

Accepted Manuscript

Multi-task Deep Convolutional Neural Network for Cancer Diagnosis

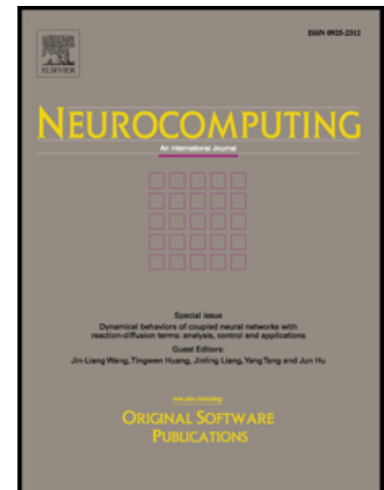
Qing Liao, Ye Ding, Zoe L. Jiang, Xuan Wang, Chunkai Zhang,
Qian Zhang

PII: S0925-2312(18)31282-7
DOI: <https://doi.org/10.1016/j.neucom.2018.06.084>
Reference: NEUCOM 20106

To appear in: *Neurocomputing*

Received date: 14 January 2018
Revised date: 13 June 2018
Accepted date: 25 June 2018

Please cite this article as: Qing Liao, Ye Ding, Zoe L. Jiang, Xuan Wang, Chunkai Zhang, Qian Zhang, Multi-task Deep Convolutional Neural Network for Cancer Diagnosis, *Neurocomputing* (2018), doi: <https://doi.org/10.1016/j.neucom.2018.06.084>



This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Multi-task Deep Convolutional Neural Network for Cancer Diagnosis

Qing Liao¹, Ye Ding², Zoe L. Jiang¹,
Xuan Wang¹, Chunkai Zhang¹, Qian Zhang³

¹ Department of Computer Science and Technology,
Harbin Institute of Technology (Shenzhen)

² Guangzhou HKUST Fok Ying Tung Research Institute,
The Hong Kong University of Science and Technology

³ Department of Computer Science and Engineering,
The Hong Kong University of Science and Technology
{liaoqing,zoejiang,ckzhang}@hit.edu,
wangxuan@cs.hitsz.edu.cn,
{qzhang,dingye}@ust.hk

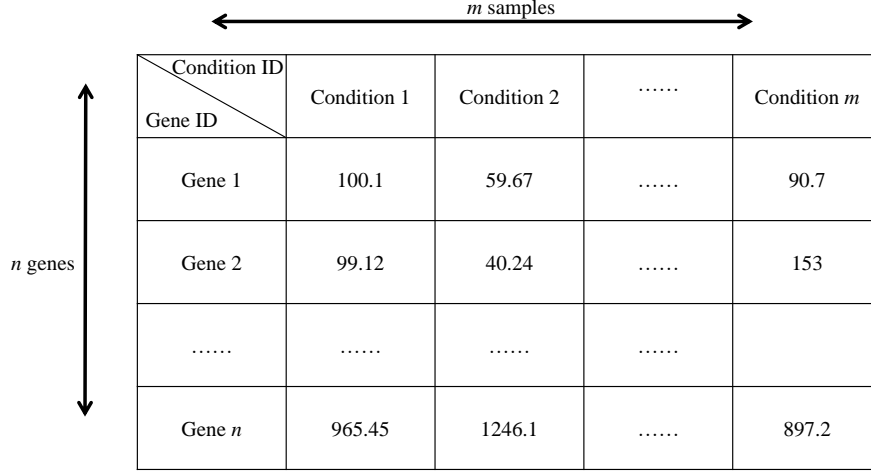
Abstract

Using computational techniques especially deep learning methods to facilitate and enhance cancer detection and diagnosis is a promising and important area. Nowadays, gene expression data has been widely used to train an effective deep neural network for precise cancer diagnosis. However, if a particular tumor has insufficient gene expressions, the trained deep neural networks may lead to a bad cancer diagnosis performance. In this paper, we propose a novel multi-task deep learning (MTDL) method to solve the data insufficiency problem. Since MTDL leverages the knowledge among the expression data of multiple cancers to learn a more stable representation for rare cancers, it can boost cancer diagnosis performance even if their expression data are inadequate. The experimental results show that MTDL significantly improves the performance of diagnosing every type of cancer when it learns from the aggregation of the expression data of twelve types of cancers.

Keywords: Multi-task learning, Deep learning, Cancer diagnosis

1. Introduction

Research on the correlation between gene expression profiles and cancer/diseases states plays an important role in biological and clinical tasks [1]. For instance, comparing genes expressed in diseased tissue and normal counterpart will advance our understanding in disease pathology, and also help to identify different type of tissues (cancerous and normal), because the gene expression data provides cues about the phenotype, function and physiological process of tissues. However, considering the amount and complexity of the gene expression data,



The diagram shows a matrix of gene expression data. A vertical double-headed arrow on the left is labeled "n genes", indicating the number of rows. A horizontal double-headed arrow at the top is labeled "m samples", indicating the number of columns. The matrix is a table with 5 rows and 5 columns. The first row is the header for conditions, and the first column is the header for genes. The data rows show expression values for Gene 1, Gene 2, an ellipsis row, and Gene n across Condition 1, Condition 2, an ellipsis column, and Condition m.

Condition ID	Condition 1	Condition 2	Condition m
Gene ID				
Gene 1	100.1	59.67	90.7
Gene 2	99.12	40.24	153
.....	
Gene n	965.45	1246.1	897.2

Figure 1: An example of gene expression data.

traditional biological experiments cannot handle such data. Figure 1 demonstrates an example of gene expression data. The gene expression data are usually organized in a matrix of n rows and m columns, where the rows represent features (genes) and the columns represent samples (for example, tissues, developmental stages and treatments). Since usually $n \gg m$, it is impossible for a biological expert to compute and compare the $n \times m$ gene expression matrix manually. Towards this end, a number of machine learning methods have been applied to analyze gene expression data and automatically classify tumors.

In the past decades, several machine learning methods have been applied to classify tissue into cancerous and normal by using microarray gene expression data. The earliest one is decision tree (DT) [2, 3] which conducted distinctive sequence features of known diseases proteins compared to all human proteins. The K-nearest neighbor (k-NN) classifier [4, 5] and naive Bayesian classifier (NB) [6] were also applied to identify human diseases by classifying multiple types of genomic data. Along this direction, Bharathi [7] applied Analysis of Variance (ANOVA) rank scheme on important genes and tested the classification capability by using Support Vector Machine (SVM). Hu *et al.* [8] proposed a Maximally Diversified Multiple Trees (MDMT) algorithm which ensembles a set of unique trees in the decision committee. Halder *et al.* [5] proposed a fuzzy k-nearest neighbor based active learning (ALFKNN) method which first applies unlabeled samples to get the labels from experts, then the labeled *informative sample* can be iteratively added to the training samples to improve the prediction accuracy. Begum *et al.* [9] incorporated AdaBoost and linear SVM (ADASVM) as a component classifier, and showed that it has higher performance than the state-of-the-art classifiers.

