Optimized multilayer perceptrons for molecular classification and diagnosis using genomic data

Zuyi Wang1,2, Yue Wang3,*, Jianhua Xuan2, Yibin Dong3, Marina Bakay1, Yuanjian Feng3, Robert Clarke4 and Eric P. Hoffman1

1Center for Genetic Medicine, Children’s National Medical Center, Washington, DC 20010, USA, 2Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064, USA, 3The Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA and 4Departments of Oncology, Physiology and Biophysics, Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC 20007, USA

ABSTRACT

Motivation: Multilayer perceptrons (MLP) represent one of the widely used and effective machine learning methods currently applied to diagnostic classification based on high-dimensional genomic data. Since the dimensionalities of the existing genomic data often exceed the available sample sizes by orders of magnitude, the MLP performance may degrade owing to the curse of dimensionality and over-fitting, and may not provide acceptable prediction accuracy.

Results: Based on Fisher linear discriminant analysis, we designed and implemented an MLP optimization scheme for a two-layer MLP that effectively optimizes the initialization of MLP parameters and MLP architecture. The optimized MLP consistently demonstrated its ability in easing the curse of dimensionality in large microarray datasets. In comparison with a conventional MLP using random initialization, we obtained significant improvements in major performance measures including Bayes classification accuracy, convergence properties and area under the receiver operating characteristic curve (Az).

Supplementary information: The Supplementary information is available on http://www.cbil.ece.vt.edu/publications.htm

Contact: yuewang@vt.edu

1 INTRODUCTION

Diagnostic classification with genomic data refers to the assignment of a particular unknown tissue sample to a known disease class based on its quantitative mRNA expression pattern from microarrays. This classification can be performed by a trained predictive classifier, such as a neural network classifier. This approach is particularly helpful for diagnosing complex genetic disease subtypes or stages whose subtle differences may be difficult to recognize by traditional clinical and pathological approaches (Bittner et al., 2000; Brown et al., 2000; Khan et al., 2001; Mjolsness and DeCoste, 2001; Ramaswamy et al., 2001; Shipp et al., 2002; West et al., 2001; Linder et al., 2004; O’Neill and Song, 2003; Wei et al., 2004). A common type of neural network classifier applied to diagnostic classification is feed-forward back-propagation multilayer perceptrons (MLP) (Fig. 1). Input vectors and the corresponding target vectors are used to train an MLP, a process that updates the weights and biases until the MLP can approximate a mapping function that associates input vectors with specific output vectors. The generalization property makes it possible to train an MLP with a representative set of input/target pairs and get good results for predicting unseen input samples. The ability of an MLP to learn complex (non-linear) and multi-dimensional mapping from a collection of examples makes it an ideal classifier for diagnostic classification (Haykin, 1999; Khan et al., 2001; O’Neill and Song, 2003; Wei et al., 2005).

Despite reported successful studies on applying MLPs to diagnoses with genomic data, such as gene expression microarray data (Khan et al., 2001; Linder et al., 2004; O’Neill and Song, 2003; Wei et al., 2005), the most critical problem, that of the curse of dimensionality, has not been effectively addressed. The curse of dimensionality is caused by the finite amount of training data available relative to the large input feature space. Accordingly, when the dimensionality increases considerably and the available information remains inadequate, the large number of model parameters in the classifier cannot be well-trained (Haykin, 1999; Jain et al., 2000). Consequently, the classifier’s performance may degrade beyond a certain point with the increasing inclusion of features or dimensions. In mRNA microarray experiments, there is typically an extremely ill-conditioned ratio of sample number (tens to hundreds) to dimension number (probe or probe sets typically >10000), which greatly augments the impact of the curse of dimensionality (Fukunaga, 1990; Haykin, 1999). In current studies, the approaches to avoid the curse of dimensionality are generally limited to directly reducing the number of inputs. The commonly applied methods include conventional dimensionality reduction methods, such as principal component analysis (Khan et al. 2001, Wei et al., 2005), t-statistics (Golub et al., 1999), correlation measure (van’t Veer et al., 2002) and an MLP training-based gene selection procedure that selects genes with greater influence on the changes of outputs in an MLP (O’Neil and Song, 2003).

