A Hybrid Dimensionality Reduction Method based on Support Vector Machine and Independent Component Analysis

Sangwoo Moon and Hairong Qi

***Abstract*—This paper presents a new hybrid dimensionality reduction method to seek projection through optimization of both structural risk (supervised criterion) and data independence (unsupervised criterion). Classification accuracy is used as a metric to evaluate the performance of the method. By minimizing the structural risk, projection originated from the decision boundaries directly improves classification performance from a *supervised* perspective. From an *unsupervised* perspective, projection can also be obtained based on maximum independence among features (or attributes) in data to indirectly achieve better classification accuracy over more intrinsic representation of the data. Orthogonality interrelates the two sets of projections such that minimum redundancy exists between the projections, leading to more effective dimensionality reduction. Experimental results show that the proposed hybrid dimensionality reduction method that satisfies both criteria simultaneously provides higher classification performance, especially for noisy data sets, in relatively lower dimensional space than various existing methods.**

***Index Terms*—Hybrid dimensionality reduction, structural risk minimization, independence maximization, projection.**

I. Introduction

Due to the increasing demand for high dimensional data analysis from various applications such as electrocardiogram (ECG) signal analysis, gene expression analysis for cancer detection/DNA forensic, and content-based image retrieval (CBIR), dimensionality reduction becomes a viable process to provide robust data representation in relatively low dimensional space. Dimensionality reduction is a process to extract essential information from data such that the high-dimensional data can be represented in a more condensed form with much lower dimensionality to both improve classification accuracy and reduce computational complexity. Conventional dimensionality reduction methods can be categorized into *stand-alone* and *hybrid* approaches. The stand-alone method utilizes a single criterion from either supervised or unsupervised perspective, where supervised approaches require the prior knowledge of class assignment for training data whereas the unsupervised methods are free from this requirement. On the other hand, the hybrid method integrates both criteria. Compared with a variety of stand-alone dimensionality reduction methods, the hybrid approach is promising as it takes advantage of both the supervised criterion that results in mapping vectors aimed for better classification accuracy and the unsupervised criterion yielding mapping vectors that better represent the original data, simultaneously. However, two issues always exist that challenge the efficiency of the hybrid approach, including (1) the difficulty in finding a subspace that seamlessly integrates both criteria in a single hybrid framework, and (2) the robustness of the performance (or the generalization capability of the algorithm) regarding noisy data. Existing hybrid approaches usually combine standalone methods of Linear Discriminant Analysis (LDA) [1], Principal Component Analysis (PCA) [1], Independent Component Analysis (ICA) [2], and their variations [3].

A. Supervised Methods

Linear Discriminant Analysis (LDA) [1] is a representative supervised dimensionality reduction method. The projection in traditional LDA [4] is obtained by maximizing the variance between classes while minimizing the variance within class so as to achieve better separability in reduced dimensional space. LDA is extended to kernel Discriminant Analysis (kDA) [5] for nonlinear data representation. Inherited from the LDA criterion are the major issues of the small sample size (S3), the common mean (CM), and the robustness problem.

The small sample size often makes the within-class variance singular, so that the LDA criterion becomes infinite regardless of the between-class variance. Face recognition, for example, is a well-known application suffering from the S3 problem due to the limited number of face samples per person. Several approaches have been introduced to overcome the S3 problem such as Shrunken Centroids Regularized Discriminant Analysis (SCRDA) [6], LDA with Generalized Singular Value Decomposition [7], Null space LDA [8], Discriminative Common Vector (DCV) [9], kernel Discriminative Common Vector (kDCV) [10], Orthogonal Centroid Method (OCM) [11], Weighted Piecewise LDA [12], and LDA over PCA [13, 14]. Compared with these LDA-based criteria using the within-class variance, support vector machine (SVM) minimizes the empirical error by maximizing the separation margin measured by the distance from the separation hyperplane to the support vectors, nearest samples of any class. The separation margin is also regularized by additional constraint based on the nature of the data to build robust model against noise. The maximum separation margin with the regularization constraint leads SVM to search for the optimal trade-off between empirical error and complexity such that the decision hyperplane in SVM delivers better generalization capability for arbitrary

S. Moon and H. Qi are with the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN, 37996 USA e-mail: {smoon3,hqi}@utk.edu.

input, resulting in robustness under noisy environment. The lack of the training samples tends to increase model estimation error since the given data does not represent the true model sufficiently. The model estimation error mostly results from the samples with relatively low probability of occurrence and the problematic samples might act as noisy samples during model estimation. In SVM, the lack of the sample data per class does not degrade the classification performance as significantly as in LDA due to the generalization of decision for arbitrary data with noise regularization.

The common mean problem is caused by non-distinguishable between-class variances from overlapped centers among different classes. As a solution, Hsieh proposed Common Mean Feature Extraction (CMFE) [15], Discriminant Analysis Feature Extraction (DACM) [15], and CMFE with Approximate Pairwise Accuracy Criterion (aPAC [16]) [17]. SVM is not influenced by the common mean problem since structural risk in SVM does not rely on the training data center.

Robustness improvement is pursued as the other critical issue in LDA for better classification performance in noisy environment. Several methods have been proposed under the LDA framework, including Asymmetric Discriminant Analysis (ADA) [3] and LDA over significant nodes [18]. Due to SVM's complexity suppression in addition to minimum empirical error, the projection vectors from SVMs deliver data representation with improved robustness compared with LDA. The robustness is enhanced especially under biased and noisy environment. According to [19], LDA can only obtain a decision boundary identical to the one from SVM when there exist sufficiently large number of observations for effective representation of the internal structure of data.

Beyond the LDA criteria, SVM-related approaches like Recursive SVM (RSVM) [20] and Large-scale Maximum Margin Discriminant Analysis (Large-scale MMDA) [21] have been applied for dimensionality reduction purpose. Both are based on a series of SVMs with orthogonality. RSVM is motivated by Recursive LDA (RLDA) [22] but utilizes SVM instead of LDA to iteratively extract the projection vector. Large-scale MMDA extracts projection with maximum separability by Core Vector Machine (CVM) which provides an approximation of SVM pursuing fast computation in large-scale dataset. Although both RSVM and large-scale MMDA utilize SVM to obtain projections resulting in no struggle with the S3 or the common mean problems under improved robustness, there exists possible redundancy issue due to the lack of analysis of the similarity among the extracted projections from the multiple series of SVMs/CVMs under orthogonal relationship.

Regression is another type of dimensionality reduction approach which finds reduced dimensional space for the input variables maximally correlated with the response variables. Regression based approaches can be categorized as supervised when the response is actually the class assignment for training data represented by the input variables. The regression model for supervised dimensionality reduction includes Partial Least Squares regression (PLS regression) [23] and kernel Partial Least Squares regression (kPLS) [24]. The aim of PLS is to find linear relationship between the explanatory input and the corresponding response using the regression model, and achieves it by projecting the data onto reduced dimensional space consisting of latent variables based on the covariance structure analysis. However, the covariance-based analysis might lead to lower classification performance compared with stronger statistical measure of independent relationship among variables in ICA. kPLS extends the correlation measurement in covariance structure by using kernel function for nonlinear data representation capability to reduced dimensional space.

B. *Unsupervised Methods*

Principal Component Analysis (PCA) [1], Independent Component Analysis (ICA) [2], regression model [23], and intrinsic data geometry preservation [25–28] are well-known unsupervised dimensionality reduction approaches to provide better data representation capability and robustness. PCA seeks a projection which maximally uncorrelates data in a least-squares sense. Kernel PCA (kPCA) [29] extends PCA to nonlinear case using kernel function. The unmixing matrix in ICA acts as a projection that maximizes the independence among features based on the independence measure such as mutual information. Unsupervised regression models the dependent relationship between two sets of variables on the unknown parameters estimated from the data without prior knowledge of class assignment such that some intrinsic nature of data can be revealed which implicitly leads to classification performance improvement. The intrinsic data geometry preservation captures the mixture of the intrinsic geometric models from the original observation and represents the captured geometry into lower dimensional space.

Since data independence represents a stronger optimization criterion than uncorrelatedness, the maximum independence in ICA provides more intrinsic information resulting generally in contributing more to performance improvement with robustness than PCA. The major issue with ICA, however, is the high computational complexity associated with the data independence measure. FastICA [2] provides fast independence measure based on non-Gaussianity whereas kernel Canonical Correlation Analysis (kCCA) and kernel Generalized Variance (kGV) [30] formulate the canonical correlation in RKHS for data independence measure. Although kCCA/kGV provides better independence representation than FastICA, the relatively high computational complexity from the nonlinear function search in RKHS makes kCCA/kGV impractical compared with FastICA. Regression approaches can be categorized as unsupervised dimensionality reduction method when the response is not required to be identical to the class assignment. Unsupervised Kernel Regression (UKR) [31] is a representative unsupervised regression method through latent variables in the nonlinear hyperplane. Although UKR estimates maximum correlation via kernel to construct lower dimensional latent space, it still optimizes correlation measure which is statistically less powerful than the independent relationship in ICA. [32] shows the effectiveness of Wiener filter not only as a noise reduction method prior to PCA/ICA but also as a stand-alone dimensionality reduction method under noisy environment. ISOMAP [25], Locally Linear Embedding (LLE) [26], Generalized PCA [27], and Self Organizing

Maps with distance preservation in Output Space [28] aim at preserving intrinsic geometry of data. Based on the kernel interpretation for manifold methods [33], ISOMAP is extended to kernel ISOMAP [34] with projection property providing projection of arbitrary data onto manifold. Although the geometry preservation approaches offer effective data visualization capability, they require prior knowledge of mixed geometric models in the observation, which is mostly unknown.

