Using deep learning to enhance cancer diagnosis and classification

Rasool Fakoor RASOOL.FAKOOR@MAVS.UTA.EDU
Faisal Ladhak FAISAL.LADHAK@MAVS.UTA.EDU
Azade Nazi AZADE.NAZI@MAVS.UTA.EDU
Manfred Huber HUBER@CSE.UTA.EDU
Computer Science and Engineering Dept, University of Texas at Arlington, Arlington, TX 76019 USA

Abstract

Using automated computer tools and in particular machine learning to facilitate and enhance medical analysis and diagnosis is a promising and important area. In this paper, we show that how unsupervised feature learning can be used for cancer detection and cancer type analysis from gene expression data. The main advantage of the proposed method over previous cancer detection approaches is the possibility of applying data from various types of cancer to automatically form features which help to enhance the detection and diagnosis of a specific one. The technique is here applied to the detection and classification of cancer types based on gene expression data. In this domain we show that the performance of this method is better than that of previous methods, therefore promising a more comprehensive and generic approach for cancer detection and diagnosis.

1. Introduction

Studying the correlation between gene expression profiles and disease states or stages of cells plays an important role in biological and clinical applications (Tan & Gilbert, 2003). The gene expression profiles can here be obtained from multiple tissue samples and by comparing the genes expressed in normal tissue with the ones in diseased tissue, one can obtain better insight into the disease pathology (Tan & Gilbert, 2003). One of the challenges that has been addressed in this way is to determine the difference between cancerous gene expression in tumor cells and the gene expression in normal, non-cancerous tissues. To address this, quite a number of machine learning classification techniques have been used to classify tissue into cancerous and normal. However, due to the high dimensionality of gene expression data (a.k.a the high dimensionality of the feature space) and the availability of only a few hundred samples for a given tumor, this application requires a number of specific considerations to deal with these data. The first challenge here is how to reduce the dimensionality of the feature space in a way that ensures that the resulting feature space still contains sufficient information to perform accurate classification. In addition, small sample sets (i.e. a small number of training examples) make the problem much harder to solve and increase the risk of overfitting. For years, many solutions have been proposed to address the cancer detection problem, most of which perform feature space reduction by deriving compact feature sets by selecting and constructing features either manually or in supervised ways. This, however, leads to the problems with those methods that they are mostly not scalable and can not be generalized to new cancer types without the re-design of new features. In addition, these techniques can not take effective advantage of tissue samples from other cancers when, for example, breast cancer detection is to be learned, being effectively restricted to only data from breast cancer and normal tissue when building the classifier. Given this restriction, in turn, likely leads to limitations in the way these methods scale to new cancer detection tasks when only a handful of samples are available.

To deal with this problem and to facilitate and develop more generalized versions of cancer classifiers, we propose in this paper a more general way of learning features by applying unsupervised feature learning and deep learning methods. We use a sparse autoencoder method to learn a concise feature representation from unlabeled data. In contrast to the previous methods where data has to be strictly from the cancer type to be detected in order to provide the appropriate label for supervised learning, the unlabeled data can here be

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

obtained by combing data from different tumor cells provided that they are generated using the same microarray platform (i.e. given that they contain the same gene expression information). For example, for the feature learning that forms the basis for prostate cancer classification we can use samples from breast cancer, lung cancer, and many other cancers which are available in that platform. The resulting features from all these sets are then used as a basis for the construction of the classifier.

The remainder of this paper is organized as follows: Section 2 provides some background about gene expression. Section 3 reviews prior research. Section 4 outlines the proposed method and Section 5 shows results of our method and compares them to results achieved using other methods. Finally, Section 6 concludes the paper.