The design parameters in training an MLP include initial values of the model parameters (synaptic weights and biases), stopping rules, MLP architecture, etc. Since no effective algorithms are...
available to search for a global optimum and traditional MLP initialization is done randomly, classification performance depends largely on the initial values of weights and biases. Furthermore, the higher complexity of the classifiers often results in more local minima in the error surface, and the classifier trainings can easily be trapped into such local minima (Raudys and Skurikhina, 1992; Raudys, 1994).

We hypothesized that developing an optimization of MLP initialization will allow the reduction of the curse of dimensionality, and therefore improve performance of the MLP. Our goal was to find an effective non-random initialization scheme that places the initial state of an MLP closer to the optimal solution that is later sought by training (Wang et al., 2004). This approach is bolstered by previous studies in statistical pattern recognition field, where it has been shown that non-random initializations of MLP weights and biases resulted in the MLP with small generalization error even when the number of samples is smaller than the number of features or dimensions (Raudys, 1994; Raudys, 1997; Raudys and Skurikhina, 1992).

2 THEORY AND METHOD

2.1 wFC-based MLP Initializations

2.1.1 Linear dimension reduction and MLP feature extraction—hidden layer initialization

The MLP offers an integrated procedure for feature extraction and Bayes classification by learning the decision boundary (Haykin, 1999). Its feed-forward auto-associative architecture can also be used to construct non-linear subspaces in a supervised or unsupervised mode (Haykin, 1999; Jain et al., 2000). The output of the hidden layer may be interpreted as a set of new features presented to the normal mixture (SFNM) model (Loog et al., 2000). The output of the hidden layer to capture the most discriminatory components or subspaces for distinguishing the classes. Based on these theoretical observations, we suggest a wFC-based initialization mechanism for the MLP hidden layer (Wang et al., 2004). To limit the complexity of the MLP, we assume that the number of neurons in the hidden layer is smaller than the number of inputs.

Given an \( m_0 \)-dimensional input \( t \)-space with \( K_0 \) classes, the multi-class LDR searches for a linear transformation \( W \) that transforms the original input space to a lower \( m_1 \)-dimensional feature \( x \)-space (\( m_1 < m_0 \)); the extracted \( k \)-space should preserve the maximum amount of class discriminating information. Since it is too complex to directly use the Bayes error as a criterion, the most common technique for finding this transformation is LDR that is based on Fisher criterion (Jain et al., 2000; Haykin, 1999). This method maximizes the ratio of the between-class scatter matrix to the within-class scatter matrix, thereby guaranteeing maximal separability. In this paper, we apply the wFC to the multi-class classification problem (Loog et al., 2001), and the wFC is defined as

\[
J_{\text{wFC}}(W) = \left( \sum_{i=1}^{K_0} \sum_{j=1}^{K_0} w_{ij} \pi_i \pi_j \left( D_{ij} \right) \right) \left( \sum_{i=1}^{K_0} \sum_{j=1}^{K_0} w_{ij} \pi_i \pi_j \left( D_{ij} \right) \right)^{-1}.
\]

where \( W \) is the linear transformation matrix, \( \pi_i \) and \( \pi_j \) are the prior probabilities of classes \( k \) and \( l \) respectively, \( S_0 = \sum_{i=1}^{K_0} \pi_i C_{ii} \) is the total within-class scatter matrix, and \( S_{kl} = (\mu_k - \mu_l)(\mu_k - \mu_l)^T \) is the between-class scatter matrix for classes \( k \) and \( l \). \( \omega(\Delta_{il}) \) is the weighting function defined as

\[
\omega(\Delta_{il}) = \frac{1}{2} \log \frac{1 + \Delta_{il}}{1 - \Delta_{il}}.
\]

where \( \Delta_{il} = \left| (\mu_k - \mu_l)^T S_{0}^{-1} (\mu_k - \mu_l) \right| \) is the Mahalanobis distance between classes \( k \) and \( l \) with class mean vector \( \mu_k \) and covariance matrix \( C_k \). It has been shown that when there are more than two classes to be classified, the conventional multi-class Fisher criterion (cFC) for deriving dimension-reduced subspace is suboptimal with respect to classification (Loog et al., 2001). The reason is that the cFC treats class pairs with various between-class distances equally. In contrast, the wFC incorporates a weight function that approximates the Bayes error rate between classes, and assigns larger weights to the closer class pairs and smaller weights to the distant pairs. Thus, in the extracted subspace found by wFC, the classes with heavy overlap gain adequate emphasis, and the distant pairs remain well separated.