C. Hybrid Methods

The hybrid dimensionality reduction consists of both supervised and unsupervised criteria so as to provide a single framework to find better data representation for classification performance improvement. There exist several hybrid approaches such as Asymmetric Principal and Discriminant Analysis (APCDA) [3], LDA over PCA, ICA augmented by LDA [35], Discriminant Nonnegative Matrix Factorization (DNMF) [36], Nonnegative Tensor Factorization (NTF) with LDA [37], supervised Mutual Information (MI)-based ICA [38], and Kernel Dimensionality Reduction (KDR) [39].

Although these hybrid approaches improve/resolve various problems associated with LDA to certain degree, there are still performance limitations. For example, among the three issues that concern the LDA algorithm (i.e., S3, common mean, and robustness), APCDA only alleviates the common mean problem and improves robustness; LDA over PCA only eliminates the S3 problem; the sequential combination of PCA, ICA, and LDA in ICA augmented by LDA improves overall classification performance but still suffer from the S3 and common mean problems; DNMF and NTF with LDA also suffer from the S3 and common mean problems;the supervised MI-based ICA does not sufficiently incorporate the nature of the data behavior along the decision hyperplane as a hybrid approach since the dimensionality reduction only relies on independence measured by mutual information between classes as well as features; and KDR extends PLS/kPLS's correlation analysis to canonical correlation analysis (CCA) in the Reproducing Kernel Hilbert Space (RKHS) to provide better statistical relationship of conditional independence between the input and the response variables. However, KDR does not provide robust data representation as shown in SVM due to the lack of generalization capability.

In this paper, we present an effective hybrid dimensionality reduction method based on Support Vector Machine (SVM) and Independent Component Analysis (ICA), referred to as SVM+ICA, to maintain high classification accuracy in lower dimensional space that is less sensitive to noise. Since SVM+ICA is not based on LDA, it does not suffer from the S3 or common mean problems inherited from the LDA criteria.

SVM maximizes separation margin between classes so as to offer projection with better generalization capability to improve classification/estimation performance for unknown samples. Since maximum margin among features provides better data representation to improve classification performance [40] and SVM projection itself is capable of building an effective subspace for dimensionality reduction [20, 21], we adopt SVM as a supervised component in the proposed hybrid algorithm. On the other hand, ICA offers projection which maximizes independence among features with better data representation [2] and has been shown [41, 42] to play an important role in classification performance improvement. Therefore, we incorporate ICA as the unsupervised component in the proposed hybrid algorithm. In order to combine projections derived from SVM and ICA into a unified framework for effective dimensionality reduction, the orthogonal relationship is sought between mapping vectors from SVM and ICA, such that contribution made by the supervised and unsupervised processes have minimum correlation, leading to much reduced dimensionality. This idea is similar to the Orthogonal Centroid Method (OCM) [43] but we replace OCM's maximum margin criterion with SVM's structural risk minimization which does not suffer from the S3 problem. Under the orthogonal relationship between SVM and ICA, ICA over the subspace orthogonal to SVM projection vectors allows us to merge two projections from both SVM and ICA into one concatenated projection matrix. Therefore, the proposed hybrid dimensionality reduction method improves classification performance with robustness resulting from minimum structural risk with independence.

The paper is organized as follows: Sec. II describes the new hybrid dimensionality reduction method. In Sec. III, the hybrid framework is extended to multiclass case. Experimental results and performance evaluation are shown in Sec. IV. The paper is concluded in Sec. V.

II. SVM+ICA

In this section, we describe the new hybrid dimensionality reduction method that consists of the simultaneous minimization of structural risk (the supervised criterion) and maximization of data independence (the unsupervised criterion), as each criterion has shown better performance individually compared to the corresponding traditional criterion, such as LDA or PCA. We refer to this method as SVM+ICA.

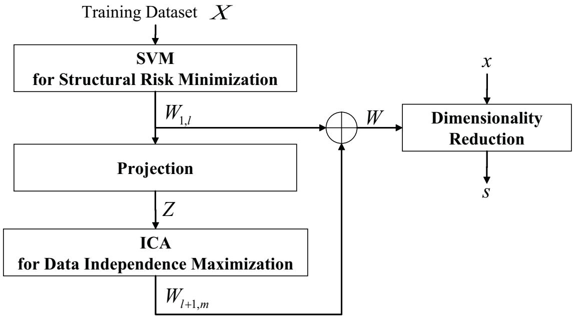


Fig. 1. The SVM+ICA framework for hybrid dimensionality reduction.

Figure 1 provides a block diagram of the proposed SVM+ICA method. It consists of three components, structural risk minimization, projection, and independence maximization. In Fig. 1, $X = \{\boldsymbol{x}_i \in R^n, \forall i\}$ represents a training data set of dimension n, which is to be reduced to another set, S, of dimension, m, where $m \ll n$, using the projection matrix, W, of m mapping column vectors constructed from the SVM+ICA process. The structural risk minimization component generates the first l mapping vectors of W, denoted as $W_{1,l}$ and the data independence maximization component yields the other $m - l$

vectors of W, denoted as $W_{l+1,m}$. This concatenation process is denoted using the symbol $\bigoplus$ in Fig. 1. Z is the projected data set from X based on $W_{1,l}$, which is to be fed to the data independent maximization component to derive $W_{l+1,m}$. We will elaborate on the rationale behind the proposed SVM+ICA framework in the following three subsections. Based on the projection matrix, W, the dimensionality reduction process can be carried out as follows,

$$\boldsymbol{s} = W^{\mathrm{T}}\boldsymbol{x} \tag{1}$$

where $\boldsymbol{s} \in R^m$ is the data samples with reduced dimension.

A. Support Vector Machine for Structural Risk Minimization

The structural risk minimization step generates $W_{1,l}$ whose column vectors represent the direction of the decision surface in classification. The concept of using the decision surface as mapping vectors for supervised dimensionality reduction is not new. LDA is the first-of-a-kind that adopts this idea. The only difference between LDA and SVM is the different objective functions they utilize to obtain the decision surface.

SVM [44] provides the decision surface satisfying minimum structural risk by maximizing the separation margin through constrained quadratic problem with duality for binary classification problems.

$$\begin{aligned} \boldsymbol{\alpha}^* &= \underset{\boldsymbol{\alpha}}{\text{argmin}} \left\{ \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}} Q \boldsymbol{\alpha} - \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{1} \right\} \\ &\text{subject to } \sum_{i=1}^{N} \alpha_i y_i = 0,\ 0 \le \alpha_i \le C \end{aligned} \tag{2}$$

where $\boldsymbol{\alpha} = [\alpha_1 \cdots \alpha_N]^{\mathrm{T}}$ and α_i, $i = 1, \cdots, N$, denotes the Lagrange multiplier corresponding to the i-th data pair $(\boldsymbol{x}_i,\ y_i)$ with $\boldsymbol{x}_i$ being the data sample and $y_i \in \{-1, 1\}$ the class index for a two-class separation problem. N is the number of samples in the training set. $Q = [q_{ij}]$ is an $N \times N$ matrix where $q_{ij} = y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$, $i, j \in \{1, \cdots, N\}$. $\mathbf{1}$ represents the column vector consisting of N-many 1's. C is the relaxation parameter which allows SVM to tolerate certain level of empirical error in decision margin during training so as to generalize the decision boundary for arbitrary input. The optimal decision surface is formed by

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = 0 \tag{3}$$

where

$$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i^* y_i \boldsymbol{x}_i \in R^n. \tag{4}$$

and $b \in R^1$ has the functionality to shift $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} = 0$ parallel to the location where the two classes denoted by $y_i \in \{-1, 1\}$ are classified with minimum structural risk. Since $\boldsymbol{w}$ is the core information of the decision making process, we utilize $\boldsymbol{w}$ as part of the overall linear mapping, W, in the proposed SVM+ICA framework. Although b in Eq. (3) also plays an important role as for classification purpose, it is not as essential as $\boldsymbol{w}$ since it does not affect the "direction" of the decision surface. Since y_i takes on two values, the problem in Eq. (2) is defined for 2-class datasets, resulting in $W_{1,l} = \boldsymbol{w}$ where $l = 1$. When dealing with multiclass problems, more than one $\boldsymbol{w}$ will be derived from the structural risk minimization process, thus more $\boldsymbol{w}$'s will be used as part of W. The extension to multiclass dataset will be discussed in Sec. III.

Using structural risk as a dimensionality reduction criterion, we expect that mappings from structural risk minimization by SVM are more robust than LDA due to the generalization capability especially when observations in the same class are biased or corrupted with noise. We also expect that structural risk based dimensionality reduction shows equal or better classification accuracy than LDA or kDA since LDA can only obtain a decision boundary identical to the one from SVM when there exist sufficiently large number of observations for effective representation of the internal structure of data [19].