2. Gene Expression

Gene expression data measures the level of activity of genes within a given tissue and thus provides information regarding the complex activities within the corresponding cells. This data is generally obtained by measuring the amount of messenger ribonucleic acid(mRNA) produced during transcription which, in turn, is a measure of how active or functional the corresponding gene is (Aluru, 2005). As cancer is associated with multiple genetic and regulatory aberrations in the cell, these should reflect in the gene expression data. To capture these abnormalities, microarrays, which permit the simultaneous measurement of expression levels of tens of thousands of genes, have been increasingly utilized to characterize the global gene-expression profiles of tumor cells and matched normal cells of the same origin. Specifically, microarrays are used to identify the differential expression of genes between two experiments, typically test vs. control, and to identify similarly expressed genes over multiple experiments. The processing pipeline of microarray data involves the raw data pre-processing to obtain a gene expression matrix, and then analyzing the matrix for differences and/or similarities of expression. The gene expression matrix, GEM, contains preprocessed expression values with genes in the rows, and experiments in the columns. Thus, each column corresponds to an array (or gene-chip) experiment, and could contain multiple experiments if there were replicates. Each row in the matrix represents a gene expression profile (Aluru, 2005).

Gene-chips can hold probes for tens of thousands of genes, whereas the number of experiments, limited by resources like time and money, is much smaller, at most in the hundreds. Thus, the gene expression matrix is typically very narrow (i.e. number of genes, n, is significantly larger than the number of experiments, m). This is known as the dimensionality curse and it is a serious problem in gene network inference (Aluru, 2005).

3. Related Work

Quite few methods have been proposed to detect cancer using gene expression data. In (C. Aliferis et al., 2003), Aliferis et al. used recursive feature elimination and univariate association filtering approaches to select a small subset of the gene expressions as a reduced feature set. Ramaswamy et al. in (Ramaswamy et al., 2001) applied recursive feature elimination using SVM to find similarly a small number of gene expressions to be used as the feature space for the classification. In (Wang et al., 2005b), Wang et al. showed that by combining a correlation-based feature selector with different classification approaches, it is possible to select relevant genes with high confidence and obtain good classification accuracy compared with other methods. Sharma et. al (Sharma et al., 2012) proposed a feature selection method aimed at finding an informative subset of gene expressions. In their method, genes were divided into small subsets and then informative genes in these smaller subsets were selected and then merged, ending up with an informative subset of genes. Nanni et. al in (Nanni et al., 2012) proposed a method for gene microarray classification that combines different feature reduction approaches. In most of those methods the focus was on how to learn features and reduce the dimensionality of the gene expression data. The majority of these methods use manually designed feature (e.g feature engineering) selectors to reduce the dimensionality of gene expression and select informative sets of genes. The potential problems with these feature selection methods are scalability and generality of features (i.e. whether the selected/designed features can be extended and applied to new classification tasks and data sets). In addition, since specific cancer data are usually rare and most of the mentioned methods can not efficiently take advantage of data from other cancers than the one to be detected or classified, these methods have to operate with very small data sets, limiting the effectiveness of the automatic feature learning approaches used. For example, prostate cancer data can not be used in selecting features for breast cancer detection, reducing the basis for feature learning. In contrast to these methods, our proposed method can use data from different cancer types in the feature learning step, promising the potential for effective feature

learning in the presence of very limited data sets.

4. Approach

Unsupervised feature learning methods and deep learning have been widely used for image and audio applications such as (Lee et al., 2009b; Huang et al., 2012), etc. In these domains, these techniques have shown a strong promise in automatically representing the feature space using unlabeled data in order to increase the accuracy of subsequent classification tasks. Using additional properties of the data, these capabilities have been further extended to facilitate learning in very high dimensional feature spaces. For example, by using image characteristics such as locality and stationary of images, Lee in (Lee et al., 2009a) proposed a method to scale the unsupervised feature learning and deep learning methods to high dimensional and full-sized images. Similarly, Le in (Le et al., 2012) applied an unsupervised feature learning method (in particular Reconstruction Independent Subspace Analysis) in the context of cancer detection by applying it in the classification of histological image signatures and classification of tumor architecture. However, to the best of our knowledge, unsupervised feature learning methods have not been applied to gene expression analysis (it should be noted that Le's method (Le et al., 2012) still has been applied to images not gene expression). Some of the reasons for this can be seen in the extremely high dimensionality of gene expression data, the lack of sufficient data samples, and the lack of global known characteristics such as locality in gene expression data which limit the applicability of techniques such as convolution or pooling which have been highly successful in the above-mentioned image data applications.