Finding a solution \( W \) that maximizes the wFC is essentially a problem of eigenvalue decomposition of the total Fisher scatter matrix.

\[
S_{0}^{-1} = \sum_{i=1}^{K_0} \sum_{j=1}^{K_0} w_{ij} \pi_i \pi_j \left( D_{ij} \right) S_{kl}.
\]

By taking only the \( m_1 \) eigenvectors corresponding to the \( m_1 \) largest eigenvalues \( (m_1 < m_0) \), we can form a transformation that not only reduces the dimensionality of the original input space, but also retains maximal class separability information. We call this procedure wFC-Discriminatory Component Analysis (wFC-DCA).

With the transformation \( W \) \((m_0 \times m_1)\) derived from LDR, the dimension-reduced feature subspace \( x \)-space with \( m_1 \) dimensions becomes \( x_i = W_i^T (t_i - b_0) \); for \( i = 1, \ldots, N \), where \( N \) is the number of samples, \( x_i \) is the representation of the sample vector \( t_i \) in the \( x \)-space with \( x_i = W_i^T (t_i - b_0) \) for \( r = 1, \ldots, m_0 \), and \( b_0 \) is the global center of the dataset. On the other hand, the outputs of the hidden layer in the MLP (Fig. 1) can be acquired as \( h_i = \phi(W_i^T \mathbf{p} - b_{1i}) \), where \( W_i \) is the set of synaptic weights connecting \( m_0 \) inputs to neuron \( n \) at the hidden layer, \( b_{1i} \) is the bias of hidden neuron \( n \) and \( \phi(\cdot) \) is an activation function (Haykin, 1999). The connection between the LDR and the MLP feature extraction mechanism now becomes clearer, suggesting that the column vectors of the LDR matrix \( W \) can be used to initialize the weights between the input and hidden layer of an MLP, \( \mathbf{w}_i = W_i \mathbf{b}_0 \), and their biases can be initialized as, \( b_{1i} = W_i^T \mathbf{b}_0 \).
features are further scaled by the activation function \( \varphi(\cdot) \) that could be linear or non-linear. It has been theoretically shown that the minimization of the Bayes error with respect to the synaptic weights and biases of the MLP is equivalent to maximizing the wFC [Equation (1)]. and it can be entirely determined by the hidden neurons (Haykin, 1999).

### 2.1.2 LDA and multi-class perceptrons—output layer initialization

Since the outputs of the hidden layer serve as new features, Fisher LDA determines a linear transformation for converting an \( m_1 \)-dimensional problem to a one-dimensional (1D) problem (Haykin, 1999). Consider the variable \( y = w^T x - b \) that is transformed from \( x \)-space to a 1D space via LDA, and the LDA is defined by

\[
J(w) = w^T S_{xL} w \quad \text{s.t.} \quad x^T S_x x > 0
\]

that is known as the generalized Rayleigh quotient. The solution that maximizes \( J(w) \) is simply \( w = S_{xL}^{-1} (\mu_1 - \mu_2) \), which is also a generalized eigenvalue problem.