In order to show that SVM provides better model estimation than LDA with imbalanced data, we prepare an example using a two-class synthetic dataset in a two-dimensional space. The data in each class consists of mixture of two Gaussians with biased number of samples corrupted by noise of SNR=5 where SNR is the signal-to-noise ratio. The two Gaussians in class 1 have 500 and 50 samples centered at $[-2\ 3]^{\mathrm{T}}$ and $[2\ 3]^{\mathrm{T}}$ respectively whereas class 2 includes two Gaussians with 50 and 500 samples centered at $[-2\ \ -3]^{\mathrm{T}}$ and $[2\ \ -3]^{\mathrm{T}}$, respectively. The covariance for the Gaussians are all identity matrices.

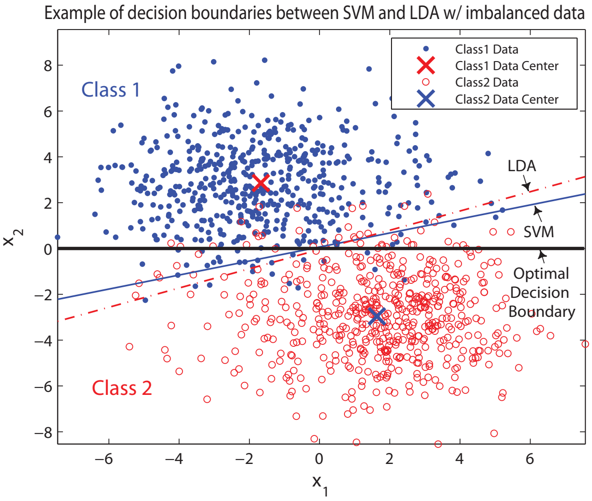


Fig. 2. Example of decision boundaries from SVM and LDA with imbalanced data

The distribution of data for the example is shown in Fig. 2, where the bullet indicates data in class 1 and the circle is for data in class 2. Due to the symmetry between class 1 and 2, the optimal decision should be made at the linear decision boundary, $\boldsymbol{x}_2 = 0$, indicated by the dark solid line in the figure, only if there exists sufficient amount of data for both class 1 and 2. We observe that the decision boundary by linear SVM is closer to the optimal than the one by LDA, which shows that SVM estimates more accurate model than LDA with imbalanced samples. This is because SVM builds decision boundary by support vectors regularized by C for effect of noise whereas LDA utilizes sample mean to form decision criteria. Consequently, the model estimation capability of LDA is determined by the accuracy of the sample mean over the true mean. The decision boundaries by both LDA and SVM should converge to the optimal when there exist sufficient amount of clean data with unbiased data distribution.

B. Projection

Intuitively, the most effective set of mapping vectors derived from the structural risk minimization process ($W_{1,l}$) and the independence maximization process ($W_{l+1,m}$) should be the ones without any redundant information for the reduced space construction spanned by $W_{1,l}$ and $W_{l+1,m}$. The least amount of redundancy results from the pair-wise orthogonality between $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ where $i \in \{1, \cdots, l\}$ and $j \in \{l+1, \cdots, m\}$. The pair-wise orthogonality is also represented by $W_{1,l} \perp W_{l+1,m}$ or equivalently $W_{l+1,m}^{\mathrm{T}} W_{1,l} = \mathbf{0}$.

The projection component, as an intermediate step in the SVM+ICA framework, allows for mapping vectors derived from structural risk minimization and independence maximization to achieve minimum correlation. It does so by projecting the given data X onto the subspace satisfying $W_{1,l}^{\mathrm{T}} \boldsymbol{x} = \mathbf{0}$, yielding the projected data, Z, such that the subsequent independence maximization process based on Z is least affected or correlated with the previous structural risk minimization process. After the projection procedure, the projected data, Z, would lose information along the direction of $W_{1,l}$, which indicates that decision information through $W_{1,l}$ is no longer valid in the projection subspace. Therefore, the projection guarantees that any mapping vectors from structural risk minimization, $W_{1,l}$, and independence maximization, $W_{l+1,m}$, are uncorrelated since $W_{l+1,m} \perp W_{1,l}$.

The projection onto the subspace, orthogonal to the decision hyperplane from structural risk minimization, $W_{1,l}$, is formulated as a constrained optimization problem as follows,

$$\begin{aligned} \boldsymbol{z}^* &= \underset{\boldsymbol{z}}{\operatorname{argmin}} \|\boldsymbol{x} - \boldsymbol{z}\|^2 \\ &\text{subject to } W_{1,l}^{\mathrm{T}} \boldsymbol{z} = \mathbf{0} \end{aligned} \tag{5}$$

where $\boldsymbol{z}$ represents the projected data onto the subspace orthogonal to $W_{1,l}$ and parallel to the decision hyperplane(s). Due to the orthogonality between $W_{1,l}$ and any components in the decision hyperplane, the structural risk minimization and independence maximization are isolated and performed one by one holding independence between any pair of $\boldsymbol{w}_i$'s and $\boldsymbol{w}_j$'s where $i \in \{1, \cdots, l\}$ and $j \in \{l+1, \cdots, m\}$.

In order to solve Eq. (5), we apply Lagrange optimization by introducing the Lagrangian multipliers, $\boldsymbol{\lambda} \in R^l$ as follows,

$$L(\boldsymbol{z}, \boldsymbol{\lambda}) = \|\boldsymbol{x} - \boldsymbol{z}\|^2 + \boldsymbol{\lambda}^{\mathrm{T}} (W_{1,l}^{\mathrm{T}} \boldsymbol{z}) \tag{6}$$

Taking the partial derivative of L with respect to $\boldsymbol{z}$ and $\boldsymbol{\lambda}$, we have

$$\partial L(\boldsymbol{z}, \boldsymbol{\lambda}) / \partial \boldsymbol{z} = -2(\boldsymbol{x} - \boldsymbol{z}^*) + W_{1,l} \boldsymbol{\lambda} = \mathbf{0} \tag{7}$$

$$\partial L / \partial \boldsymbol{\lambda} = W_{1,l}^{\mathrm{T}} \boldsymbol{z}^* = \mathbf{0} \tag{8}$$

By summarizing Eqs. (7) and (8), we have

$$\begin{bmatrix} 2I_n & W_{1,l} \\ W_{1,l}^{\mathrm{T}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}^* \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} 2\boldsymbol{x} \\ \mathbf{0} \end{bmatrix} \tag{9}$$

where I_n is the identity matrix of n dimension. The $\boldsymbol{z}^*$'s form the projected dataset Z which will be used by the subsequent independent maximization process.

C. Independence Maximization

As unsupervised dimensionality reduction component in the proposed SVM+ICA framework, independence maximization is applied over the projected data, Z. Independence maximization searches for a linear (possibly) non-orthogonal coordinate system whose axes are determined by both the second and higher order statistics of the original data. Since independence maximization is known as a method providing better data representation than other conventional techniques such as PCA, higher classification accuracy is expected, leading to the adoption of independence maximization in the proposed hybrid dimensionality reduction framework. To find mappings which maximize independence, we adopt the approximated negative entropy criterion introduced in [2], also referred to as FastICA, due to well-justified statistical theory and computational efficiency. The FastICA algorithm involves two sequential processes, the *one unit (weight vector) estimation* and the *decorrelation* among weight vectors. The one unit process estimates the weight vectors as follows,

$$\boldsymbol{w}_i^+ = E\left\{\boldsymbol{z} g(\boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{z})\right\} - E\left\{g'(\boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{z})\right\} \boldsymbol{w}_i \tag{10}$$

where $\boldsymbol{w}_i^+$ is the temporal approximation of the independent component with $i \in \{l+1, \cdots, m\}$. g is the derivative of the non-quadratic function introduced in [2], and $g(u) = \tanh(au)$. g' is the derivative of g, and $g'(u) = a\mathrm{sech}^2(au)$.

The purpose of the decorrelation process is to keep different weight vectors from converging to the same maximum. The deflation scheme based on symmetric decorrelation [45] helps remove dependency among $\boldsymbol{w}_i^+$'s as follows,

$$W_{l+1,m} = W_{l+1,m}^+ \left[({W_{l+1,m}^+}^{\mathrm{T}} W_{l+1,m}^+)^{-\frac{1}{2}} \right]^{\mathrm{T}} \tag{11}$$

where $W_{l+1,m}$ represents decorrelated mappings based on $W_{l+1,m}^+ = [\boldsymbol{w}_{l+1}^+ \cdots \boldsymbol{w}_m^+]$ from independence maximization.

III. Generalization of The SVM+ICA Framework

The proposed hybrid dimensionality reduction in Sec. II is limited for 2-class datasets only because of the usage of SVM in Eq. (2) with two-class label. In this section, we discuss the multiclass extension of SVM+ICA and issues associated with the extension. For a two-class problem, there is only one mapping vector generated from the SVM component, i.e., $l = 1$ for $W_{1,l}$. For multiclass datasets, $l > 1$, hence decorrelation among the mapping vectors need to be conducted.

A. Multiclass Generalization of SVM

There are two distinctive multiclass SVM approaches, referred to as one-against-all and one-against-one [46]. The one-against-all approach compares data in a single class with all the others to generate the decision boundary. This method builds c-many decision boundaries from c-many one-against-all data combinations, where c denotes the number of classes. The one-against-one approach creates decision boundaries from all possible pair of classes. It basically generates ${}_cC_2$-many decision boundaries where ${}_cC_2$ represents the number of combinations of c-many classes taken 2 at a time. For a 2-class pattern, one-against-one is equivalent to one-against-all.

