracy. To examine the influence of the different feature reduction methods and to provide comprehensive analysis, the classification performance using the four different classifiers was evaluated via: (i) area



(a) KNN classifier results: Accuracy vs. Methods

(b) Bayes net classifier results: Accuracy vs. Methods

(c) J48 classifier results: Accuracy vs. Methods

(d) Naive Bayes classifier results: Accuracy vs. Methods

(e) Softmax classifier results: Accuracy vs. Methods

Figure 4: Dimension reduction results (classification accuracy) for different classifiers. We compare our multi-objective AE (MOAE-2c) with four other groups: single objective AEs; scalarized two criterion AEs; sequential multi-objective AE; and principal component analysis. The error bars represent the standard error over 10 cross-validation rounds.

under the receiver operator characteristic (AUROC); (ii) area under precision-recall curve (AUPRC); (iii) F-measure; and (iv) accuracy percentage (Table 5). Note that the best results of each algorithm during the 70 executions (reduction steps) are reported. The values marked in bold represent the best results and underlined values are the second best. The ROC, PRC and F-measure values are averaged over all classes. The last column of Table 5 shows the accuracy rate when all features are used, i.e. feature transformation without any reduction.

Accuracy comparison with related works reported in literature. Table 6 summarizes selected state of the art methods applied to this classification problem. It is important to note that most of these methods have been tested with only 2 to 4 classes and most focused only on data with either lesion or tissue classification problems. In contrast, we addressed the 12 hybrid classes based on BI-RADS with both lesion and tissue problems. Even with a larger set of features and more classes, which makes the classification task more complex, our method shows improved overall accuracy. Note that all the methods listed in Table 6, have been discussed in the introduction section.

Table 6: Comparison with performance reported in related literature

Author and (year)	Problem	No. of classes	Classifier	Data	Features	Accuracy
Bria et al. (2014)[8]	Lesion	2	Cascade boosting model	1599, private	Shape, topological, score, texture	93%
Mohanty et al. (2013)[7]	Lesion	2	Association rule mining	265, DDSM	GLCM	96.6%
Subashini et al. (2010)[10]	Tissue	3	SVM with RBF kernel	43, private	ROI	95%
Verma et al. (2010)[11]	Lesion	2	Soft-clustered direct learning	200, DDSM	ROI	97%
Oliver et al. (2010)[9]	Both	4	PCA and Bayesian combination of kNN and C4.5	184, private	ROI	92% 94%
Sadaf et al. (2011)[12]	Both	10	Non-statistical analysis	127, private	ROI	91%
Deserno et al. (2011)[13]	Both	12	Support vector machine (Gaussian kernel)	2796, IRMA	ROI	80%
Present work	Both	12	Naive Bayes, J48, KNN, Bayes net	949, INbreast, DDSM	GLCM, SFTA	96.28%, 98.45%

6. Discussion

Classification Performance I. From Figure 4 we can see that AEBP-2c results in a greater accuracy than AEBP-1c, suggesting that adding the MCE criterion improves the classification. As expected, the SdAE-2c method achieved