The neuron in the output layer behaves similarly as a perceptron that can be considered as a decision making element that bears a close resemblance to the Bayes classifier, and has been generalized to multiple classes (Haykin, 1999). Specifically, the outputs of the neurons in the output layer are computed as, \( y_i = \varphi(w^T_i a - b_i) \) for \( i = 1, \ldots, m_2 \), where \( a \) is the output vector of the hidden layer, \( w_i \) is the set of weights connecting the hidden layer and the output neuron \( i \), \( b_i \) is the bias of output neuron \( i \) and \( m_2 \) is the number of the output neurons, i.e. the number of classes (Fig. 1). Consider a two-class case with a linear activation function \( \varphi(\cdot) \), we have \( y = w^T x - b \) with \( w = S_{xL}^{-1} (\mu_1 - \mu_2) \) and \( b = w^T b_0 \), where \( b_0 = (\mu_1 + \mu_2)/2 \). We can use two output neurons to derive a class-dependent representation by rearranging the output as, \( y = w^T x - b = (w^T_1 x - b_1) - (w^T_2 x - b_2) = y_1 - y_2 \), where \( w_1 = S_{xL}^{-1} \mu_1 \), \( w_2 = S_{xL}^{-1} \mu_2 \), \( b_1 = w^T b_0 \) and \( b_2 = w^T b_0 \), so we have \( y_1 = w_1^T x - b_1 \) and \( y_2 = w_2^T x - b_2 \). Figure 2a illustrates such an interpretation. Based on the above derivation, the class-dependent Fisher linear discriminant transformation \( w \) can be again used to initialize the weights between the hidden and output neurons as \( w_i = S_{xL}^{-1} \mu_i \), and the biases of the output neurons can be, \( b_{2i} = w_i^T b_0 \) for \( i = 1, 2 \). Accordingly, for a three-class case, it is straightforward to have \( w_i = S_{xL}^{-1} \mu_i \), \( b_{1i} = w_i^T b_0 \), \( w_{i1} = S_{xL}^{-1} \mu_1 \), \( b_{12} = w_{i2}^T b_0 \) and \( w_{i2} = S_{xL}^{-1} \mu_2 \), \( b_{23} = w_{i3}^T b_0 \), where \( b_0 = (\mu_1 + \mu_2 + \mu_3)/3 \). Figure 2b depicts this case. Notice that such an initialization is readily applicable to single-layer perceptrons.

### 2.1.3 Determining the size of the hidden layer

The wFC-based MLP initialization method may also suggest a suitable number of hidden neurons, a key component of MLP architecture. Neural networks, like other flexible non-linear estimation methods, are vulnerable to problems of under-fitting and over-fitting (Haykin, 1999; Ripley, 1996). The over-fitting problem occurs more easily when the number of samples in the training set is small and the network is relatively large, which is the case for most genomic data. Therefore, it is important to use a network that is just large enough to provide an adequate fit. The resulting subspace represented by the outputs of the hidden layer should maintain as much class separability as possible (Haykin, 1999): the retained partial separability is given by \( J_{wFC} = (W) \) [Equation (1)]. Hence, it is appropriate to let the number of pseudo genes (i.e. \( m_2 \), the number of hidden neurons) be the number of significant eigenvalues derived from wFC-DCA because the eigenvalues represent class separability in feature space. In this study, we select the dominant eigenvalue subset that contains 99% of the total separability, and let the number of hidden neurons be equal to the number of selected eigenvalues.

### 2.2 Selection of MLP inputs

Input selection is a prerequisite for diagnostic classification using genomic data; we apply our newly developed two-step wFC-based input selection method (Xuan et al., 2004) that shares the same theoretical basis (wFC) with the proposed MLP initialization approach. First, we rank all genes based on their individual discriminatory power measured by the 1D wFC (Xuan et al., 2004); a gene will be selected as an individually discriminatory gene (IDG) if its discriminatory power is above an empirical threshold. Second, from the IDG pool, we select jointly discriminatory gene (JDI) subsets (with various sizes) whose joint discriminatory power is the maximum among all sets of the same size. The joint discriminatory power is also determined by the multi-dimensional version of wFC [Equation (1)]. Furthermore, the JDI sets are refined by testing on a trained MLP, which ultimately determines the ‘optimal’ diagnostic gene subset that minimizes the generalization error. From the curve of classification rate versus JDI subsets, we pick the optimal JDI subset that corresponds to the maximal classification rate as the final inputs for the MLP. This step boosts the MLP performance, and also determines its number of inputs (\( m_0 \), Fig. 1).

### 3 EXPERIMENTAL VERIFICATION

#### 3.1 Data

To highlight the biological and clinical relevance, we chose diagnostic tasks that are difficult for standard clinical and pathological methods alone. The following list summarizes the microarray datasets tested in this study.