The one-against-all provides relatively small number of projection vectors than one-against-one, resulting in lower dimensional data representation since $c \leq {}_cC_2$ for $c \geq 3$. However, the one-against-all also requires at least equal or more amount of training data per SVM compared with the one-against-one, resulting in higher computational complexity to solve the quadratic problem in Eq. (2) since the number of unknown variables, α's, increases proportionally to the number of training samples.

In this paper, we adopt the one-against-all technique to extend SVM+ICA to handle multiclass datasets, where $l = c$ in $W_{1,l}$. Independence maximization is subsequently applied to extract the other $\boldsymbol{w}_j$'s, $j \in \{l+1, \cdots, m\}$ over the projected data Z. Since there is more than one $\boldsymbol{w}_i$ generated from the SVM component, some post-processing procedures, i.e., eligibility test and decorrelation, need to be conducted.

B. *Eligibility Test*

Although there are l-many $\boldsymbol{w}_i$'s, $i \in \{1, \cdots, l\}$ from the structural risk minimization process, not all of them need to be adopted for the purpose of dimensionality reduction. The eligibility of $\boldsymbol{w}_i$ from SVM is evaluated based on the importance of the information from the classification perspective. $\boldsymbol{w}_i$'s resulting in poor classification performance do not include meaningful information of decision for the given dataset and therefore should be discarded.

The eligibility test discards $\boldsymbol{w}_i$ when $r_i \leq \gamma$ where $r_i \in [0, 1]$ is the classification accuracy corresponding to $\boldsymbol{w}_i$ in the one-against-all classification and γ is a threshold. In this paper, γ is set to 0.5 to accept $\boldsymbol{w}_i$ only when r_i is higher than the classification accuracy of 50% as 1-D random walk with equal probability for both directions (left/right) on 1-D plane [47].

C. *Asymmetric Decorrelation for Redundancy Removal*

The redundancy removal procedure is embedded in ICA through the decorrelation process. This section focuses on how to remove redundancy in $\boldsymbol{w}_i$'s, $i \in \{1, \cdots, l\}$ generated from the risk minimization process.

Although symmetric decorrelation in Eq. (11) removes redundancy among $\boldsymbol{w}_i$'s in $W_{l+1,m}$ from independence maximization, it cannot be applied to $\boldsymbol{w}_i$'s in $W_{1,l}$ from SVM since the direction of $\boldsymbol{w}_i$'s in $W_{1,l}$ provides core information of decision hyperplanes. Instead of symmetric decorrelation, in this paper, we introduce the concept of *asymmetric decorrelation* to alleviate redundancy among $\boldsymbol{w}_i$'s in $W_{1,l}$.

Asymmetric decorrelation between two vectors, $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ is conducted based on two metrics, the angular distance between the vectors, a_{ij}, and the classification performance using the vectors, r_i and r_j. r_i is obtained based on w_i over training dataset as follows,

$$r_i {=} \frac{1}{2N} \sum_{k=1}^{N} \left| y_k^{(i)} - \mathrm{sgn}\left(\boldsymbol{w}_i^{\mathrm{T}} \boldsymbol{x}_k + \boldsymbol{b}_i\right) \right| \tag{12}$$

where $y_k^{(i)} \in \{-1, 1\}$ denotes x_k's desired output for the i-th SVM. b_i represents a bias for w_i. $\mathrm{sgn}(\cdot)$ is the sign function. We formulate the following pseudo-metric to reflect the joint effect of angular distance and classification performance,

$$d_{ij} = a_{ij} \frac{r_i}{r_j} \frac{\gamma}{\pi} \tag{13}$$

where $a_{ij} = \arccos\left(\boldsymbol{w}_i{}^{\mathrm{T}} \boldsymbol{w}_j / (||\boldsymbol{w}_i||||\boldsymbol{w}_j||)\right)$ is the angular distance between $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$ which is symmetric, satisfying $a_{ij} = a_{ji} \in [0, \pi]$. r_i and r_j are the classification accuracies using $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$, respectively. $\gamma = 0.5$ and π are normalization factors such that $d_{ij} \in [0, 1]$. d_{ij} represents how close $\boldsymbol{w}_i$ is to $\boldsymbol{w}_j$. It is asymmetric due to $d_{ij} \neq d_{ji}$ with $a_{ij} = a_{ji}$ and $r_i \neq r_j$. $d_{ij} < d_{ji}$ represents the situation where $\boldsymbol{w}_i$ becomes less meaningful than $\boldsymbol{w}_j$ due to $r_i < r_j$. Since smaller d_{ij} denotes more redundancy between $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$, we utilize d_{ij} as a pseudo-metric for redundancy removal process.

Algorithm 1 Pseudocode for redundancy removal

```
Begin
Require: $\delta$, $W_{1,l} = [\boldsymbol{w}_1 \cdots \boldsymbol{w}_l]$, $r_i$, where $i \in \{1, \cdots, l\}$
  Set $I = \{1, \cdots, l\}$
  Evaluate all $d_{ij}$ for $i, j \in I$, $i \neq j$
  while $d_{uv} \leq \delta$ where $d_{uv} = \min_{i,j}(d_{ij})$ for $i, j, u, v \in I$
  AND $n(I) \geq 1$ do
    $I = I - \{u\}$
  end while
  return $\boldsymbol{w}_i$'s for $i \in I$
End
```

Algorithm 1 shows the pseudocode for the redundancy removal process for $\boldsymbol{w}_i$ from structural risk minimization. $\delta \in [0, 1]$ is a threshold for asymmetric decorrelation to guide the decision whether to discard $\boldsymbol{w}_i$. $n(I)$ denotes the number of elements in the set, I. $\boldsymbol{w}_i$ is treated as redundant and eliminated by the process *one by one* only when $d_{ij} \leq \delta$. This process stops when there exists no $d_{ij} \leq \delta$ or no $\boldsymbol{w}$ remains. After the removal process, $W_{1,l}$ includes equal ($\delta = 0$) or less ($0 < \delta \leq 1$) number of $\boldsymbol{w}_i$'s. In case of $\delta = 1$, $W_{1,l}$ becomes empty matrix so that dimensionality reduction in SVM+ICA depends only on ICA. After redundancy removal, the dimensionality from SVM becomes $l \leq c$.

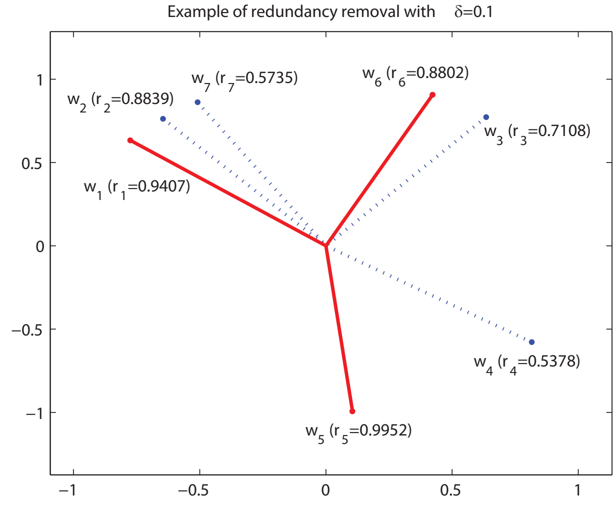


Fig. 3. Example of redundancy removal

Figure 3 shows an example for the redundancy removal process with the threshold, $\delta = 0.1$. There are 7 $\boldsymbol{w}_i$'s initially, $i \in I = \{1, \cdots, 7\}$ generated from the risk minimization component. The solid lines denote $\boldsymbol{w}_i$'s survived at the end of the redundancy removal process whereas the dotted lines indicate $\boldsymbol{w}_i$'s eliminated during the redundancy removal process. In the first iteration, the minimum asymmetric decorrelation is found between $\boldsymbol{w}_2$ and $\boldsymbol{w}_7$ with $\min(d_{ij}) = d_{72} = 0.0174$

resulting in the elimination of $\boldsymbol{w}_7$ due to $r_7 < r_2$. In the second iteration, $\boldsymbol{w}_2$ is removed based on $\min(d_{ij}) = d_{21} = 0.0275$ with $r_2 < r_1$. $\boldsymbol{w}_3$ is eliminated with $\min(d_{ij}) = d_{36} = 0.0322$ with $r_3 < r_6$ in the third iteration. $\boldsymbol{w}_4$ is the last one to be discarded with $\min(d_{ij}) = d_{45} = 0.0729$, $r_4 < r_5$ in the fourth iteration. The removal process terminates at the beginning of the fifth iteration since $\min(d_{ij}) = d_{61} = 0.1968 > \delta$ so there only survive $\boldsymbol{w}_1$, $\boldsymbol{w}_5$, and $\boldsymbol{w}_6$ which are sufficiently far away from each other with relatively higher classification accuracies.

D. Conducting Dimensionality Reduction

The dimensionality reduction by SVM+ICA is performed by linear projection,

$$\boldsymbol{s} = W^{\mathrm{T}}\boldsymbol{x} \tag{14}$$

where $W = [W_{1,l} W_{l+1,m}]$ and m is the number of dimensions to be reduced to. When $m \leq l$, the dimensionality reduction is driven only by $\boldsymbol{w}_i$'s from SVM without ICA. When $m > l$, $(m - l)$-many mapping vectors from ICA will be added to the mapping matrix, W. There is the potential of sparse representation of W to reduce computational cost of the dimensionality reduction process by utilizing fast sparse matrix-vector product computation. For example, we can extend the conjugate gradient method for fast computation of gradient-based search in SVM and ICA as well as the projection process in Eq. (9). For the sparse W, however, both SVM and ICA must be capable of building sparse projections/bases which is beyond the scope of this paper.