Table 5: Classification performance II results obtained by different classifiers

Classifier	Methods	AUROC	AUPRC	F-measure	Number of features	Accuracy of reduced data	Accuracy of whole data
KNN	AEBP-1c	0.9 ± 0.1	0.672 ± 0.275	0.640.27	37	75.431	43.1034
	AEBP-2c	0.85 ± 0.16	0.526 ± 0.286	0.504 ± 0.334	44	66.2577	60.1227
	AEEA-1c	0.973 ± 0.028	0.884 ± 0.094	0.819 ± 0.126	17	82.353	72.1362
	AEEA-2c	0.82 ± 0.14	0.432 ± 0.282	0.426 ± 0.303	9	59.5092	50.3067
	AERBM-1c	0.98 ± 0.024	0.904 ± 0.096	0.826 ± 0.097	44	82.662	79.257
	SdAE-2c	0.905 ± 0.101	0.742 ± 0.192	0.758 ± 0.19	45	85.3448	73.2759
	PCA	0.991 ± 0.014	0.944 ± 0.01	0.888 ± 0.124	26	88.854	79.5666
	MOEA-2c	**0.992 ± 0.026**	**0.975 ± 0.028**	**0.962 ± 0.055**	21	**96.285**	**93.4985**
Bayes net	AEBP-1c	0.99 ± 0.01	0.863 ± 0.128	.78 ± 0.17	37	83.6207	37.069
	AEBP-2c	0.517 ± 0259	0.476 ± 0.362	0.467 ± 0362	44	69.3252	61.3497
	AEEA-1c	0.991 ± 0.012	0.935 ± 0.087	0.917 ± 0.11	64	91.641	80.4954
	AEEA-2c	0.88 ± 0.08	0.391 ± 0.217	0.335 ± 0.299	8	54.6012	46.6258
	AERBM-1c	0.985 ± 0.019	0.897 ± 0.081	0.84 ± 0.12	64	84.52	82.9721
	SdAE-2c	0.945 ± 0.149	0.814 ± 0.303	0.802 ± 0.274	28	87.931	76.7241
	PCA	0.999 ± 0.002	0.988 ± 0.018	0.96 ± 0.024	65	95.97	95.9752
	MOEA-2c	**1 ± 0.0009**	**0.997 ± 0.013**	**0.984 ± 0.034**	47	**98.452**	**98.1424**
J48	AEBP-1c	0.87 ± 0.08	0.614 ± 0.199	0.747 ± 0.13	37	81.0345	43.9655
	AEBP-2c	0.88 ± 0.13	0.577 ± 0.316	0.551 ± 0.341	52	71.1656	63.8037
	AEEA-1c	0.971 ± 0.027	0.889 ± 0.077	0.92 ± 0.057	38	91.95	85.4489
	AEEA-2c	0.82 ± 0.17	0.443 ± 0.279	0.462 ± 0.314	14	62.5767	46.0123
	AERBM-1c	0.974 ± 0.029	0.903 ± 0.077	0.92 ± 0.08	42	92.26	89.7833
	SdAE-2c	0.949 ± 0.065	0.788 ± 0.175	0.852 ± 0.131	49	87.931	76.7241
	PCA	0.979 ± 0.027	0.929 ± 0.059	0.953 ± 0.037	23	95.356	94.7368
	MOEA-2c	**0.991 ± 0.017**	**0.956 ± 0.053**	**0.971 ± 0.038**	61	**97.214**	**97.2136**
Naive Bayes	AEBP-1c	0.97 ± 0.04	0.872 ± 0.14	0.86 ± 0.096	37	88.7931	35.7759
	AEBP-2c	0.93 ± 0.06	0.565 ± 0.351	0.503 ± 0.305	47	67.4847	63.1902
	$AEEA-1c$	0.94 ± 0.077	0.697 ± 0.241	0.659 ± 0.24	64	65.015	50.774
	AEEA-2c	0.85 ± 0.1	0.425 ± 0.219	0.381 ± 0.257	60	48.4663	45.8344
	AERBM-1c	0.923 ± 0.079	0.633 ± 0.22	0.607 ± 0.201	48	59.752	55.7276
	SdAE-2c	0.962 ± 0.061	0.705 ± 0.273	0.657 ± 0.273	24	77.5862	62.069
	PCA	0.991 ± 0.016	0.952 ± 0.093	0.905 ± 0.13	53	90.093	87.9257
	MOEA-2c	**0.999 ± 0.003**	**0.994 ± 0.02**	**0.985 ± 0.023**	36	**98.452**	**97.2136**
Softmax	AEBP-1c	0.784 ± 0.132	0.303 ± 0.224	0.274 ± 0.344	64	51.7241	42.2414
	AEBP-2c	0.928 ± 0.047	0.534 ± 0.275	0.566 ± 0.27	51	70.5521	63.8037
	AEEA-1c	0.948 ± 0.032	0.653 ± 0.212	0.616 ± 0.249	57	73.9938	57.8947
	AEEA-2c	0.832 ± 0.101	0.329 ± 0.249	0.248 ± 0.317	16	51.5337	39.2638
	AERBM-1c	0.894 ± 0.084	0.577 ± .0.219	0.539 ± 0.242	33	65.3251	59.1331
	SdAE-2c	0.991 ± 0.008	0.846 ± 0.203	0.712 ± 0.365	51	88.7931	72.4138
	PCA	0.978 ± 0.044	0.916 ± 0.076	0.901 ± 0.077	34	93.1034	90.0862
	MOEA-2c	**0.998 ± 0.015**	**0.974 ± 0.042**	**0.973 ± 0.033**	33	**98.1424**	**97.8328**