However, existing methods face two challenges which makes it difficult to

be directly applied to cancer diagnosis. The first challenge is *curse of dimensionality* [10] of the feature space in the gene expression data. For example, gene expression of thyroid cancer in [11] consists of 367 samples, each having about $O(n^4)$ features (genes). Hence, the high dimensionality of the feature space will increase the risk of overfitting because the number of genes are much larger than the number of samples. The second challenge is the insufficiency problem of tissue samples which makes all existing methods difficult to learn good representations of cancers, because the training data of cancers are often rare. The insufficiency problem is caused by two reasons: 1) the data samples of rare cancers (some only occurring for few people in each year) are much difficult to obtain than common cancers, and such insufficiency of cancer samples means that classifiers cannot be trained from stable representations of cancer patterns; 2) the data samples from different gene expression platforms are difficult to be integrate together, and such difficulty of integration aggravates the insufficiency problem. For example, Leukemia is one of the most common type of cancer in the world which has caused 353,500 deaths in 2015 [12]. Many research groups study Leukemia for the health of human beings. However, the data sources from different research groups cannot be integrated together, because of different experimental settings, and choose different gene features and even tag with different cancer labels. In our tested cancer datasets, there are two types of Leukemia data from two data sources. One Leukemia dataset [13] classifies Leukemia samples into two types (NPM1+ and NPM1-). Another Leukemia dataset [14] classifies samples into four types (MP, HDMTX, HDMTX+MP and LDMTX+MP). The first Leukemia dataset is classified based on gene point and second dataset is classified based on drug responses in human being's leukemia cells. Moreover, the first Leukemia dataset [13] has 54,675 gene features and the second Leukemia dataset [14] has 12,600 gene features because of different collection of gene features to investigate an identical cancer according to different background knowledges.

In this paper, we propose a novel Multitask Deep Learning (MTDL) method which can solve the insufficiency problem of tissue samples and reduce the bad influence of the high dimensionality problem of the feature space. MTDL is inspired by multi-task learning and deep learning. Specifically, MTDL can not only classify the small-scale datasets from different cancers simultaneously but also employ closely related datasets to help learning a better representation and boosting the classification performance. Our contributions are summarized as follows:

1. We propose a novel MTDL method which can directly solve the insufficiency problem of tissue samples and reduce the bad influence of the high dimensionality problem of the feature space.
2. MTDL can integrate various cancer datasets from different sources to enhance the classification performance, even if the types of tumors, features and labels are different.
3. MTDL can simultaneously utilize multiple cancer datasets so that hidden representations can provide more information to small-scale cancer

datasets and enhance the classification performance.

The remainder of this paper is organized as follows: Section 2 briefly reviews prior research works and related techniques. Section 3 illustrates the details of multi-task deep learning (MTDL) method. Section 4 evaluates the classification performance of MTDL and the representative algorithms using twelve real-world cancer datasets. Finally, the conclusions are given in Section 5.

2. Related Work

2.1. Dimensionality Problem of Feature Spaces

Reducing high dimensionality of feature space in gene expression is important for accurate classification of phenotype of samples, because tens of thousands of gene with only a small number of samples will increase the risk of overfitting. There are two approaches to solve the curse of dimensionality problem, feature extract (FE) and feature selection (FE). The former methods reduce the dimensionality by building new gene features from combinations (linear or nonlinear) of the original features and the later methods try to find a small subset of original features which can maintain the information content of the gene expression data. For example, Fakoor *et al.* [15] and Liu *et al.* [16] firstly utilized PCA to project gene expression data onto a low-dimensional subspace to relieve the imbalance influence between gene features and samples, and then utilized deep learning methods to enhance cancer diagnosis and classification performance. Guan and Tao [17, 18] built several nonnegative matrix factorization (NMF) models to efficiently reduce the dimensionality of the matrix feature and [19] proposed Gausse-seidel NMF (GSNMF) to relieve the impact of high dimensionality of gene feature space in the gene expression data. Moreover, Isabelle *et al.* [20] applied the support vector machine (SVM) to select gene features to improve cancer classification. Alexander *et al.* [21] applied mutual information (MI) method to select the top 10 genes and Chris Ding [22] proposed a minimum redundancy maximum relevance (MRMR) method which can use a smaller feature set to effectively cover the same space that a larger conventional feature set does. In a summary, above methods try to reduce the dimensionality of the feature space to avoid the overfitting problem of the classifier model. Different with dimensionality reduction methods, our method utilizes related gene expression datasets to increase the number of tissue samples to solve the overfitting problem.

2.2. Insufficiency Problem of Tissue Samples

Collecting adequate samples of cancer is expensive and time consuming, which leads to an insufficiency problem of tissue sample. Comparing with single-task learning, multi-task learning and transfer learning can use related cancer datasets to improve the classification performance of cancer tasks when the number of train samples is rather insufficient. Recently, transfer learning and multi-task learning techniques are proposed which can solve the problem in