IV. Experimental Results and Analysis

We utilize two data sets, 'Arrhythmia' from UCI Machine Learning databases [48] and 'Cancer' from the Center for Genome Research at MIT Whitehead Institute [49] to demonstrate the effectiveness of the proposed hybrid dimensionality reduction method in terms of classification accuracy and its robustness toward noisy data input.

The Arrhythmia dataset includes 452 instances of 16 types of cardiac arrhythmia in R^{279}. We only utilize 13 types in the Arrhythmia dataset due to no instances found for the excluded 3 types. The number of samples per class is heavily biased from 2 to 245 with missing elements. We replace the missing elements with a random number generated between the minimum and maximum readings of that attribute. The cancer dataset uses 16063 tumor gene expression signatures to distinguish 14 different types of cancers. This dataset does not contain any missing attributes. The 144 training and 46 testing data are given. For noisy environment construction, we add Gaussian noise to individual dimension independently for the entire data with Signal-to-Noise Ratio (SNR) from 5[dB] to 50[dB]. The noiseless data has SNR of ∞[dB].

In addition to the Arrhythmia and Cancer data sets, the 'Arcene' dataset [48] is also utilized for in-depth analysis of the distinctive behavior of supervised and unsupervised methods over various class-wise training data balance. The Arcene dataset aims at 2-class positive/negative cancer detection and consists of 100 (44 positive/56 negative) training, 100 (44/56) validation, and 700 (310/390) test samples in R^{10000}. We utilized the validation samples for testing since the ground truths of the test samples are not unveiled.

We adopt the k-nearest neighbor (kNN) algorithm as classifier and use classification accuracy as a performance measure of dimensionality reduction method due to its non-parametric nature. For Arrhythmia dataset, the half of dataset is randomly selected for training since the minimum number of sample per class is 2. We utilize the given training data for the cancer dataset. The SVM+ICA model estimation is based on cross validation with the range of parameters: $C = \{10^{-3}, 10^{-2}, \cdots, 10^{3}\}$, $\delta = \{0, 0.005, \cdots, 0.3\}$, and $m = \{1, 2, \cdots, 100\}$. The number of neighbors, $k = \{1, 2, \cdots, 50\}$ is fine-tuned after the cross validation to be completed on C, δ, and m with $k = 10$. The selected parameters over cross validation and fine-tuned k are shown in Table I.

TABLE I
Selected Parameters for SVM+ICA

	Arrhythmia Dataset				Cancer Dataset			
SNR[dB]	C	δ	m	k	C	δ	m	k
05	10^{-3}	0.180	28	7	10^{-3}	0.2498	45	2
10	10^{-3}	0.214	13	5	10^{-1}	0.2490	40	2
20	10^{-3}	0.180	16	7	10^{-0}	0.2520	36	5
30	10^{-3}	0.200	44	6	10^{+1}	0.2530	34	5
40	10^{-3}	0.200	16	5	10^{+2}	0.2530	38	5
50	10^{-3}	0.180	16	5	10^{+3}	0.2520	38	5
∞	10^{-3}	0.180	18	9	10^{+3}	0.2530	38	4

For SVM in the proposed SVM+ICA, we tested traditional Sequential Minimal Optimization (SMO) [50] with L1-norm, SVMLight [51] with L2-norm and LibSVM [52] with L2-norm. Among the three implementations, LibSVM is utilized to build SVM+ICA model due to the best overall classification performance with relatively less sensitive behavior to parameter change. For ICA in SVM+ICA, we adopt FastICA as described in Sec. II-C.

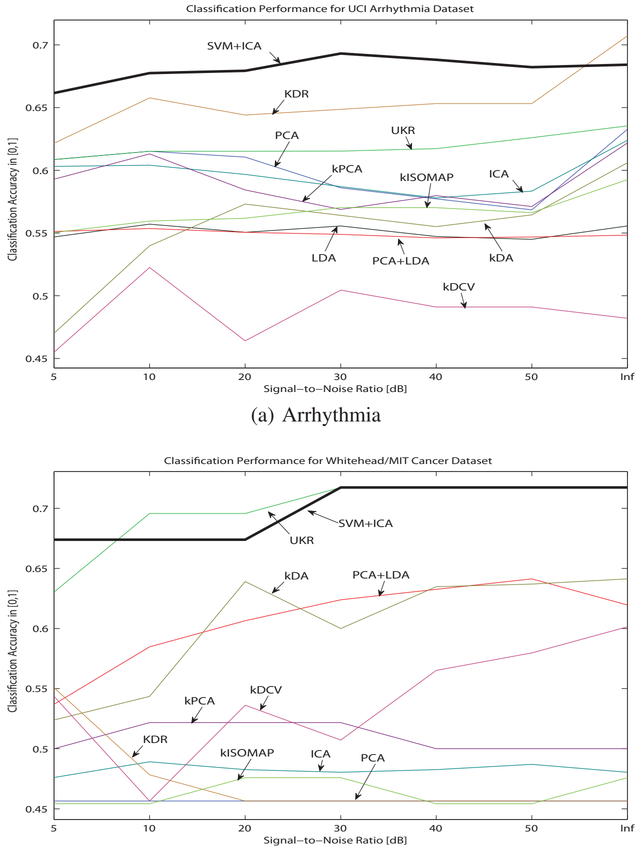
Table II shows the model estimation time spent over various algorithms. The second column of the table indicates the time spent by single parameter set, the third column is the number of parameters utilized in grid search (the numbers within the parenthesis indicate the number of possible values that each parameter can take), and the last column is the time spent for the whole grid search procedure. It is calculated by multiplying the computational time spent by single parameter setting and the number of possible combinations of all parameter values. For example, the total time spent by SVM+ICA is 15.87[s]×($100 \times 7 \times 61$)=188.2[hr].

TABLE II
Model Estimation Time with Cross Validation and Grid Search

Algorithm	Single Estimation	Parameters	Grid Search
PCA	0. 10 [s]	1 (100)	10.0 [s]
LDA	0. 25 [s]	1 (100)	25.0 [s]
PCA+LDA	0. 33 [s]	1 (100)	33.0 [s]
UKR	222. 56 [s]	0	223.0 [s]
kPCA	0. 36 [s]	2 (100×7)	252.0 [s]
kDA	0. 89 [s]	2 (100×7)	623.0 [s]
kDCV	1. 64 [s]	2 (100×7)	0.3 [hr]
ICA	6. 57 [s]	2 (100×3)	0.6 [hr]
kISOMAP	9. 06 [s]	2 (100×7)	1.8 [hr]
SVM+ICA	15. 87 [s]	3 ($100 \times 7 \times 61$)	188.2 [hr]
KDR	14. 25 [hr]	3 ($100 \times 7 \times 4$)	39900.0 [hr]

All algorithms except UKR use m as a parameter where m can take 100 values in $\{1, 2, \cdots, 100\}$. UKR does not require grid search due to its internally implemented parameter search capability. ICA uses an additional parameter, $a \in \{1, 1.5, 2\}$ for $g(u)$ in Eq. (10). In addition, since we select Gaussian kernel for all kernel-based approaches, i.e., kPCA, kDA, kDCV, kISOMAP, these algorithms have an additional parameter, kernel width, that takes 7 values in $\{10^{-3}, 10^{-2}, \cdots, 10^{3}\}$. KDR also includes the maximum number of annealing process as parameter that takes four possible values from $\{2, 4, 6, 8\}$. SVM+ICA, as described in Table I, has 3 parameters, m, C, δ for grid search. Due to the excessive computational burden of grid search, we first apply coarse search and then fine-tune the model in the given parameter ranges. Widening the current grid search resolution increases excessively the computational burden without sensible performance improvement.

A. *Comparison of Different Approaches*



(a) Arrhythmia

(b) Cancer

Fig. 4. Comparison of classification performance in noisy environment

Figure 4 denotes the classification performance over noisy environment with various noise levels. We compare the classification performance of SVM+ICA with that of PCA, kPCA, ICA, UKR, kISOMAP for unsupervised, LDA, kDA, kDCV for supervised, and PCA+LDA, KDR for hybrid approaches. For kernel-based method such as kPCA, UKR, kDA, kDCV, and KDR, we utilize Gaussian kernel with the kernel width selected in the range of $[0.01,\infty]$ to estimate model which delivers the best classification performance. Since the cancer dataset has 144 training samples in 16063-dimensional space, its covariance matrix becomes 16063×16036 which is not applicable. Instead of direct covariance calculation, we apply Eigenface [53] to reduce the computational complexity when implementing the PCA and PCA+LDA approaches. We exclude LDA for the cancer dataset due to the difficulty in eigenvalue decomposition from the 16063×16063 matrix. Additionally, kPCA and kDA do not utilize covariance matrix, but the Gram matrix of 144×144.