In the method proposed here, we try to address this dimensionality problem in the area of gene expression data. In our method, we first reduce the dimensionality of the feature space using PCA, and then apply the result of PCA as a compressed feature representation which still encodes the data available in the sample set, along with some randomly selected original gene expressions (i.e. original raw features) as a more compact feature space to either a one or a multi-layered sparse auto-encoder to find a sparse representation for data that will then be used for the classification. This overall approach to building and training a system to detect and classify cancer from gene expression data is shown in Figure 1. As shown in the figure, the approach proposed here consists of two parts, the feature learning phase and the classifier learning phase.

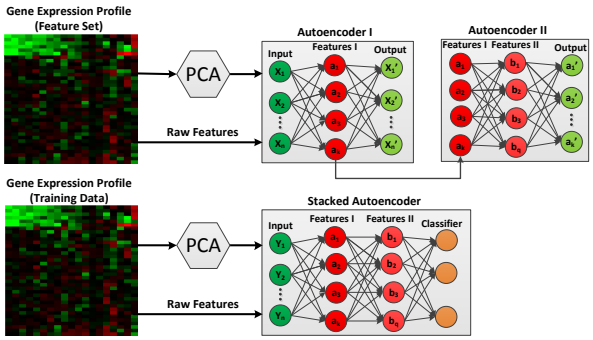

Figure 1. Overall Approach.

4.1. Feature Learning

Our proposed feature learning approach uses two phases, at first, PCA-based phase aimed at reducing the dimensionality of the feature space while maintaining the information content of the data. Second, based on an augmented form of the PCA features in addtion to some random raw features, develops a sparse encoding of the data samples to obtain high level and complex features for use by the classification approach. The main reason for this two-phase approach is that since the dimensionality of gene expression data is extremely high (on the order of 20000 to 50000 features) and these contain redundant and noisy data, we apply PCA to reduce the dimensionality of data without a significant loss of information. However, there is a problem with directly using the PCA components as features for classification. PCA performs a linear transformation on the data. In other words, after applying PCA, the resulting extracted features are simply a linear function of the original input data (Raina et al., 2007). However, in order to provide an opportunity to also capture the non-linearity of the relations between expressions of different genes, a different feature learning approach is needed. To facilitate this and to obtain more discriminating features, we use an unsupervised feature learning method in the second stage and, in order to provide it with the opportunity to capture additional non-linear relations that are hidden by the PCA features, we randomly add some of the original raw features to the PCA features to form an augmented basis for the second stage feature learning algorithms.

For the second phase of the feature learning appraoch, we use the framework of the sparse autoencoder (Coates et al., 2011; Bengio et al., 2007; Ng). The autoneocder neural network is an unsupervised feature learning method in which the input is used as the target for the output layer (Ng). In this way

it learns a function $h_{w,b}(x) \approx x$ that represents an approximation of the input data constructed from a limited number of feature activations represented by the hidden units of the network. The sparse autoencoder is constructed by three layers in the neural network (i.e input layer, hidden layer, and output layer) in which the hidden layer contains K nodes. The units in the hidden layer force the network to learn a representation of the input with only K hidden unit activations, representing K features. To train the network it uses the back-propagation method to minimize the squared reconstruction error with an additional sparsity penalty (Coates et al., 2011; Raina et al., 2007):

$$min_{b,a} \sum_{i}^{m} \parallel x_u^{(i)} - \sum_{j}^{K} a_j^{(i)} b_j \parallel_2^2 + \beta \parallel a^{(i)} \parallel_1 \qquad (1)$$

$$s.t. \ \parallel b_j \parallel_2 \leq 1, \forall j \in 1, .., s$$

where $x_u^{(i)}$ is unlabeled training example, b is basis vector, and a is the vector of activations of the basis (Raina et al., 2007). The sparsity penalty included in the form of the one-norm of the activation vector, a, here biases the learner towards features, b_j, that allow the data items to be represented using a combination of a small number of these features. The rationale for using a sparse encoding for feature learning in the gene regulation data is that features that allow for a sparse representation are more likely to encode discriminatory and functionally monolithic properties of the original data and thus are more likely to form a good basis for classification learning. Within the neural network, the sigmoid function has been selected as the activation function g:

$$g(z) = \frac{1}{1 + exp(-z)}$$

As an additional option and for further comparison we have used a stacked auto-encoder with two layers in which greedy layer-wise learning has been used to train the deep network (Bengio et al., 2007). In this greedy layer-wise approach , we separately train each network. However, the output of the first network functions is an input to the second network.