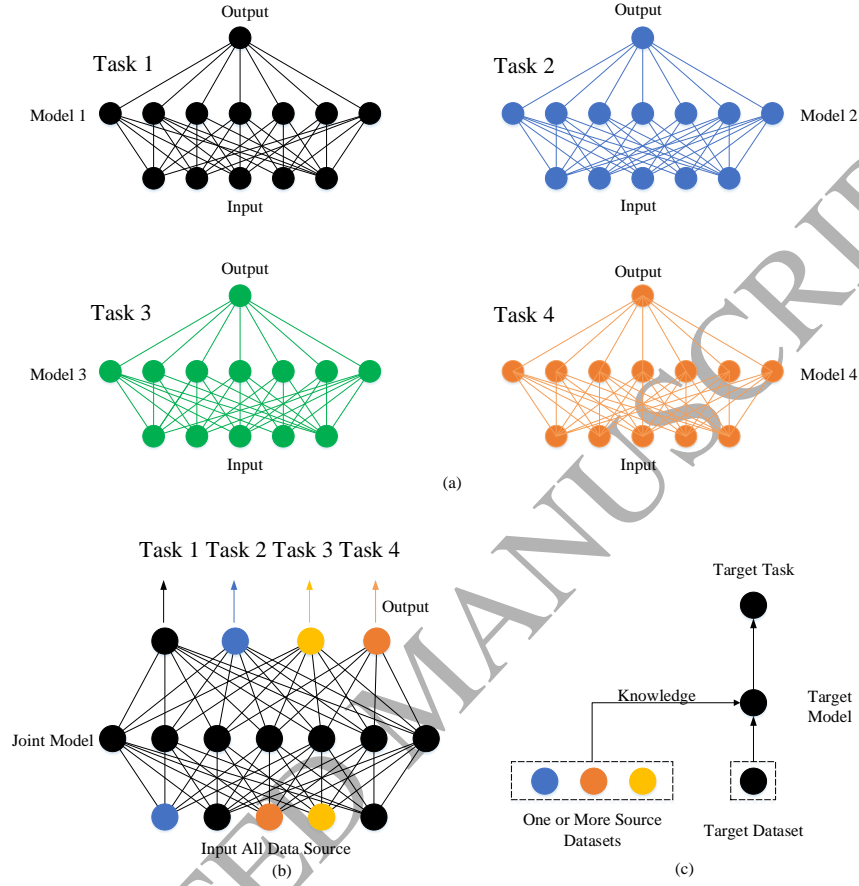


Figure 2: (a) Single-Task Learning, (b) Multi-Task Learning and (c) Transfer Learning

computer vision [23, 24, 25, 26, 27, 28, 29], bioinformatics [30, 31, 32], and climate analytics [33].

Figure 2 shows the differences between the learning processes of single-task, multi-task and transfer learning techniques. In the single-task learning (Figure 2 (a)), all four datasets have no connection and each task will be trained separately, because single-task learning assumes that the training samples are drawn independently from a particular distribution. Hence the single-task learning totally ignores the relationship among related tasks. In the multi-task learning (Figure 2 (b)), it assumes that tasks may be correlated which implies that the information learned from one task can be leveraged to another. Hence the multi-task learning learns a joint model of all tasks simultaneously. In the transfer learning (Figure 2(c)), it transfers the knowledge (parameters or representa-

tions) from source datasets to the target dataset to enhance the performance. Multi-task learning and transfer learning [34] have been thoroughly proven to improve the generalization performance significantly when there is not enough number of samples to train individual task [35].

As mentioned before, transfer learning extracts the knowledge from source tasks and transfer the knowledge to a target task [36], even if the training data and testing data have different domains, tasks and distributions. And the goal of transfer learning is it only requires good performance on the target task, because it more cares about the target dataset rather than source datasets. Recently, several transfer learning works have achieved successes in bioinformatics, especially for biological image analysis [37, 38, 39]. For example, Ravi K Samala *et al.* [37] developed a computer-aided detection (CAD) system which can use deep convolution neural network (DCNN) to transfer knowledge from mammograms digital images to digital breast tomosynthesis (DBT). Hariharan *et al.* [39] investigated the process of transferring a CNN, which uses ImageNet dataset as a trainset and medical images as a testset.

On the other hand, multi-task learning is close to transfer learning, which it tries to learn multiple datasets (tasks) simultaneously even all the datasets are different. But the goal of multi-task learning is it requires all the tasks have good experimental performance rather than only requires good performance on the target task, because all the datasets are important to the multi-task learning model. Multi-task learning is more suitable for our study because we try to achieve satisfactory classification performance on each cancer dataset. With the big success of deep learning technique in image processing [40, 41, 42, 43, 44] and pattern recognition [45, 46, 47, 48], more and more researchers incorporate multi-task learning and deep learning techniques together in computer vision [23, 24, 25, 31, 49, 50] and bioinformatics [30, 51, 38, 39] since these three years. In the computer vision field, Zhang *et al.* [30] proposed a tasks-constrained deep convolution network (TDCDN) model to jointly optimal facial landmark detection with multiple related tasks, e.g., head pose estimation task and facial attribute inference task. Similarity, Rejeev *et al.* [24] proposed a HyperFace architecture based on CNN which can simultaneously conduct face detection task, facial landmark localization task, head pose estimation task and gender recognition task from a given image. Abrar *et al.* [25] proposed a multi-task CNN model to better predict attributes in images, for example, whether wearing necktie or wearing a blue dress, by using deep CNN. In the bioinformatics field, Zhang *et al.* [30] proposed a deep model based on transfer learning and multi-task learning for biological image analysis on the domain-specific biological images. Ravi *et al.* [51] proposed a multi-task transfer learning DCNN which translates the knowledge from non-medical images to medical diagnostic tasks and simultaneously learning auxiliary tasks. Although above deep learning models based on multi-task learning show their successes in computer vision and biomedical image analysis in the past five years, all the existing methods learn representations and do multiple tasks separately. Specifically, these methods learn parameters via transfer learning method and learn representation via convolution neural networks (CNN) in the first stage. And then to use multi-task

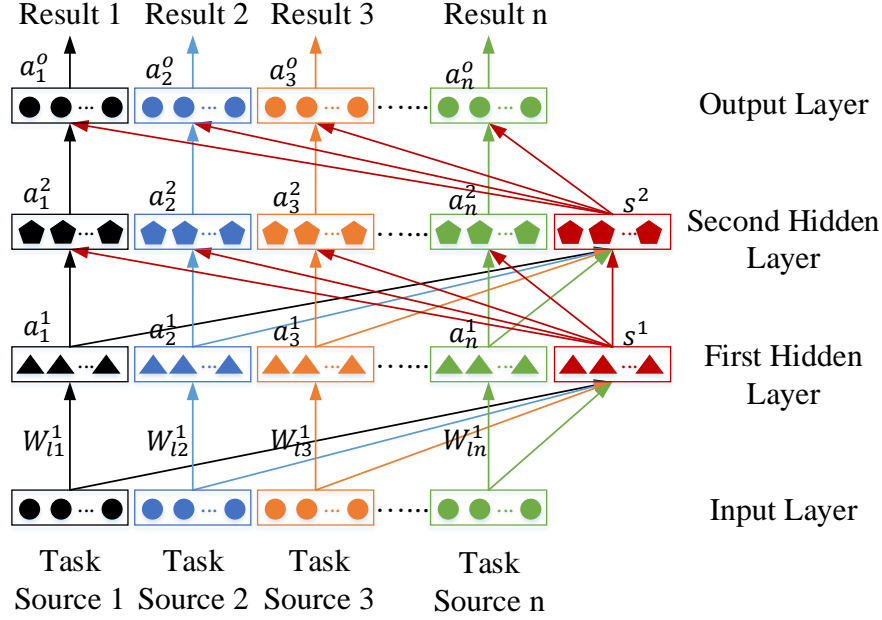


Figure 3: The proposed multi-task deep neural network structure. In the two hidden layers, the red units denote the shared hidden units and the units of the rest color denote the local hidden units.

learning technique to produce each result of task via learning a joint model from multiple datasets simultaneously. Different from all the existing multi-task deep learning works, we propose a novel multi-task deep learning (MTDL) algorithm which incorporates multi-task learning during deep learning. The advantages of MTDL lie in the following aspects: 1) it can learn more shared representations in each layers; and 2) it preserves the local representations of each dataset (task).