We make two observations from Fig. 4. First of all, no matter what the SNR level is, the proposed SVM+ICA presents, in most cases, the highest classification accuracy, demonstrating its supremacy over tested supervised, unsupervised or hybrid approaches. This remains true for both the balanced and imbalanced datasets. Second, for imbalanced datasets like the Arrhythmia, the unsupervised methods such as PCA, kPCA, ICA, UKR, and kISOMAP work better than the supervised methods such as LDA, kDA, kDCV. And for the cancer dataset with relatively well-balanced data distribution, the supervised methods generally outperform the unsupervised methods. UKR shows competitive result due to the regression which is known as a superior model selection scheme with small number of training samples in high dimensional space although it is an unsupervised method. For KDR, the performance is degraded for Cancer dataset since the gradient search in KDR suffers from the local minimum problem to become severe by imbalanced dataset. In both cases, the proposed SVM+ICA, with its seamless integration of the supervised SVM and unsupervised ICA, presents the best overall performance. The non-monotone behavior of the performance is observed mainly from the algorithms with poor performance due to less accurate model estimation resulting in the inconsistent performance trend as shown in Fig. 4. Another presentation of the overall summary of Fig. 4 is provided in Table III with the classification accuracy and the corresponding reduced dimensionality.

Most of the computational burden in the proposed algorithm results from SVM. The succeeding projection and ICA processes require relatively lower computational cost than SVM due to the complexity of quadratic problem solver. For example, the processes of SVM, projection, and ICA take 5.1, 0.2, and 3.5 seconds on the Arrhythmia dataset and 175.9, 21.2, and 12.0 seconds on the Cancer dataset, respectively for the proposed SVM+ICA model construction in Table III. The examples show first that SVM always consumes more computational time than the others. Secondly, projection becomes more complex than ICA when training data dimension increases due to the quadratic growth of the matrix size in Eq. (9) whereas the computational complexity in ICA relies on the iterative gradient search for each one of the components to be extracted and ICA in SVM+ICA model does not require huge number of bases. The grid parameter search, which iterates multiple times of training process on parameter grid, is utilized for SVM+ICA to find the best parameter set.

Figure 5 shows the relationship between class-wise data balance and classification performance over the Arcene dataset. To demonstrate the effect of data balance on different dimensionality reduction algorithms, we manually construct a training set using the Arcene data where we keep all the negative samples from the training set (56 out of 100) but use different percentage of the positive samples from the training set over all negative samples. We refer to this percentage as

TABLE III
THE OVERALL CLASSIFICATION PERFORMANCE SUMMARY WITH REDUCED DIMENSIONALITY (THE TWO NUMBERS WITHIN THE PARENTHESES INDICATE THE NUMBER OF PROJECTION VECTORS FROM SVM AND ICA, RESPECTIVELY)

Dataset	Method	Classification Accuracy [%] with Reduced Dimensionality for Various Signal-to-Noise Ratios						
		SNR=5 [dB]	SNR=10 [dB]	SNR=20 [dB]	SNR=30 [dB]	SNR=40 [dB]	SNR=50 [dB]	SNR=∞
Arrhythmia	PCA	60.9 (**10**)	61.5 (**10**)	61.1 (**10**)	58.6 (**10**)	57.7 (**10**)	56.9 (**10**)	63.3 (15)
	LDA	54.7 (39)	55.7 (64)	55.1 (78)	55.6 (59)	54.7 (68)	54.5 (93)	55.6 (45)
	PCA+LDA	55.1 (43)	55.4 (36)	55.1 (29)	54.9 (37)	54.6 (36)	54.7 (40)	54.8 (38)
	ICA	60.3 (19)	60.4 (18)	59.7 (34)	58.7 (25)	57.8 (27)	58.3 (20)	62.4 (21)
	kPCA	59.3 (**10**)	61.3 (**10**)	58.4 (60)	56.9 (30)	58.0 (65)	57.1 (55)	62.2 (**10**)
	kDA	47.0 (12)	54.0 (**10**)	57.3 (12)	56.4 (12)	55.5 (**10**)	56.5 (12)	60.6 (12)
	kISOMAP	55.0 (10)	56.0 (15)	56.2 (15)	57.0 (55)	57.0 (25)	56.6 (30)	59.3 (30)
	UKR	60.9 (50)	61.5 (90)	61.5 (40)	61.5 (45)	61.7 (50)	62.6 (80)	63.5 (50)
	kDCV	45.5 (13)	52.2 (**10**)	46.4 (**10**)	50.5 (**10**)	49.1 (13)	49.1 (13)	48.1 (13)
	KDR	62.2 (35)	65.8 (25)	64.4 (60)	64.9 (30)	65.3 (65)	65.3 (65)	**70.7** (60)
	SVM+ICA	**66.2** (28) (13+15)	**67.8** (13) (12+1)	**67.9** (16) (13+3)	**69.3** (44) (12+32)	**68.8** (16) (13+3)	**68.2** (16) (12+4)	68.4 (18) (13+5)
Cancer	PCA	45.7 (**5**)	45.7 (10)	45.7 (**5**)	45.7 (**5**)	45.7 (**5**)	45.7 (**5**)	45.7 (**5**)
	PCA+LDA	53.7 (34)	58.5 (31)	60.7 (22)	62.4 (17)	63.3 (20)	64.1 (18)	62.0 (17)
	ICA	47.6 (38)	48.9 (51)	48.3 (39)	48.0 (34)	48.3 (36)	48.7 (45)	48.0 (42)
	kPCA	50.0 (50)	52.2 (65)	52.2 (60)	52.2 (65)	50.0 (60)	50.0 (60)	50.0 (60)
	kDA	52.4 (12)	54.4 (11)	63.9 (12)	60.0 (13)	63.5 (12)	63.7 (12)	64.1 (13)
	kISOMAP	45.4 (15)	45.4 (40)	47.6 (55)	47.6 (50)	45.4 (70)	45.4 (60)	47.6 (30)
	UKR	63.0 (55)	**69.6** (45)	**69.6** (75)	**71.7** (60)	**71.7** (70)	**71.7** (55)	**71.7** (40)
	kDCV	54.4 (14)	45.7 (**8**)	53.6 (14)	50.7 (14)	56.5 (13)	58.0 (13)	60.1 (14)
	KDR	55.7 (27)	47.8 (18)	45.7 (30)	45.7 (20)	45.7 (28)	45.7 (20)	45.7 (27)
	SVM+ICA	**67.4** (45) (14+31)	67.4 (40) (13+27)	67.4 (36) (9+27)	**71.7** (34) (9+25)	**71.7** (38) (9+29)	**71.7** (38) (9+29)	**71.7** (38) (9+29)

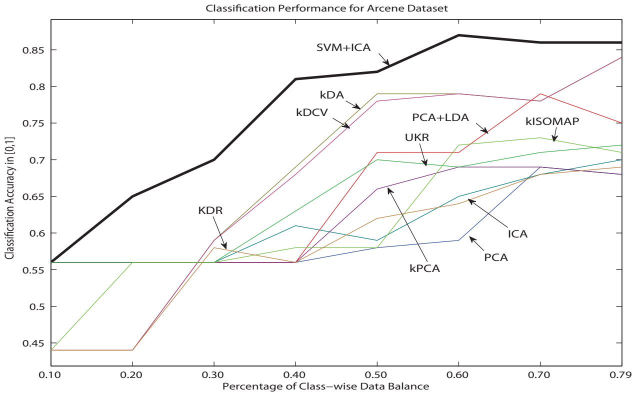
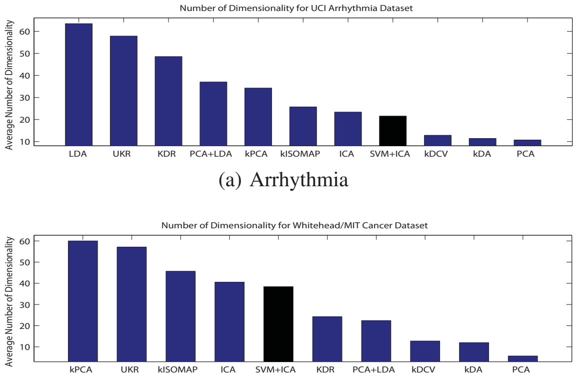


Fig. 5. Behavior of dimensionality reduction methods over various class-wise training data balance



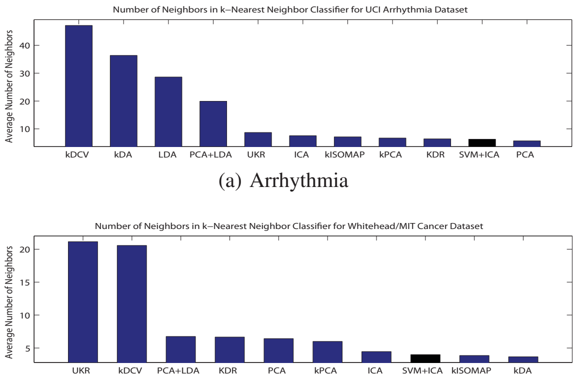
(a) Arrhythmia

(b) Cancer

Fig. 6. Comparison of reduced dimensionality

"percentage of class-wise data balance," on x-axis in Fig. 5 and the percentage is obtained by the number of the selected positive samples divided by the entire negative samples. For fair performance comparison, we apply fixed dimensionality of 50 for all methods. In Fig. 5, the unsupervised methods show better performance than the supervised methods in the 10-20% range (i.e., when the data is very imbalanced) while the supervised methods outperform the unsupervised methods on the percentage beyond 30%. The hybrid PCA+LDA is placed mostly over the unsupervised methods, but lower than the supervised methods due to the limitations of LDA criteria compared with kDA and kDCV alleviating the limitations as described in Sec. I. In KDR, slower performance improvement is observed starting from 30% balance resulting from KDR's objective function not pursuing direct classification but component extraction, usually requiring more dimensions for reduced subspace than kDCV and kDA. SVM+ICA always shows equal or better performance than all the others over various class-wise data balance.