higher accuracy over traditional AEs by sequentially minimizing first the reconstruction error followed by classification error. However, our MOAE-2c achieved superior results to the SdAE-2c by optimizing both reconstruction and classification error for for the best Pareto-solutions. We consistently see that our MOAE-2c produces more accurate overall results than the other tested methods. Although some classifiers (e.g. J48) only demonstrated modest accuracy improvements using MOAE-2c when compared to PCA, overall we notice consistent improvements in accuracy for all tested classifiers using our proposed MOAE-2c. The high accuracies obtained by other methods (e.g. PCA) show that our model did not suffer from over-fitting. The proposed MOAE-2c, for all classifiers, demonstrated superior performance over other methods. The accuracy rates of, 98.45% with 47 features, 96.90% with 21 features, 97.21% with 61 features, 98.45% with 36 features, and 98.14% with 33 features were obtained by the Bayes net, KNN, J48, Naive Bayes, and Softmax classifiers respectively. We note that incorporating multiple objectives via scalarization may not always perform better than a single objective as seen in the case of AEEA-1c outperforming AEEA-2c. The problem with AEEA-2c (and even AEBP-2c), which uses scalarization, is the need to decide on the weight values of the scalarization function (equation 3). A better choice of weight for AEEA-2c may have resulted in a superior performance over AEEA-1c at the expense of having to optimizing this scalarization weight. The proposed method (MOAE-2c), however, outperforms other types of single- and multi-objective optimizations of AEs: scalarization and sequential, which is accomplished by optimizing for Pareto-optimal solutions (MOAE-2c) does not require a balancing weight between MCE and MRE.

Classification Performance II. We note from Table 5 That by reducing the number of features, we managed to enhance the mean classification accuracy significantly. Notably, our proposed method reduces the dimensionality from 70 to 21 features and achieves an accuracy 8.05% greater than the second best value using PCA with a KNN classifier (96.29% vs. 88.85%). In transformation without reduction (last column), the improvement over the second best is 13.93%. This demonstrates the success of our method even without feature reduction. Our other metrics show consistent improvements as well. When using Bayes net, the proposed method achieved AUROC of 1 and similar results are obtained for Naive Bayes and J48 classifiers.

Accuracy comparison with related works reported in literature. Examining Table 6, we note that only Deserno et al. [13] tackled the problem of classifying 12 classes (like we do). Other works [7–11] gave excellent result $> 90\%$ however all of those classified less than or equal to 4 classes. Verma et al. [11] produced the highest accuracy 97%, but they used only 200 images and 2 classes. We also highlight that all these methods were tested on a single data set, but the proposed method was tested using two different data-sets.

7. Conclusion

A novel feature transformation and reduction method was presented to tackle mammography classification. Our experiments with our proposed multi-objective auto-encoder (MOAE-2c) suggests that training a deep network to minimize both classification (MCE) and reconstruction (MRE) error using Pareto-optimal solutions leads to more discriminant features. We observed improvements classification performance for our 12 class tissue and lesion problem when compared to other state-of-the-art methods. In the current paper, we demonstrated that the multi-objective optimization extension of the traditional autoencoders, which seeks Pareto-optimal solutions via NSGA-II method, results in more discriminative features than two widely used methods: PCA and stacked denoising auto-encoders. An advantage of NSGA-II over other evolutionary multi-objective optimization methods is its elitism feature, which does not allow an already found Pareto-optimal solution to be removed from the selection pool. However, the NSGA-II suffers from slow convergence [56]. The convergence speed decreases as the number of the hidden layers and their nodes in autoencoders increase. This is particularly important to address in future work, e.g. via leveraging GPU implementation, as we plan to test deeper networks with more complex architectures on raw pixel values. Future work will explore finding optimum parameters for the proposed MOAE-2c and applying it to other biomedical imaging applications.

Acknowledgements

The authors would like to thank the Breast Research Group, INESC Porto, Portugal and Prof. TM Deserno, Dept. of Medical Informatics, RWTH Aachen, Germany for providing the two useful data-sets of INbreast and IRMA version of DDSM respectively.

References

[1] A. S. Majid, E. S. de Paredes, R. D. Doherty, N. R. Sharma, X. Salvador, Missed breast carcinoma: Pitfalls and pearls 1, Radiographics 23 (4) (2003) 881–895.

[2] J. de Nazaré Silva, A. O. de Carvalho Filho, A. C. Silva, A. C. de Paiva, M. Gattass, Automatic detection of masses in mammograms using quality threshold clustering, correlogram function, and svm, Journal of digital imaging (2014) 1–15.

[3] T. Ayer, M. U. Ayvaci, Z. X. Liu, O. Alagoz, E. S. Burnside, Computer-aided diagnostic models in breast cancer screening, Imaging in medicine 2 (3) (2010) 313–323.

[4] C. Balleyguier, S. Ayadi, K. Van Nguyen, D. Vanel, C. Dromain, R. Sigal, Birads classification in mammography, European journal of radiology 61 (2) (2007) 192–194.