3. Multi-task Deep Learning

In this paper, we propose a novel multi-task deep learning (MTDL) algorithm for cancer classification. The structure of the network is shown in Figure 3. The proposed MTDL shares information across different tasks by setting a *shared hidden units*. In Figure 3, the red shapes signify shared hidden units of all the task sources in each layer, and the triangle, square and pentagon of rest colors signify local hidden units of each task source in different layers. In this work, we design two hidden layers and one soft-max output layer. We choose ReLU and sigmoid as activation functions in the hidden layers and the output layer, respectively. In machine learning, the non-saturating nonlinearity ReLU [52] is much faster than the saturating nonlinearities like tanh and sigmoid in

computing the gradient. Therefore, the convolutional neural networks (CNN) with ReLU unit can be trained several times faster than their alternatives such as that with tanh unit. We chose ReLU as activation function for fast training. The Sigmoid function is usually utilized to get labels in the output layer because it only outputs 0 or 1 and acts as a classifier. Hence, we chose the sigmoid function to represent the classification results of multiple tasks.

The model receives n groups of input units and each group corresponds to one task. In the first hidden layer, there are n groups of local hidden units correspond to n task source and a single group of shared units, which receives from all n groups of input units. Similar to the first hidden layer, the second hidden layer contains n groups of local hidden units and one group of shared units. In contrast to the first hidden layer, each group of local hidden units in the second hidden layer receives not only the activations of corresponding local hidden units in the first hidden layer, but also the activations of the shared units in the first hidden layer. The shared hidden units in the second hidden layer receive activations of the whole units in the first hidden layer, including local hidden units and shared hidden units.

Let x_1, x_2, \dots, x_n denote the inputs of n tasks. For the first hidden layer, the activations of each group of local hidden units a_i^1 are calculated by

$$a_i^1 = \sigma(W_{li}^1 x_i + b_i^1), i = 1, \dots, n \quad (1)$$

where the upscript of a_i^1 denotes the index of layer and the subscript of a_i^1 denotes the index of task source. Activation function is the rectified linear unit (ReLU), i.e., $\sigma(x) = \max(0, x)$, and W_{li}^1 is the local weight of edge between the #1 task source and the local hidden units a_i^1 . Moreover b_i^1 is the bias for the i -th group of local hidden units. The activations of the first shared hidden units s^1 are calculated by

$$s^1 = \sigma\left(\sum_{i=1}^n W_{si}^1 x_i + b_s^1\right), \quad (2)$$

where W_{si}^1 is the shared weight of edge between a_i^1 and s^1 . The b_s^1 is the bias of the shared hidden units s^1 , and the activation function is ReLU. For the second hidden layer, the activations of each group of local hidden units a_i^2 are calculated by

$$a_i^2 = \sigma(W_{li}^2 a_i^1 + W_{si}^2 s^1 + b_i^2), i = 1, \dots, n \quad (3)$$

where W_{li}^2 is local weight of edge between a_i^1 and a_i^2 . The W_{si}^2 is the share weight of edge between all the first local unit and the s^2 . The b_i^2 is the bias for the i -th group of local hidden units in the second layer. The activations of the shared hidden units s^2 are calculated by

$$s^2 = \sigma\left(\sum_{i=1}^n W_{si}^2 a_i^1 + W_s^2 s^1 + b_s^2\right), \quad (4)$$

Table 1: Summary of the gene expression datasets.

Cancer	Description	#Features	#Samples	Labels
Task 1	Acute Myeloid Leukemia [53]	54613	2341	1=AML, 2=MDS
Task 2	Adenocarcinoma [54]	34749	193	1=adenocarcinoma, 2=squamous cell carcinoma
Task 3	Breast Cancer [55]	30006	1047	1=non-IBC, 2=IBC
Task 4	Leukemia [13]	54675	2284	1=NPM1+, 2=NPM1-
Task 5	Leukemia [14]	12600	658	1=MP, 2=HDMTX, 3=HDMTX+MP, 4=LDMTX+MP
Task 6	Acute Myeloid Leukemia [56]	12625	625	1=Complete Remission, 2=Relapse
Task 7	Seminoma [57]	12625	618	1=state I, 2=state II and III
Task 8	Ovarian Cancer [58]	15154	153	1=cancer, 2=normal
Task 9	Colon Cancer [59]	2000	32	1=cancer, 2=non-cancer
Task 10	Medulloblastoma [60]	7129	30	1=class 0, 2=class 1
Task 11	Prostate Cancer [61]	12600	102	1=tumor, 2=normal
Task 12	Leukemia [62]	54613	2389	1=NPM1+, 2=NPM1-

where W_s^2 is the weight of edge between s^1 and s^2 . For the output layer, the output a_i^o of each task is calculated by

$$a_i^o = \text{sigmoid}(W_{li}^o a_i^2 + W_{si}^o s^2 + b_i^o), i = 1, \dots, n, \quad (5)$$

where W_{li}^o is the local weight of edge between a_i^2 and a_i^o . W_{si}^o is the weight of edge between s^2 and a_i^o . Moreover, b_i^o is the bias for the output units of the i -th task, and the activation function is defined as $\text{sigmoid}(x) = 1/(1 + e^{(-x)})$. The advantages of setting local units and the shared units are that each task can learn private representation from its local units perform classification. Different task learns separate private representation, because the local units preserve the feature of each separate task. On the other hand, the shared units learn shared representation from the whole datasets to leverage the information obtained from the microarray system. It is the shared units that boost the performance of each task because they leverage information through all tasks. In summary, MTDL can not only to preserve each tasks local features but also utilize the shared knowledge to provide stable features for all tasks. The following experiments confirm this point.