We further compare the average reduced dimensionality for all methods in Fig. 6. Since we fix the maximum dimensionality of kDA to the number of classes minus 1 which is equivalent to the rank of the between-class scatter matrix for multiclass datasets, the maximum dimensionality of kDA for the Arrhythmia and the cancer dataset is 12 and 13, respectively. Additionally, kDCV is designed to provide discriminative common vectors which should be less than or equal to the number of classes. Therefore, the upper bound of reduced dimensional space by kDCV is also limited by the number of classes, 13 and 14 for Arrhythmia and Cancer data, respectively. However, we do not apply the upper bound restriction to LDA or PCA+LDA so that we can observe the behavior of classification performance improvement with the introduction of information from the null space to compensate for the linear model used in LDA against the nonlinear model used in kDA. For the imbalanced Arrhythmia dataset, in general, supervised/hybrid methods such as LDA, KDR, and PCA+LDA return higher dimensionality than unsupervised methods such as kPCA, kISOMAP, and ICA with SVM+ICA standing close to unsupervised approaches. kDA and kDCV have relatively lower dimensionality than LDA and PCA+LDA due to the application of the upper limit. On the contrary, the balanced cancer dataset shows higher dimensionality from the unsupervised approaches but lower dimensionality from supervised approaches, with SVM+ICA standing in between. PCA is an exception here due to early saturation with poor classification performance.



(a) Arrhythmia

(b) Cancer

Fig. 7. Number of neighbors in kNN

We also use the number of neighbors (k) in kNN to observe the performance sensitivity to different data distribution patterns in the dataset. In Fig. 7a, it is clear that supervised kDCV, kDA, and LDA achieve their highest classification accuracies with large ks whereas unsupervised methods such as UKR, ICA, kISOMAP, kPCA, and PCA utilize ks of approximately no more than 10. kNN classifier utilizes large k to achieve the best performance by smoothing the noisy information from the possible non-effective data representation provided by the supervised methods with relatively low performance for the imbalanced Arrhythmia dataset in Fig. 4a. SVM+ICA requires intermediate k between the supervised and unsupervised methods, although k for SVM+ICA is close to the unsupervised approaches. For the cancer dataset, all methods achieve their highest classification accuracies with relatively small ks where $k < 7$ except for UKR and kDCV, as shown in Fig. 7b whereas SVM+ICA presents the smallest k among all.

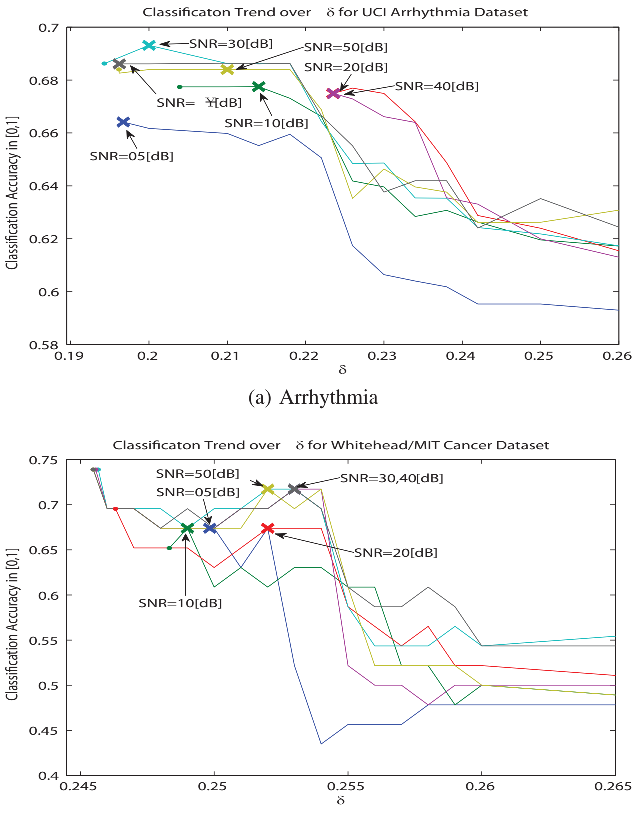
B. Class-wise Performance Comparison

In order to study the effect of imbalanced vs. balanced data distribution, we study the class-wise classification performance using SVM+ICA.

Table IV shows the performance summary of SVM+ICA for the Arrhythmia dataset. The 1st, 2nd, and 10-th classes include more than 9% of total number of data samples. Due to the sufficient number of data for training, the overall performance on this dataset is mostly dependent on the performance of the three classes. However, classes 5, 7, 8, 9, 11, 12, and 13 only have tiny portion of data samples so that SVM+ICA fails to construct appropriate dimensionality reduction model, resulting in poor classification accuracies.

Table V shows the performance summary of SVM+ICA for the cancer dataset. The cancer dataset has relatively balanced amount of data per class compared with the Arrhythmia dataset in Table IV. Since there exists no significant data imbalance in the cancer dataset, the poor performance from the 2nd and 10th classes is expected due to less informative training samples in the classes to reveal the nature of dataset by SVM.

C. Effect of Parameter Selection



(a) Arrhythmia

(b) Cancer

Fig. 8. Trend of classification accuracy corresponding to δ in SVM+ICA

Figure 8 shows the trend of classification accuracy corresponding to δ over various SNR's. By introducing δ, the redundancy removal threshold for SVM features, we provide explicit correspondence of δ with the number of projection vectors from SVM. SVM+ICA selects δ that generates the highest classification accuracy, represented by 'X' mark, so that any projection vectors from SVM with a value of δ lower than the 'X' mark are eliminated by the redundancy removal process. We observe from Figs. 8a and 8b that there exists more redundancy among projection vectors from SVMs obtained by the balanced cancer dataset due to denser projection vectors from SVMs between the lower bound of δ and the selected δ, resulting in more projection vectors to be eliminated as shown in Table V compared with Table IV.

Overall, we conclude that SVM+ICA provides reduced dimensional data representation in relatively low dimensional space by effectively combining both SVM and ICA. The classifier achieves better classification accuracy in the reduced dimensional dataset with robustness against noise.

V. Conclusions

This paper proposed a hybrid dimensionality reduction algorithm, SVM+ICA. The algorithm provides projection that minimizes SVM-based structural risk in supervised manner and maximizes ICA-based data independence in unsupervised manner. Due to the power of structural risk minimization to pursue minimized empirical error and complexity in conjunction with independence maximization to find maximally independent features, SVM+ICA offers projection vectors as a

TABLE IV
CLASSIFICATION PERFORMANCE SUMMARY OF SVM+ICA FOR THE ARRHYTHMIA DATASET

Class	1	2	3	4	5	6	7	8	9	10	11	12	13	Total	
samples	245 (54.2%)	44 (9.7%)	15 (3.3%)	15 (3.3%)	13 (2.9%)	25 (5.5%)	3 (0.7%)	2 (0.4%)	9 (2.0%)	50 (11.1%)	4 (0.9%)	5 (1.1%)	22 (4.9%)	452 (100%)	
SNR [dB]	Classification Accuracy [%]														m
5	93.9	38.6	73.3	46.7	0	8.0	0	0	66.7	52.0	0	0	0	66.2	28(13+15)
10	95.9	31.8	93.3	33.3	7.7	0	0	0	66.7	58.0	25.0	0	4.6	67.8	13(12+1)
20	94.7	47.7	73.3	53.3	0	4.0	0	0	22.2	64.0	0	0	0	67.9	16(13+3)
30	95.1	40.9	86.7	60.0	0	4.0	0	0	66.7	66.0	0	0	0	69.3	44(12+32)
40	94.7	43.2	73.3	66.7	0	4.0	0	0	44.4	64.0	25.0	0	4.6	68.8	16(13+3)
50	96.7	34.1	86.7	40.0	0	4.0	0	0	66.7	56.0	25.0	0	4.6	68.2	16(12+4)
∞	96.3	47.7	80.0	53.3	0	0	0	0	55.6	52.0	0	0	4.6	68.4	18(13+5)

TABLE V
THE CLASSIFICATION PERFORMANCE SUMMARY OF SVM+ICA FOR THE CANCER DATASET

Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total	
samples	11 (5.8%)	10 (5.3%)	11 (5.8%)	11 (5.8%)	22 (11.5%)	11 (5.8%)	10 (5.3%)	10 (5.3%)	30 (15.7%)	11 (5.8%)	11 (5.8%)	11 (5.8%)	11 (5.8%)	20 (10.5%)	190 (100%)	
SNR [dB]	Classification Accuracy [%]															m
5	33.3	0	33.3	100.0	100.0	66.7	100.0	50.0	83.3	0	66.7	33.3	100.0	100.0	67.4	45(14+31)
10	0	0	33.3	100.0	100.0	66.7	100.0	100.0	83.3	0	66.7	33.3	100.0	100.0	67.4	40(13+27)
20	33.3	0	33.3	100.0	100.0	66.7	50.0	100.0	66.7	0	100.0	33.3	100.0	100.0	67.4	36 (9+27)
30	33.3	0	33.3	100.0	100.0	100.0	100.0	100.0	66.7	0	100.0	33.3	100.0	100.0	71.7	34 (9+25)
40	33.3	0	33.3	100.0	100.0	100.0	100.0	100.0	66.7	0	100.0	33.3	100.0	100.0	71.7	38 (9+29)
50	33.3	0	33.3	100.0	100.0	100.0	100.0	100.0	66.7	0	100.0	33.3	100.0	100.0	71.7	38 (9+29)
∞	33.3	0	33.3	100.0	100.0	100.0	100.0	100.0	66.7	0	100.0	33.3	100.0	100.0	71.7	38 (9+29)

mapping from observations to reduced dimensional space including advantages from both approaches simultaneously. The hybrid algorithm gives linear projection to obtain prominent features for better classification performance. Experimental results showed that SVM+ICA outperforms other approaches, including supervised, unsupervised, and conventional hybrid methods in terms of providing better classification performance with relatively low dimensional space, especially in noisy environment.