4.2. Classifier Learning

In order to perform the task of cancer detection and cancer type classification, the features learned in the proposed unsupervised feature learning approach are subsequently used with a set of labeled data for specific cancer types to learn a classifier. For the results in this paper we used softmax regression as the learning approach for the classifier.

For comparison in the experiments presented in this paper, a sparse autoencoder with one layer and one with two layers (aka. stacked autoencoder) has been used as the unsupervised feature learning method to learn a sparse representation from unlabeled data which then served as the input representation for classifier learning using the softmax regression classifier. In addition, we performed an additional experiment in which we used the fine-tuning method (Bengio et al., 2007) in order to tune the weights of the features of the stacked autoencoder to better match the requirements of the classification task. In this method, the weights of the features learned by the unsupervised feature learner are tuned through the classifier using labeled data. While this makes the features less generic by tuning them towards the specific classification task, it also promises the possibility of higher classification accuracy in some situations.

Overall, the method presented in this paper takes its strength from the combination of dimensionality reduction through PCA and unsupervised non-linear sparse feature learning for the construction of effective features for general classification tasks. This methodology allows for the effective use of unlabeled data, and thus of microarray data unrelated to the specific classification task, to assist in and improve the classification accuracy. As mentioned earlier, since the number of gene expression data samples for a specific cancer type is generally low, other cancer data from the same platform (i.e. with the same genes in the microarray) are a good candidate to be used in this method as unlabeled data for feature learning. One of the significant advantages of this approach as compared to most previous work is that it generalizes the feature sets across different types of cancers. For instance, data from prostate, lung, and other cancers can be used as unlabeled data for feature learning in a breast cancer detection or classification problem. The potential of this is further demonstrated by the results of Lu et. al in (Lu et al., 2007) who showed via comprehensive gene analyses that it is possible to find common cancer genes across different cancer data. This finding intensifies our argument for having generalized feature sets across data from various cancer types.

5. Results

To demonstrate the feasibility and applicability of the proposed method, we first obtained 13 different data sets from various papers/sources as summarized in Table 1. In Table 1, columns 2 and 3 show the dimensionality of the data. Column 3 shows labeled data which is used for training the classifier. For the fea-

ture learning, we use the unlabeled data in column 2. Some of the feature sets have been expanded to include samples of various different types of cancer from other data sets of the same microarray platform. This feature expansion provides the ability to the feature learning algorithm to learn more generalized features that are not specific to an individual cancer but rather reflect features of interest in general cancers.

Due to the high dimensionality of the data, the data was, as described in the approach section, preprocessed by applying PCA to reduce its dimensionality. Three different of sparse encoders have been used to learn features: sparse autoencoder which contains just one hidden layer, two layer stacked autoencoder, and stacked autoencoder with fine-tuning, which is trained based on a greedy layer-wise approach. The fine-tuning method uses labeled data to tune the parameters from the stacked autoencoder during the classifier training stage. Columns 2, 3, and 4 in Table 2 correspond to results of each of these methods, respectively.

To evaluate the robustness of the classifier, we performed *10-fold* cross-validation and results are presented in terms of the average classification accuracy. In addition, the standard deviation of the classification accuracy across the different learning trials is represented in the table. We also compare our proposed algorithm against two baselines: SVM with Gaussian kernel, and Softmax Regression. Note, that these methods also use the principal component projections as features to address the very high dimensionality and relatively small number of samples in the datasets. Table 2 reports the results of the data sets where only the result of the better of the two baseline algorithms is reported. From this we can see that the proposed method which uses the sparse autoencoder features derived from PCA and randomly selected raw features outperforms the baseline algorithms which do not use unsupervised sparse features. The only exceptions are the second and ninth data sets, for which our method does not outperform the baseline algorithms. We believe that we can actually improve those results by adding more unlabeled data to the feature sets. We were unable to do so because the platforms of the data sets were either a very specialized micro-array for which there was not a lot of samples available, or we could not find data from the same platform.