4. Experiments

4.1. Datasets

To demonstrate the feasibility and practicability of the proposed method, we have collected 12 different datasets from various papers/sources as summarized in Table 1.¹

In Table 1, there are two Acute Myeloid Leukemia datasets (task 1 and 6) collected from different sources, but it is not possible to integrate task 1 and 6 to enlarge the Acute Myeloid Leukemia datasets and solve the insufficiency

¹All the data are available in <http://liaoqing.me>.

Table 2: The accuracies of classifying 12 cancers by using DNN on each task without the shared hidden layers.

Cancer	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	Mean	Std. Dev.	Median
Task 1	0.6996	0.6996	0.747	0.6996	0.6552	0.7552	0.697	0.6496	0.7552	0.6947	0.70527	0.03743	0.6996
Task 2	1	1	1	1	1	1	1	1	1	1	1	0	1
Task 3	1	1	1	1	1	1	1	1	1	1	1	0	1
Task 4	0.5817	0.5817	0.5817	0.5817	0.6234	0.5817	0.5817	0.5817	0.6298	0.5836	0.59087	0.0189	0.5817
Task 5	0.2335	0.2502	0.2335	0.2169	0.2501	0.2335	0.2501	0.2335	0.2169	0.2169	0.23351	0.01357	0.2335
Task 6	0.9	0.9	0.9	0.9	0.8	0.9	0.9	0.9	0.9	0.8	0.88	0.04216	0.9
Task 7	0.7335	0.7668	0.8001	0.8001	0.7001	0.6335	0.6001	0.7001	0.6001	0.7668	0.71012	0.07706	0.7168
Task 8	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0	0.68
Task 9	0.9667	1	1	1	1	0.9667	0.9667	1	1	1	0.99001	0.01609	1
Task 10	0.9	0.9	0.8667	0.8667	0.8001	0.9333	0.9333	0.9333	0.9	0.9333	0.89667	0.04285	0.9
Task 11	1	1	1	1	1	1	1	1	1	1	1	0	1
Task 12	0.6871	0.6871	0.6871	0.6871	0.6871	0.6871	0.6871	0.6871	0.7176	0.6828	0.68972	0.00989	0.6871

problem in cancer diagnosis. It is because 1) task 1 and 6 contain different labels based on different domain knowledges, where labels *AML* and *MDS* in task 1 represent different stages of the Acute Myeloid Leukemia disease, but labels *Complete Remission* and *Relapse* in task 6 represent different symptoms after initial treatment, which are different from task 1; and 2) Task 1 and 6 contain different types of gene features from different data sources, where there are 54,613 gene features collected in task 1 [53], but there are 12,625 gene features collected in task 6 [56]. Hence, the insufficiency problem still exists in common cancer diseases, because different research groups choose different labels and gene features to investigate samples even for a same cancer.

As elaborated in Section 2.2 it is difficult for traditional methods to solve the insufficiency problem in high-performance cancer diagnosis. In this paper, the multi-task deep learning method utilizes all the datasets simultaneously, which can preserve the local representation of each dataset, and learn a shared representation from all datasets to enhance the classification accuracies of all tasks.

4.2. Classification Performance

We evaluate the effectiveness of MTDL by comparing it with two traditional deep learning methods, i.e., deep neural network (DNN) and sparse auto-encoder. We utilize 10-fold leave-one-out cross-validation to evaluate DNN, i.e., we firstly divide each dataset into ten folds, wherein nine folds for training and one fold for testing. To eliminate the influence of random partition, we repeat such trial ten times and output the averaged accuracy as the final result. The biggest difference between the proposed method and two representative methods is that the proposed method utilizes the shared knowledge from multiple datasets to learn more useful representations but the representative methods learn local representation from each individual dataset. In our experiment, the proposed method can learn on the 12 datasets simultaneously, but the representative methods learn on each dataset separately.

Table 2 gives the classification accuracies of 12 cancers by a traditional DNN on the gene expression data of each cancer tissue. As mentioned before, the DNN model receives 12 cancer datasets separately and output classification results one by one. Since traditional DNN model ignores the similarity information between

Table 3: The accuracies of classifying 12 cancers by using sparse auto-encoder on each task without the shared hidden layers.

Cancer	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	Mean	Std. Dev.	Median
Task 1	0.6996	0.6996	0.747	0.6996	0.6552	0.7552	0.697	0.6496	0.7552	0.6947	0.70527	0.03743	0.6996
Task 2	1	1	1	1	1	1	1	1	1	1	1	0	1
Task 3	1	1	1	1	1	1	1	1	1	1	1	0	1
Task 4	0.5817	0.5817	0.5817	0.5817	0.6234	0.5817	0.5817	0.5817	0.6298	0.5836	0.59087	0.0189	0.5817
Task 5	0.2335	0.2502	0.2335	0.2169	0.2501	0.2335	0.2501	0.2335	0.2169	0.2169	0.23351	0.01357	0.2335
Task 6	0.9	0.9	0.9	0.9	0.8	0.9	0.9	0.9	0.9	0.8	0.88	0.04216	0.9
Task 7	0.7335	0.7668	0.8001	0.8001	0.7001	0.6335	0.6001	0.7001	0.6001	0.7668	0.71012	0.07706	0.7168
Task 8	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0.68	0	0.68
Task 9	0.9667	1	1	1	1	0.9667	0.9667	1	1	1	0.99001	0.01609	1
Task 10	0.9	0.9	0.8667	0.8667	0.8001	0.9333	0.9333	0.9333	0.9	0.9333	0.89667	0.04285	0.9
Task 11	1	1	1	1	1	1	1	1	1	1	1	0	1
Task 12	0.6871	0.6871	0.6871	0.6871	0.6871	0.6871	0.6871	0.6871	0.7176	0.6828	0.68972	0.00989	0.6871

Table 4: The accuracies of classifying 12 cancers at the same time by using MTDL with the shared hidden layers.