- **(1)** Limb-girdle muscular dystrophy (LGMD, provided by Children National Medical Center, Center for Genetic Medicine): four diagnostic groups, Fukutin related protein deficiency (FKRP) (homozygous missense for glycosylation enzyme, limb-girdle muscular dystrophy sub-type, \( n = 7 \)), Becker muscular dystrophy (BMD, hypomorph for dystrophin, \( n = 5 \)), Dysferlin deficiency (putative vesicle traffic defect, \( n = 10 \)), and Calpain III deficiency (\( n = 10 \)), total 32 samples, 22,283 genes.
- **(2)** Leukemia (Kohlmann et al., 2004): three diagnostic groups, T-ALL (\( n = 9 \)), MLL (\( n = 10 \)) and BCR-ABL (\( n = 15 \)), total 34 samples, 312 genes.
- **(3)** Central nervous system (CNS) cancer (Pomeroy et al., 2002): five diagnostic groups, Medulloblastomas (\( n = 60 \)), Malignant glioma (\( n = 10 \)), Rhabdoid tumours (\( n = 10 \)), Normal cerebella (\( n = 4 \)), Supratentorial PNET (\( n = 6 \)), total 90 samples, 7,129 genes.

#### 3.2 Results

The experiments were designed to show the impact of the proposed MLP optimization method on two major aspects of MLP performance: prediction accuracy and training efficiency. For the
prediction accuracy, we examined classification rate and area under the curve (Az) from receiver operating characteristic (ROC) analysis; to probe the training property we recorded initial error (mean squared error, MSE) between target and output before training, final error (MSE) after training, total number of epochs needed for convergence and percentage of converged training.

In all experiments with MLP training and testing, we applied 100 iterations of stratified 3-fold cross validations in order to ensure reliability, and all performance measures were calculated based on the results from the cross validations. In the stratified 3-fold cross validation, the dataset is randomly divided into three subsets of equal size, and the proportion of each class in each subset remains the same as that in the entire set. In each fold, one of the subsets is used for testing and the rest are combined for training; in each iteration, the training is repeated until all subsets have been used for testing.

The optimized MLPs (oMLP, wFC-based initialization) consistently outperformed conventional MLPs (cMLP, conventional random initialization) for all different tested JDG subsets (Fig. 3). We selected 200 JDG subsets consisting of 1–200 genes as inputs of the MLPs. Figure 3 plotted the curves of the classification rate from the test set (those samples not used for MLP training) versus JDG subsets, which is part of the step 2 in the two-step input selection procedure. To determine the optimal JDG subset among the 200 candidate subsets, the oMLP and cMLP were trained with the same training set and tested with the same test set in each fold for fairness and reliability. The search of the optimal JDG subset was considered sufficient when the classification rate of the oMLP did not increase substantially and the classification rate of the cMLP decreased consistently over 20 JDG sets. The oMLP was able to maintain high classification rate as the size of the JDG increased, whereas the cMLP performance degraded. Moreover, the smaller standard deviation of the oMLP classification rate across all cross validations indicated that the oMLP provided more stable performance (Table 1).

Additionally, as ROC analysis (Metz, 1986) has been widely recognized as the most meaningful assessment of medical diagnostic performance (Metz, 1986), we also evaluated relative prediction performance of the oMLP and cMLP using a one-against-rest ROC analysis (Hand and Till, 2001) that was specifically designed for the multi-class classification. The ROC analysis offers a description of the trade-offs between true positive fraction (TPF) and false positive fraction (FPF) of a detection test as the decision threshold varies. In the one-against-rest ROC analysis, the approximated posterior probabilities (the outputs of an MLP) of test samples were recorded, and a two-class ROC analysis was applied to all combinations of one-class against the rest classes. For example, there will be n ROC curves for an n-class classification task. A ROC curve plots TPF versus FPF; generally the larger the index, Az, the better the prediction performance of the classifier. With the optimal JDG subset as inputs, the oMLP had greater Az values for all one-against-rest combinations than the cMLPs, therefore showed better overall performance (Fig. 4). Within each individual case, the larger difference between the prediction
The accuracies of the oMLP and cMLP correspond to the larger differences in $A_z$ values (Fig. 4 and Table 1).