[5] F. Abu-Amara, I. Abdel-Qader, Hybrid mammogram classification using rough set and fuzzy classifier, Journal of Biomedical Imaging 2009 (2009) 17.

[6] K. Ganesan, U. Acharya, C. K. Chua, L. C. Min, K. T. Abraham, K. B. Ng, Computer-aided breast cancer detection using mammograms: a review, Biomedical Engineering, IEEE Reviews in 6 (2013) 77–98.

[7] A. K. Mohanty, M. R. Senapati, S. K. Lenka, An improved data mining technique for classification and detection of breast cancer from mammograms, Neural Computing and Applications 22 (1) (2013) 303–310.

[8] A. Bria, N. Karssemeijer, F. Tortorella, Learning from unbalanced data: A cascade-based approach for detecting clustered microcalcifications, Medical image analysis 18 (2) (2014) 241–252.

[9] A. Oliver, X. Lladó, J. Freixenet, R. Martí, E. Pérez, J. Pont, R. Zwiggelaar, Influence of using manual or automatic breast density information in a mass detection cad system, Academic radiology 17 (7) (2010) 877–883.

[10] T. Subashini, V. Ramalingam, S. Palanivel, Automated assessment of breast tissue density in digital mammograms, Computer Vision and Image Understanding 114 (1) (2010) 33–43.

[11] B. Verma, P. McLeod, A. Klevansky, Classification of benign and malignant patterns in digital mammograms for the diagnosis of breast cancer, Expert systems with applications 37 (4) (2010) 3344–3351.

[12] A. Sadaf, P. Crystal, A. Scaranelo, T. Helbich, Performance of computer-aided detection applied to full-field digital mammography in detection of breast cancers, European journal of radiology 77 (3) (2011) 457–461.

[13] T. M. Deserno, M. Soiron, J. E. E. de Oliveira, A. de A Araujo, Towards computer-aided diagnostics of screening mammography using content-based image retrieval, in: Graphics, Patterns and Images (Sibgrapi), 2011 24th SIBGRAPI Conference on, IEEE, 2011, pp. 211–219.

[14] V. Vapnik, The nature of statistical learning theory, Springer Science & Business Media, 2013.

[15] I. Jolliffe, Principal component analysis, Wiley Online Library, 2002.

[16] R. A. Fisher, The use of multiple measurements in taxonomic problems, Annals of eugenics 7 (2) (1936) 179–188.

[17] R. Batuwita, V. Palade, Class imbalance learning methods for support vector machines, Imbalanced learning: Foundations, algorithms, and applications (2013) 83–99.

[18] G. E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural computation 18 (7) (2006) 1527–1554.

[19] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, et al., Greedy layer-wise training of deep networks, Advances in neural information processing systems 19 (2007) 153.

[20] C. Poultney, S. Chopra, Y. L. Cun, et al., Efficient learning of sparse representations with an energy-based model, in: Advances in neural information processing systems, 2006, pp. 1137–1144.

[21] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, S. Bengio, Why does unsupervised pre-training help deep learning?, The Journal of Machine Learning Research 11 (2010) 625–660.

[22] J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, Q. V. Le, A. Y. Ng, On optimization methods for deep learning, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 265–272.

[23] Y. Bengio, Learning deep architectures for ai, Foundations and trends® in Machine Learning 2 (1) (2009) 1–127.

[24] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, Pattern Analysis and Machine Intelligence, IEEE Transactions on 35 (8) (2013) 1798–1828.

[25] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Networks 61 (2015) 85–117.

[26] M. Ranzato, M. Szummer, Semi-supervised learning of compact document representations with deep networks, in: Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 792–799.

[27] Y. Bengio, Y. LeCun, et al., Scaling learning algorithms towards ai, Large-scale kernel machines 34 (5).

[28] P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, in: Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 1096–1103.

[29] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, The Journal of Machine Learning Research 11 (2010) 3371–3408.

[30] S. Rifai, P. Vincent, X. Muller, X. Glorot, Y. Bengio, Contractive auto-encoders: Explicit invariance during feature extraction, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 833–840.

[31] S. Rifai, G. Mesnil, P. Vincent, X. Muller, Y. Bengio, Y. Dauphin, X. Glorot, Higher order contractive auto-encoder, in: Machine Learning and Knowledge Discovery in Databases, Springer, 2011, pp. 645–660.

[32] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, C. D. Manning, Semi-supervised recursive autoencoders for predicting sentiment distributions, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 151–161.