Cancer	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	Mean	Std. Dev.	Median
Task 1	1	1	0.9526	1	1	1	1	1	0.95	1	0.99026	0.02054	1
Task 2	1	1	1	1	1	1	1	1	1	1	1	0	1
Task 3	1	1	1	0.9	1	0.95	1	1	1	1	0.985	0.03375	1
Task 4	1	0.9462	0.923	1	0.9462	1	1	1	0.95	1	0.97654	0.03112	1
Task 5	1	0.9333	0.9333	0.8666	1	0.8666	0.95	1	1	0.9333	0.94831	0.05241	0.94165
Task 6	1	1	1	1	1	1	1	1	1	1	1	0	1
Task 7	1	1	1	1	0.95	1	1	1	1	1	0.995	0.01581	1
Task 8	1	1	0.88	0.94	1	0.96	0.96	1	0.99	1	0.973	0.03945	0.995
Task 9	1	0.9667	1	1	1	1	1	1	1	1	0.99667	0.01053	1
Task 10	1	0.9667	1	0.9333	1	1	1	1	1	1	0.99	0.02250	1
Task 11	1	1	1	1	1	1	1	0.9667	0.9333	1	0.99	0.02250	1
Task 12	1	1	0.9696	0.9652	0.9392	1	0.9696	0.9696	1	0.9652	0.97784	0.02103	0.9696

related cancer datasets and the classification performance will be rather poor if the cancer samples are insufficient. We can find the DNN has very high accuracy results in Task 2, 3, 9, 10 and 11, because there are only two labels in these cancer datasets and these samples are very easy to assign to each labels. By contrary, Leukemia datasets (Task 4, 5 and 12) do not achieve good classification performances comparing with other cancer datasets. There are two reasons: 1) it is difficult to accurately diagnose Leukemia because the pattern of each Leukemia label are ambiguous; and 2) Task 5 has more labels, i.e., four labels, than other tasks so that its' accuracy is much lower than other tasks.

The accuracy results of 12 cancers by sparse auto-encoder of each cancer dataset in Table 3 are similar because the sparse auto-encoder cannot yet to use shared knowledge to improve the classification performance. Table 4 gives the classification accuracies of twelve cancers by MTDL. Unlike DNN and sparse auto-encoder, MTDL processes multiple cancer datasets (tasks) simultaneously so that MTDL can fully take advantages of similar hidden information between all the datasets (cancers) to improve each tasks performance. Table 4 shows that the accuracies of Leukemia datasets (Task 4, 5 and 12) have great improvements (more than 20%) compared with DNN and sparse auto-encoder, because MTDL can utilize these datasets simultaneously to reduce the bad influence of insufficiency problem in the Leukemia dataset. Moreover, MTDL can also achieve more than 20% improvement in both Task 7 and Task 8, because other cancer datasets provide more representative information via shared layer to help Task

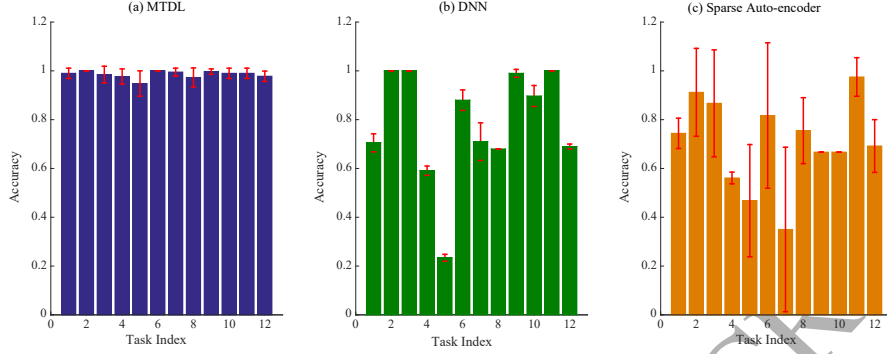


Figure 4: The accuracies of classifying 12 cancers by using MTDL, DNN, and sparse auto-encoder.

7 and Task 8 to learn a better representation and enhance the classification result. At last, MTDL has satisfactory performance on the rest cancer datasets, some datasets have clean pattern to classify so that the classification result of all the three methods are good. However, MTDL still has the highest performance in most datasets because MTDL simultaneously utilizes all the cancer datasets and learning shared representations through the shared layers can enhance the classification performance of most tasks.

In Figure 4, we demonstrate accuracy results of all the comparing methods on 12 datasets and the red line in each subfigure shows the standard deviation value of accuracy in each task. We can find that our method not only has the best performance in most datasets but also has more stable performance, because the standard deviation values of our method are much smaller than the compared methods.

5. Conclusion

Gene expression data plays an important role in precision medicine for cancer treatment, which simultaneously measures the expression levels of thousands of genes. However, for cancer classification, the gene expression data of a particular cancer might be limited. We propose a novel multi-task deep learning method (MTDL) to classify multiple cancers simultaneously and enhance the classification performance of each cancer by leveraging the knowledge through shared layers. MTDL method can process multiple datasets without pre-integrate at the same time even if each dataset has different class labels, features and samples. With the help of knowledge transfer, the classification accuracies of twelve cancers with few samples per cancer are significantly improved. In the future, MTDL method will be further extended to different types of datasets, including gene expression profiles, medical images, etc.

Acknowledgments

This work is supported in part by National Natural Science Foundation of China under grant No.61702134 and No.U1711261.

Reference

References

- [1] A. C. Tan, D. Gilbert, Ensemble machine learning on gene expression data for cancer classification, *Applied Bioinformatics* 2 (3 Suppl) (2003) S75.
- [2] N. Lpezbigas, C. A. Ouzounis, Genome-wide identification of genes likely to be involved in human genetic disease, *Nucleic Acids Research* 32 (10) (2004) 3108.
- [3] M. B. A. Snousy, H. M. El-Deeb, K. Badran, I. A. A. Khlil, M. B. A. Snousy, H. M. El-Deeb, K. Badran, I. A. A. Khlil, Suite of decision tree-based classification algorithms on cancer gene expression data, *Egyptian Informatics Journal* 12 (2) (2011) 73–82.
- [4] J. M. Keller, M. R. Gray, J. A. Givens, A fuzzy k-nearest neighbor algorithm, *IEEE Transactions on Systems Man and Cybernetics SMC-15* (4) (2012) 580–585.
- [5] A. Halder, S. Dey, A. Kumar, *Active Learning Using Fuzzy k-NN for Cancer Classification from Microarray Gene Expression Data*, Springer India, 2015.
- [6] P. Helman, R. Veroff, S. R. Atlas, C. Willman, A bayesian network classification methodology for gene expression data., *Journal of Computational Biology A Journal of Computational Molecular Cell Biology* 11 (4) (2004) 581–615.
- [7] A. Bharathi, A. M. Natarajan, Cancer classification of bioinformatics data using anova 2 (3) (2010) 369–373.
- [8] H. Hu, J. Li, H. Wang, G. Daggard, M. Shi, A maximally diversified multiple decision tree algorithm for microarray data classification, 2006, pp. 35–38.
- [9] S. Begum, D. Chakraborty, R. Sarkar, Cancer classification from gene expression based microarray data using svm ensemble, in: *International Conference on Condition Assessment Techniques in Electrical Systems*, 2016, pp. 13–16.
- [10] R. Bellman, R. E. Bellman, *Adaptive control processes: a guided tour*, Vol. 4, Princeton university press, 1961.