The evaluations of training properties on the oMLP and cMLP with the optimal JDG subset as inputs clearly demonstrated the effectiveness of the proposed initialization approach (Table 1 and Fig. 5). The smaller averages of initial and final MSE and the smaller STD of the final MSE in the oMLP trainings, also shown by the training curves (Fig. 5), provided clear evidence that the proper initialization offered a better starting training point so that the trainings were led to a better and less diverse convergence point. In addition, we monitored whether each training process converged by recording the percentage of converged trainings. Note that a training process is considered as converged only if it meets the error goal or is stopped by a standard early stop procedure we applied in all MLP trainings to prevent over-training. The result showed that 100% of the oMLP trainings converged, but a number of cMLP trainings were eventually terminated by a preset maximal number of epochs (Table 1). Moreover, the smaller average and STD of the number of total epochs needed by the oMLP to achieve convergence further confirmed that the oMLP needed less computational resources to reach higher classification rate (Table 1).

The two-step input selection procedure is effective and computationally feasible in handling a large number of genes so that the curse of dimensionality problem is significantly reduced to a more manageable scale. The considerable change of the classification rate over the entire curve (Fig. 3) confirmed that the content and size of the inputs strongly influenced MLP performance. Particularly, since it shares the same theoretical criterion with the proposed MLP initialization method (wFC), their joint influence is augmented.

We further compared the oMLP with two of the most commonly applied classifiers, $K$-nearest neighbor (KNN) and one-versus-rest support vector machine (OVR-SVM) that is a typical type of multi-class SVMs (Ramaswamy et al., 2001; Statnikov et al., 2005). Both KNN and OVR-SVM undertook rigorous optimizations for seeking optimal performance. We determined the parameter $K$ in the KNN model based on 100 iterations of 3-fold cross validations. Each SVM unit in the OVR-SVM was tested for seven different kernel functions (linear, second and third order polynomials, and Gaussians with scale factors, 0.01, 0.1, 0.5 and 1.0), and five penalty values $C = 0.001, 0.01, 0.1, 1.0$ and 10.0. The KNN took the optimal JDG set as inputs; the OVR-SVM took two types of inputs, optimal JDG set and all genes. In summary, the OVR-SVM with optimal JDG set as inputs and the oMLP provided excellent and comparable accuracies and the OVR-SVM demonstrated small improvements, whereas the KNN and cMLP generally showed inferior performances (Table 2). Detailed results of these comparative experiments can be found in Appendix A.

![Fig. 4](https://academic.oup.com/bioinformatics/article-abstract/22/6/755/296044)

**Fig. 4.** The one-against-rest ROC curves of the oMLP and cMLP with the optimal JDG set as inputs, (a) LGMD, (b) leukemia and (c) CNS cancer. Each dataset has several sets of ROC curves for the oMLP and cMLP, and the number of sets is equal to the number of classes in each dataset. The corresponding $A_z$ values for the oMLP and cMLP are displayed with the curves. The oMLP consistently showed superior performance over the cMLP for all datasets with larger $A_z$ values.

### 4 CONCLUSIONS AND DISCUSSIONS

By suggesting an initialization technique based on the wFC and the link between the MLP mechanism and Fisher LDA, together with the input selection procedure, we offer an efficient and practical MLP prototype that can ease the curse of dimensionality in multi-class high-dimensional genomic data classification and provide excellent generalization performance. The wFC-based initialization procedure initiates the MLP close to the optimal condition.
and-conquer principle to the dataset and accordingly introduce be required. An elegant yet simple method is to apply divide-through dual-purpose training, or one additional hidden layer may be necessary to transform a non-linearly separable problem into a linear classification. This may be achieved by the existing hidden layer transforming a non-linearly separable problem into a linear problem after dimensionality reduction according to Cover’s theorem on the separability of patterns. Therefore, the hidden layer of an untrained oMLP is able to extract precise discriminant components when the distribution of each diagnostic group is affected by the deviation of the distribution of a diagnostic group from a standard single multivariate Gaussian to a mixture of Gaussian (Appendix C). Although the wFC may only find less precise discriminant components when the distribution of each class cannot be closely modeled by a single Gaussian distribution, some pseudo-classes to the output layer, such that all class-pairs become linearly separable. Notice that the discrete decision fusion can be readily and effortlessly done without using any combiner, since the pseudo-classes belong to some of the known classes a priori. It is important to note that a net reduction in MLP complexity can still be achieved when \( m_0 \) is large, since the total number of weights in a two-layer MLP is \( m_1 (m_0 + m_2) \) such that the reduction due to \( m_1 \) surpasses the generally limited increase due to \( m_2 \). Refinements, allowing a co-determination of \( m_1 \) and \( m_2 \), may further reduce the curse of dimensionality and improve the generalization performance.