[33] H. Almousli, P. Vincent, Semi supervised autoencoders: Better focusing model capacity during feature extraction, in: Neural Information Processing, Springer, 2013, pp. 328–335.

[34] N. Andrei, A scaled bfgs preconditioned conjugate gradient algorithm for unconstrained optimization, Applied Mathematics Letters 20 (6) (2007) 645–650.

[35] Y. Yang, G. Shu, M. Shah, Semi-supervised learning of feature hierarchies for object detection in a video, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, 2013, pp. 1650–1657.

[36] D.-H. Lee, Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks, in: Workshop on Challenges in Representation Learning, ICML, Vol. 3, 2013.

[37] J. Weston, F. Ratle, H. Mobahi, R. Collobert, Deep learning via semi-supervised embedding, in: Neural Networks: Tricks of the Trade, Springer, 2012, pp. 639–655.

[38] D. E. Goldberg, J. H. Holland, Genetic algorithms and machine learning, Machine learning 3 (2) (1988) 95–99.

[39] R. T. Marler, J. S. Arora, Survey of multi-objective optimization methods for engineering, Structural and multidisciplinary optimization 26 (6) (2004) 369–395.

[40] K. Van Moffaert, T. Brys, A. Chandra, L. Esterle, P. R. Lewis, A. Nowé, A novel adaptive weight selection algorithm for multi-objective multi-agent reinforcement learning, in: Neural Networks (IJCNN), 2014 International Joint Conference on, IEEE, 2014, pp. 2306–2314.

[41] M. Caramia, P. Dell´ Olmo, Multi-objective optimization, Multi-objective Management in Freight Logistics: Increasing Capacity, Service Level and Safety with Optimization Algorithms (2008) 11–36.

[42] M. Chen, Z. Xu, K. Weinberger, F. Sha, Marginalized denoising autoencoders for domain adaptation, arXiv preprint arXiv:1206.4683.

[43] B. L. Betechuoh, T. Marwala, T. Tettey, Autoencoder networks for hiv classification, CURRENT SCIENCE-BANGALORE- 91 (11) (2006) 1467.

[44] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: Nsga-ii, Evolutionary Computation, IEEE Transactions on 6 (2) (2002) 182–197.

[45] N. H. Eltonsy, G. D. Tourassi, A. S. Elmaghraby, A concentric morphology model for the detection of masses in mammography, Medical Imaging, IEEE Transactions on 26 (6) (2007) 880–889.

[46] M. Elter, E. Haßlmeyer, A knowledge-based approach to the cadx of mammographic masses, in: Medical imaging, International Society for Optics and Photonics, 2008, pp. 69150L–69150L.

[47] A. Tagliafico, G. Tagliafico, S. Tosto, F. Chiesa, C. Martinoli, L. E. Derchi, M. Calabrese, Mammographic density estimation: comparison among birads categories, a semi-automated software and a fully automated one, The Breast 18 (1) (2009) 35–40.

[48] W. Ho, P. Lam, Clinical performance of computer-assisted detection (cad) system in detecting carcinoma in breasts of different densities, Clinical Radiology 58 (2) (2003) 133–136.

[49] S. Obenauer, C. Sohns, C. Werner, E. Grabbe, Impact of breast density on computer-aided detection in full-field digital mammography, Journal of digital imaging 19 (3) (2006) 258–263.

[50] H. He, E. Garcia, et al., Learning from imbalanced data, Knowledge and Data Engineering, IEEE Transactions on 21 (9) (2009) 1263–1284.

[51] G. Ou, Y. L. Murphey, Multi-class pattern classification using neural networks, Pattern Recognition 40 (1) (2007) 4–18.

[52] S. Maldonado, R. Weber, F. Famili, Feature selection for high-dimensional class-imbalanced data sets using support vector machines, Information Sciences 286 (2014) 228–246.

[53] M. Wasikowski, X.-w. Chen, Combating the small sample class imbalance problem using feature selection, Knowledge and Data Engineering, IEEE Transactions on 22 (10) (2010) 1388–1400.

[54] A. F. Costa, G. Humpire-Mamani, A. J. M. Traina, An efficient algorithm for fractal analysis of textures, in: Graphics, Patterns and Images (SIBGRAPI), 2012 25th SIBGRAPI Conference on, IEEE, 2012, pp. 39–46.

[55] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, J. S. Cardoso, Inbreast: toward a full-field digital mammographic database, Academic radiology 19 (2) (2012) 236–248.

[56] K. Deb, A. R. Reddy, Reliable classification of two-class cancer data using evolutionary algorithms, BioSystems 72 (1) (2003) 111–129.