- [11] E. K. Alexander, G. C. Kennedy, Z. W. Baloch, E. S. Cibas, D. Chudova, J. Diggans, L. Friedman, R. T. Kloos, V. A. Livolsi, S. J. Mandel, Pre-operative diagnosis of benign thyroid nodules with indeterminate cytology, *Endocrinology* 13 (5) (2012) 705–715.
- [12] T. Vos, C. Allen, M. K. Arora, R. M. Barber, Z. A. Bhutta, A. C. Brown, Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015, *The Lancet* 388 (10053) (2016) 1545–1602.
- [13] H. U. Klein, C. Ruckert, A. Kohlmann, L. Bullinger, C. Thiede, T. Haferlach, M. Dugas, Quantitative comparison of microarray experiments with published leukemia related gene expression signatures., *BMC Bioinformatics* 10 (1) (2009) 1–11.
- [14] C. M. Y. Wl, C. H. Pui, J. R. Downing, C. Cheng, C. W. Naeve, M. V. Relling, W. E. Evans, Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells, *Nature Genetics* 34 (1) (2003) 85–90.
- [15] R. Fakoor, F. Ladhak, A. Nazi, M. Huber, Using deep learning to enhance cancer diagnosis and classification, in: *The International Conference on Machine Learning*, 2013.
- [16] J. X. Liu, Y. Xu, C. H. Zheng, H. Kong, Z. H. Lai, Rnpca-based tumor classification using gene expression data, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 12 (4) (2015) 964–970.
- [17] N. Guan, D. Tao, Z. Luo, B. Yuan, Online nonnegative matrix factorization with robust stochastic approximation, *IEEE Transactions on Neural Networks and Learning Systems* 23 (7) (2012) 1087–1099.
- [18] N. Guan, D. Tao, Z. Luo, B. Yuan, Nnmf: an optimal gradient method for nonnegative matrix factorization, *IEEE Transactions on Signal Processing* 60 (6) (2012) 2882–2898.
- [19] Q. Liao, N. Guan, Z. Qian, Gauss-seidel based non-negative matrix factorization for gene expression clustering, in: *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 2364–2368.
- [20] I. Guyon, J. Weston, S. D. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, *Machine Learning* 46 (2002) 389–422.
- [21] S. Espezua, E. Villanueva, C. D. Maciel, A projection pursuit framework for supervised dimension reduction of high dimensional small sample datasets, *Neurocomputing* 149 (PB) (2015) 767–776.

- [22] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, in: *Proceedings of the Bioinformatics Conference*, 2003., 2003, pp. 523–528.
- [23] Z. Zhang, P. Luo, C. L. Chen, X. Tang, Facial landmark detection by deep multi-task learning, in: *European Conference on Computer Vision*, 2014, pp. 94–108.
- [24] R. Ranjan, V. M. Patel, R. Chellappa, Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition, *IEEE Transactions on Pattern Analysis & Machine Intelligence* PP (99) (2016) 1–1.
- [25] A. H. Abdulnabi, G. Wang, J. Lu, K. Jia, Multi-task cnn model for attribute prediction, *IEEE Transactions on Multimedia* 17 (11) (2016) 1949–1959.
- [26] Q. Yao, X. Jiang, M. Gong, X. You, Y. Liu, D. Xu, Efficient group learning with hypergraph partition in multi-task learning, *Pattern Recognition* (2012) 9–16.
- [27] Z. He, X. Li, X. You, D. Tao, Y. Y. Tang, Connected component model for multi-object tracking, *IEEE transactions on image processing* 25 (8) (2016) 3698–3711.
- [28] X. You, Q. Li, D. Tao, W. Ou, M. Gong, Local metric learning for exemplar-based object detection, *IEEE Transactions on Circuits and Systems for Video Technology* 24 (8) (2014) 1265–1276.
- [29] Z. He, S. Yi, Y.-M. Cheung, X. You, Y. Y. Tang, Robust object tracking via key patch sparse representation, *IEEE transactions on cybernetics* 47 (2) (2017) 354–364.
- [30] W. Zhang, R. Li, T. Zeng, Q. Sun, S. Kumar, J. Ye, S. Ji, Deep model based transfer and multi-task learning for biological image analysis, in: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1475–1484.
- [31] T. Zeng, S. Ji, Deep convolutional neural networks for multi-instance multi-task learning, in: *IEEE International Conference on Data Mining*, 2015, pp. 579–588.
- [32] C. Sotiriou, S.-Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, P. Martiat, S. B. Fox, A. L. Harris, E. T. Liu, Breast cancer classification and prognosis based on gene expression profiles from a population-based study, *Proceedings of the National Academy of Sciences* 100 (18) (2003) 10393–10398.

- [33] A. R. Gonalves, F. V. J. Zuben, A. Banerjee, Multi-task sparse structure learning with gaussian copula models, *Journal of Machine Learning Research* 17 (2016) 1–30.
- [34] T. Liu, D. Tao, M. Song, S. J. Maybank, Algorithm-dependent generalization bounds for multi-task learning, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2016) 1–1.
- [35] R. Caruana, Multitask learning, *Machine Learning*.
- [36] S. J. Pan, Q. Yang, A survey on transfer learning, *IEEE Transactions on Knowledge and Data Engineering* 22 (10) (2010) 1345–1359.
- [37] R. K. Samala, H. P. Chan, L. Hadjiiski, M. A. Helvie, J. Wei, K. Cha, Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography, *Medical Physics* 43 (12) (2016) 6654.
- [38] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, I. C. Chang, Deep learning of feature representation with multiple instance learning for medical image analysis, in: *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 1626–1630.
- [39] H. Ravishankar, P. Sudhakar, R. Venkataramani, S. Thiruvankadam, P. Anang, N. Babu, V. Vaidya, Understanding the mechanisms of deep transfer learning for medical images (2017) 188–196.
- [40] D. Tao, C. Hong, J. Yu, J. Wan, M. Wang, Multimodal deep autoencoder for human pose recovery, *IEEE Transactions on Image Processing* 24 (12) (2015) 5659–5670.
- [41] C. Hong, J. Yu, D. Tao, M. Wang, Image-based 3d human pose recovery by multi-view locality sensitive sparse retrieval 62 (6) (2015) 3742–3751.
- [42] X. You, W. Guo, S. Yu, K. Li, J. C. Principe, D. Tao, Kernel learning for dynamic texture synthesis, *IEEE Transactions on Image Processing* 25 (10) (2016) 4782–4795.
- [43] M. Tzelepi, A. Tefas, Deep convolutional learning for content based image retrieval, *Neurocomputing* 275.
- [44] X. You, R. Wang, D. Tao, Diverse expected gradient active learning for relative attributes., *IEEE Transactions on Image Processing* 23 (7) (2014) 3203–3217.
- [45] X. Wang, L. Gao, J. Song, X. Zhen, N. Sebe, H. T. Shen, Deep appearance and motion learning for egocentric activity recognition, *Neurocomputing*.
- [46] Z. He, X. You, Y. Y. Tang, Writer identification of chinese handwriting documents using hidden markov tree model, *Pattern Recognition* 41 (4) (2008) 1295–1307.