A complex multi-class classification task is beyond the capability of a single classifier. It is remarkable that the single classifier, oMLP, can compete with the OVR-SVM built with a collection of single binary SVMs and show comparable outstanding performance when the number of classes is relatively small (\( \leq 5 \), more experimental results in Appendix A). However, the OVR-SVM is generally expected to outperform most existing classifiers as the number of classes increases (Statnikov et al., 2005).

As another verification of the effectiveness of the MLP initialization, we tested and compared the untrained oMLP and cMLP, and the results showed that the untrained oMLP considerably outperformed the untrained cMLP (Appendix B). Even without training, the hidden layer of an untrained oMLP is able to extract discriminant features derived from the wFC, then the neurons in the output layer can perform linear one-versus-rest classifications based on these extracted features. We used linear transfer functions in the hidden neurons and log-sigmoid transfer function in the output neurons. Hence, an untrained oMLP closely resembles LDA, and the initial condition of the oMLP (i.e. performance of the untrained oMLP) reflects the performance of LDA.

Using simulation experiments, we demonstrated that the capability of the proposed MLP optimization method is not significantly affected by the deviation of the distribution of a diagnostic group from a standard single multivariate Gaussian to a mixture of Gaussian (Appendix C). Although the wFC may only find less precise discriminant components when the distribution of each class cannot be closely modeled by a single Gaussian distribution, such loss of information is expected to be small and can be well-compensated by further training of weights and biases that offers extensive degrees of freedom in modeling decision boundary.

### Table 2. The classification rate of the model with the best performance for the KNN and OVR-SVM. The results are listed as average (STD)

<table>
<thead>
<tr>
<th>Data</th>
<th>KNN (Optimal IDG set)</th>
<th>OVR-SVM (All genes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LGMD</td>
<td>41.33 (12.66)</td>
<td>50.94 (12.58)</td>
</tr>
<tr>
<td></td>
<td>100.00 (0.00)</td>
<td>95.34 (5.97)</td>
</tr>
<tr>
<td>Leukemia</td>
<td>88.39 (8.73)</td>
<td>95.10 (5.97)</td>
</tr>
<tr>
<td></td>
<td>Linear, Gaussian 1.0</td>
<td>Linear Linear</td>
</tr>
<tr>
<td>CNS cancer</td>
<td>86.59 (4.65)</td>
<td>89.13 (3.49)</td>
</tr>
<tr>
<td></td>
<td>Gaussian 1.0</td>
<td>Linear</td>
</tr>
</tbody>
</table>

Fig. 5. The training curves of the oMLP and cMLP with the optimal JDG set as inputs, (a) LGMD, (b) leukemia and (c) CNS cancer. The training properties, e.g. the convergence speed, initial and the final errors, can be clearly observed from the figures. The trainings of the oMLP usually started from smaller initial errors and converged to smaller final errors, whereas the cMLP training started from and converged to larger and more diverse errors. All these improved properties supported the advantage of the wFC-based initialization over the conventional random initialization.
ACKNOWLEDGEMENTS

This study was supported in part by the National Institutes of Health grants under CA109872, CA096483 and EB000830, and DOD/CDMRP grant under BC030280. Z.W. was also supported by the Crystal Ball of Virginia Beach VAs and the Muscular Dystrophy Association.

Conflict of Interest: none declared.

REFERENCES