6. Conclusion

In this paper, we propose a method to enhance cancer diagnosis and classification from gene expression data using unsupervised and deep leaning methods. The proposed method, which uses PCA to address the very high dimensionality of the initial raw feature space followed by sparse feature learning techniques to construct discriminative and sparse features for the final classification step, provides the potential to overcome problems of traditional approaches with feature dimensionality as well as very limited size data sets. It does this by allowing data from different cancers and other tissue samples to be used during feature learning independently of their applicability to the final classification task. Applying this method to cancer data and comparing it to baseline algorithms, our method not only shows that it can be used to improve the accuracy in cancer classification problems, but also demonstrates that it provides a more general and scalable approach to deal with gene expression data across different cancer types.

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. of the National Academy of Sciences of the United States of America*, 96:6745–6750, Jun. 1999.

Aluru, S. *Handbook of computational molecular biology*, volume 9. Chapman & Hall/CRC, 2005.

Bengio, Yoshua, Lamblin, Pascal, Popovici, Dan, and Larochelle, Hugo. Greedy layer-wise training of deep networks. pp. 153–160, 2007.

C. Aliferis, I. Tsamardinos, P. Massion N. Fananapazir D. Hardin, A. Statnikov, N. Fananapazir, and Hardin, D. Machine learning models for classification of lung cancer and selection of genomic markers using array gene expression data. In *FLAIRS Conf.*, 2003.

Cheok, M. H., Yang, W., Pui, C. H., Downing, J. R., Cheng, C., Naeve, C. W., Relling, M. V., and Evans, W. E. Treatment-specific changes in gene expression discriminate in vivo drug response in human leukemia cells. *Nat Genet*, 34:85–90, May 2003.

Coates, A., Lee, H., and Ng, A. Y. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 2011.

Fujiwara, T., Hiramatsu, M., Isagawa, T., Ninomiya, H., Inamura, K., Ishikawa, S., Ushijima, M., Matsuura, M., Jones, M.H., Shimane, M., Nomura, H., Ishikawa, Y., and Aburatani, H. ASCL1-coexpression profiling but not single gene expression

Table 1. Dataset

ID	Dataset	Feature Matrix	Data Matrix	Data Labels
1	AML (Mills et al., 2009)	54613×2341	54613×183	1=AML, 2=MDS
2	Adenocarcinoma (Fujiwara et al., 2011)	34749×193	34749×28	1=adenocarcinoma , 2=squamous cell carcinoma
3	Breast Cancer (Woodward et al., 2013)	30006×1047	30006×1047	1=non-IBC, 2=IBC
4	Leukemia (Klein et al., 2009)	54675×2284	54675×125	1=NPM1+, 2=NPM1-
5	Leukemia (Cheok et al., 2003)	12600×658	12600×60	1=MP,2=HDMTX, 3=HDMTX+MP, 4=LDMTX+MP
6	AML (Yagi et al., 2003)	12625×625	12625×27	1=Complete Remission, 2=Relapse
7	Breast Cancer (Wang et al., 2005a)	22277×2301	22277×143	1=ER+, 2=ER-
8	Seminoma (Gashaw et al., 2005)	12625×618	12625×20	1=stage I, 2=state II and III
9	Ovarian Cancer (Petricoin et al., 2002)	15154×153	15154×100	1=cancer, 2=normal
10	Colon Cancer (Alon et al., 1999)	2000×32	2000×30	1=cancer, 2=non-cancer
11	Medulloblastoma (Pomeroy et al., 2002)	7129×30	7129×30	1=class0, 2=class1
12	Prostate Cancer (Singh et al., 2002)	12600×102	12600×34	1=tumor, 2=normal
13	Leukemia (Verhaak et al., 2009)	54613×2389	54613×230	1=NPM1+, 2=NPM1-