- [47] J. M. Dudik, J. L. Coyle, A. El-Jaroudi, Z. H. Mao, M. Sun, E. Sejdi, Deep learning for classification of normal swallows in adults, *Neurocomputing*.
- [48] S. Yi, Z. Lai, Z. He, Y. M. Cheung, Y. Liu, Joint sparse principal component analysis, *Pattern Recognition* 61 (2017) 524–536.
- [49] S. Yi, Z. He, Y. M. Cheung, W. S. Chen, Unified sparse subspace learning via self-contained regression, *IEEE Transactions on Circuits & Systems for Video Technology PP* (99) (2017) 1–1.
- [50] Q. Liu, X. Lu, Z. He, C. Zhang, W. S. Chen, Deep convolutional neural networks for thermal infrared object tracking, *Knowledge-Based Systems*.
- [51] R. K. Samala, H. P. Chan, L. M. Hadjiiski, M. A. Helvie, K. Cha, C. Richter, Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms, *Physics in Medicine & Biology* 62 (23).
- [52] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: *International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.
- [53] K. I. Mills, A. Kohlmann, P. M. Williams, L. Wiecek, W.-m. Liu, R. Li, W. Wei, D. T. Bowen, H. Loeffler, J. M. Hernandez, et al., Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of aml transformation of myelodysplastic syndrome, *Blood* 114 (5) (2009) 1063–1072.
- [54] T. Fujiwara, M. Hiramatsu, T. Isagawa, H. Ninomiya, K. Inamura, S. Ishikawa, M. Ushijima, M. Matsuura, M. H. Jones, M. Shimane, Ascl1-coexpression profiling but not single gene expression profiling defines lung adenocarcinomas of neuroendocrine nature with poor prognosis., *Lung Cancer* 75 (1) (2012) 119–125.
- [55] W. A. Woodward, S. Krishnamurthy, H. Yamauchi, R. Elzein, O. Dai, E. Kitadaï, S. Niwa, M. Cristofanilli, P. Vermeulen, L. Dirix, Genomic and expression analysis of microdissected inflammatory breast cancer, *Breast Cancer Research and Treatment* 138 (3) (2013) 761–72.
- [56] T. Yagi, A. Morimoto, M. Eguchi, S. Hibi, M. Sako, E. Ishii, S. Mizutani, S. Imashuku, M. Ohki, H. Ichikawa, Identification of a gene expression signature associated with pediatric aml prognosis, *Blood* 102 (5) (2003) 1849.
- [57] I. Gashaw, R. Grmmer, L. Klein-Hitpass, O. Dushaj, M. Bergmann, R. Brehm, R. Grobholz, S. Kliesch, T. P. Neuvians, K. W. Schmid, Gene signatures of testicular seminoma with emphasis on expression of ets variant gene 4, *Cellular and Molecular Life Sciences Cmls* 62 (19-20) (2005) 2359–2368.

- [58] E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, Use of proteomic patterns in serum to identify ovarian cancer., *Lancet* 359 (9306) (2002) 572–7.
- [59] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, A. J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proceedings of the National Academy of Sciences* 96 (12) (1999) 6745–6750.
- [60] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. Kim, L. C. Goumnerova, P. M. Black, C. Lau, et al., Prediction of central nervous system embryonal tumour outcome based on gene expression, *Nature* 415 (6870) (2002) 436–442.
- [61] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D’Amico, J. P. Richie, et al., Gene expression correlates of clinical prostate cancer behavior, *Cancer cell* 1 (2) (2002) 203–209.
- [62] R. G. Verhaak, B. J. Wouters, C. A. Erpelinck, S. Abbas, H. B. Beverloo, S. Lugthart, B. Löwenberg, R. Delwel, P. J. Valk, Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling, *haematologica* 94 (1) (2009) 131–134.

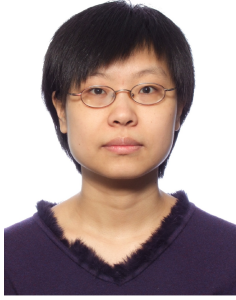


Qing Liao received her Ph.D. degree in computer science and engineering in 2016 supervised by Prof. Qian Zhang from the Department of Computer Science and Engineering of the Hong Kong University of Science and Technology. She is currently an assistant professor with School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), China. Her research interests include artificial intelligence and bioinformatics.



Ye Ding received his Ph.D. degree in computer science and engineering in 2014 supervised by Prof. Lionel M. Ni from the Department of Computer

Science and Engineering of the Hong Kong University of Science and Technology. He is currently a research associate in Fok Ying Tung Graduate School at The Hong Kong University of Science and Technology. His research interests are spatial-temporal data analytics and big data. He is a member of IEEE since 2013.



Zoe L. Jiang received the Ph.D. degree from the University of Hong Kong, Hong Kong, in 2010. She is currently an assistant professor with School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), China. Her research interests include computer vision and pattern recognition.

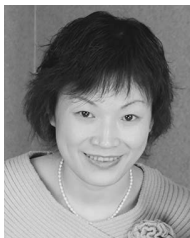


Xuan Wang received his M.S. and Ph.D. degrees in Computer Sciences from Harbin Institute of Technology in 1994 and 1997 respectively. He is a professor and Ph.D. supervisor in the Computer Application Research Center, School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), China. His main research interests include artificial intelligence, computer vision, computer network security and computational linguistics.



Chunkai Zhang is currently an associate professor with School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), China. His

research interests include computer vision and data mining.



Qian Zhang received the B.S., M.S., and Ph.D. degrees from Wuhan University, China, in 1994, 1996, and 1999, respectively, all in computer science. She joined Hong Kong University of Science and Technology in 2005 where she is now a full Professor in the Department of Computer Science and Engineering. Before that, she was in Microsoft Research Asia, Beijing, from July 1999.