REFERENCES

[1] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, 2001.

[2] A. Hyvarinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[3] X. Jiang, "Asymmetric principal component and discriminant analyses for pattern classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, pp. 931–937, 2009.

[4] R. Fisher, "The statistical utilization of multiple measurements," *Annals of Eugenics*, vol. 8, pp. 376–386, 1938.

[5] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. R. Mullers, "Fisher discriminant analysis with kernels," in *Proc. IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing IX*, 1999, pp. 41–48.

[6] Y. Guo, T. Hastie, and R. Tibshirani, "Regularized linear discriminant analysis and its application in microarrays," *Biostatistics*, vol. 8, no. 1, pp. 86–100, 2007.

[7] J. Ye, R. Janardan, C. H. Park, and H. Park, "An optimization criterion for generalized discriminant analysis on undersampled problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, pp. 982–994, 2004.

[8] L.-F. Chen, H.-Y. M. Liao, M.-T. Ko, J.-C. Lin, and G.-J. Yu, "A new LDA-based face recognition system which can solve the small sample size problem," *Pattern Recognition*, vol. 33, no. 10, pp. 1713–1726, 2000.

[9] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana, "Discriminative common vectors for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 1, pp. 4–13, 2005.

[10] H. Cevikalp, M. Neamtu, and M. Wilkes, "Discriminative common vector method with kernels," *IEEE Trans. on Neural Netw.*, vol. 17, no. 6, pp. 1550–1565, 2006.

[11] H. Park, M. Jeon, and J. B. Rosen, "Lower dimensional representation of text data based on centroids and least squares," *BIT Numerical Mathematics*, vol. 43, no. 2, pp. 427–448, 2003.

[12] M. Kyperountas, A. Tefas, and I. Pitas, "Weighted piecewise LDA for solving the small sample size problem in face verification," *IEEE Trans. Neural Netw.*, vol. 18, no. 2, pp. 506–519, 2007.

[13] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, 1997.

[14] J. Yang and J.-y. Yang, "Why can LDA be performed in PCA transformed space?" *Pattern Recognition*, vol. 36, no. 2, pp. 563–566, 2003.

[15] P.-F. Hsieh and D. Landgrebe, "Linear feature extraction for multiclass problems," in *Proc. IEEE Int. Geoscience and Remote Sensing Symp.*, vol. 4, 1998, pp. 2050–2052.

[16] M. Loog, R. P. W. Duin, and R. Haeb-Umbach, "Multiclass linear dimension reduction by weighted pairwise Fisher criteria," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 7, pp. 762–766, 2001.

[17] P.-F. Hsieh, D.-S. Wang, and C.-W. Hsu, "A linear feature extraction for multiclass classification problems based on class mean and covariance discriminant information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2,

pp. 223–235, 2006.

[18] Y. Xu, J.-y. Yang, and J. Yang, “A reformative kernel Fisher discriminant analysis,” *Pattern Recognition*, vol. 37, no. 6, pp. 1299–1302, 2004.

[19] A. Shashua, “On the relationship between the support vector machine for classification and sparsified Fisher’s linear discriminant,” *Neural Processing Lett.*, vol. 9, no. 2, pp. 129–139, 1999.

[20] Q. Tao, D. Chu, and J. Wang, “Recursive support vector machines for dimensionality reduction,” *IEEE Trans. Neural Netw.*, vol. 19, no. 1, pp. 189–193, 2008.

[21] I. W.-H. Tsang, A. Kocsor, and J. T.-Y. Kwok, “Large-scale maximum margin discriminant analysis using core vector machines,” *IEEE Trans. Neural Netw.*, vol. 19, no. 4, pp. 610–624, 2008.

[22] C. Xiang, X. A. Fan, and T. H. Lee, “Face recognition using recursive Fisher linear discriminant,” *IEEE Trans. Image Processing*, vol. 15, no. 8, pp. 2097–2105, 2006.

[23] H. Wold, “Estimation of Principal Components and Related Models by Iterative Least Squares,” in *Multivariate Analysis*. Academic Press, 1966, pp. 391–420.

[24] R. Rosipal and L. J. Trejo, “Kernel partial least squares regression in reproducing kernel Hilbert space,” *J. Mach. Learn. Res.*, vol. 2, pp. 97–123, 2002.

[25] J. Tenenbaum, V. d. Silva, and J. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[26] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[27] R. Vidal, Y. Ma, and S. Sastry, “Generalized principal component analysis (GPCA),” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1945–1959, 2005.

[28] P. Campoy, “Dimensionality reduction by self organizing maps that preserve distances in output space,” in *Proc. Int. Joint Conf. on Neural Networks*, 2009, pp. 432–438.

[29] B. Scholkopf, A. Smola, and K. Muller, “Nonlinear Component Analysis as a Kernel Eigenvalue Problem,” *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

[30] F. R. Bach and I. Jordan, Michael, “Kernel independent component analysis,” *J. Mach. Learn. Res.*, vol. 3, pp. 1–48, 2002.

[31] P. Meinicke, S. Klanke, R. Memisevic, and H. Ritter, “Principal surfaces from unsupervised kernel regression,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 9, pp. 1379–1391, 2005.

[32] S. Bourennane, C. Fossati, and A. Cailly, “Improvement of classification for hyperspectral images based on tensor modeling,” *IEEE Geoscience and Remote Sensing Lett.*, vol. 7, no. 4, pp. 801–805, 2010.

[33] J. Ham, D. D. Lee, S. Mika, and B. Scholkopf, “A kernel view of the dimensionality reduction of manifolds,” in *Proc. Int. Conf. on Machine Learning*, 2004.

[34] H. Choi and S. Choi, “Robust kernel isomap,” *Pattern Recognition*, vol. 40, no. 3, pp. 853–862, 2007.

[35] K. Kwak and W. Pedrycz, “Face recognition using an enhanced independent component analysis approach,” *IEEE Trans. Neural Netw.*, vol. 18, pp. 530–541, 2007.

[36] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas, “Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification,” *IEEE Trans. Neural Netw.*, vol. 17, no. 3, pp. 683–695, 2006.

[37] S. Zafeiriou, “Discriminant nonnegative tensor factorization algorithms,” *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 217–235, 2009.

[38] J. M. Leiva-Murillo and A. Artes-Rodriguez, “Maximization of mutual information for supervised linear feature extraction,” *IEEE Trans. Neural Netw.*, vol. 18, no. 5, pp. 1433–1441, 2007.

[39] K. Fukumizu, F. R. Bach, and M. I. Jordan, “Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces,” *J. Mach. Learn. Res.*, vol. 5, pp. 73–99, 2004.

[40] R. Gilad-Bachrach, A. Navot, and N. Tishby, “Margin based feature selection - theory and algorithms,” in *Proc. Int. Conf. on Machine Learning*, 2004, pp. 43–50.

[41] J. Yang, D. Zhang, and J.-y. Yang, “Is ICA significantly better than PCA for face recognition?” in *Proc. Int. Conf. on Computer Vision*, vol. 1, 2005, pp. 198–203.

[42] J. Yang, D. Zhang, and J.-Y. Yang, “Constructing PCA baseline algorithms to reevaluate ICA-based face-recognition performance,” *IEEE Trans. Syst. Man Cybern. B, Cybern.*, vol. 37, no. 4, pp. 1015–1021, 2007.

[43] J. Ye, “Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems,” *J. Mach. Learn. Res.*, vol. 6, pp. 483–502, 2005.

[44] V. Vapnik, “An overview of statistical learning theory,” *IEEE Trans. Neural Netw.*, vol. 10, pp. 988–999, 1999.

[45] J. Karhunen, E. Oja, L. Wang, R. Vigario, and J. Joutsensalo, “A class of neural networks for independent component analysis,” *IEEE Trans. Neural Netw.*, vol. 8, no. 3, pp. 486–504, 1997.

[46] C.-W. Hsu and C.-J. Lin, “A comparison of methods for multiclass support vector machines,” *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, 2002.

[47] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[48] A. Asuncion and D. Newman, “UCI machine learning repository.” [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html

[49] “Cancer diagnosis dataset.” [Online]. Available: http://www-genome.wi.mit.edu/cancer

[50] J. C. Platt, “Sequential minimal optimization: A fast algorithm for training support vector machines,” Microsoft Research, Technical Report MSR-TR-98-14, 1998.

[51] T. Joachims, *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 2002.

[52] C.-C. Chang and C.-J. Lin, “LIBSVM: A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[53] M. Turk and A. Pentland, “Eigenfaces for recognition,” *J. Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.