Table 2. Results

ID	Sparse Autoencoder	Stacked Autoencoder	Stacked Autoencoder with Fine Tunning	PCA + Softmax / SVM (with Gaussian kernel)
1	74.36±0.062	51.35±0.019	**95.15±0.047**	94.04±0.03
2	91.67±0.18	87.5±0.16	87.5±0.16	**93.33±0.14**
3	**86.67±0.219**	63.33±0.204	83.33±0.272	85.0±0.241
4	56.09±0.024	56.09±0.024	**93.65±0.049**	92.95±0.09
5	**46.76±0.23**	33.71±0.038	33.71±0.038	46.33±0.18
6	**81.67±0.298**	55.0±0.137	55.0±0.137	73.33±0.196
7	**85.454±0.10**	73.48±0.02	73.48±0.020	84.07±0.069
8	35.0±0.337	56.67±0.161	**80.0±0.258**	76.67±0.251
9	75.45±0.135	55.03±0.0336	99.0±0.032	**100.0±0.0**
10	66.67±0.0	66.67±0.0	**83.33±0.176**	83.33±0.236
11	66.67±0.0	66.67±0.0	**76.67±0.225**	76.67±0.274
12	**97.5±0.079**	73.33±0.102	73.33±0.102	94.167±0.124
13	69.18±0.108	65.66±0.01	**91.26±0.055**	90.39±0.081

profiling defines lung adenocarcinomas of neuroendocrine nature with poor prognosis. *Lung Cancer*, Jul. 2011. ISSN 01695002.

Gashaw, I., Grümmer, R., Klein-Hitpass, L., Dushaj, O., Bergmann, M., Brehm, R., Grobholz, R., Kliesch, S., Neuvians, T., Schmid, K., Ostau, C., and Winterhager, E. Gene signatures of testicular seminoma with emphasis on expression of ets variant gene 4. *Cellular and Molecular Life Sciences*, 62 (19):2359–2368, Oct. 2005.

Huang, G. B., Lee, H., and Learned-Miller, E. Learning hierarchical representations for face verification with convolutional deep belief networks. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 2518–2525, 2012.

Klein, H.U., Ruckert, C., Kohlmann, A., Bullinger, L., Thiede, C., Haferlach, T., and Dugas, M. Quantitative comparison of microarray experiments with published leukemia related gene expression signatures. *BMC Bioinformatics*, 10, 2009.

Le, Q.V., Han, J., Gray, J.W., Spellman, P.T., Borowsky, A., and Parvin, B. Learning invariant features of tumor signatures. In *ISBI*, pp. 302–305. IEEE, 2012. ISBN 978-1-4577-1858-8.

Lee, H., Grosse, R., Ranganath, R., and Ng, A.Y. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proc. of the 26th Int'l Conf. on Machine Learning*, pp. 609–616, 2009a.

Lee, H., Largman, Y., Pham, P., and Ng, A.Y. Unsu-

pervised feature learning for audio classification using convolutional deep belief networks. In *Advances in Neural Information Processing Systems 22*, pp. 1096–1104. 2009b.

Lu, Y., Yi, Y., Liu, P., Wen, W., James, M., Wang, D., and You, M. Common human cancer genes discovered by integrated gene-expression analysis. *PLoS ONE*, pp. e1149, 11 2007.

Mills, K.I., Kohlmann, A., Williams, P.M., Wieczorek, L., Liu, W., Li, R., Wei, W., Bowen, D.T., Loeffler, H., Hernandez, J.M., Hofmann, W., and Haferlach, T. Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of aml transformation of myelodysplastic syndrome. *Blood*, 114:1063–72, 2009.

Nanni, L., Brahnam, S., and Lumini, A. Combining multiple approaches for gene microarray classification. *Bioinformatics*, 28:1151–1157, Apr. 2012. ISSN 1367-4803.

Ng, A.Y. Unsupervised feature learning and deep learning @ONLINE. URL http://ufldl.stanford.edu/.

Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., and Liotta, L. A. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359(9306):572–577, Feb 2002.

Pomeroy, S.L., Tamayo, P., Gaasenbeek, M., Sturla, L.M., Angelo, M., Mclaughlin, M.E., Kim, J.Y.H., Goumnerova, L.C., Black, P.M., Lau, C., Allen, J.C., Zagzag, D., Olson, J.M., Curran, T., Wetmore, C., Biegel, J. A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D.N., Mesirov, J.P., Lander, E.S., and Golub, T.R. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, Jan. 2002.

Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A.Y. Self-taught learning: transfer learning from unlabeled data. In *Proc. of the 24th Int'l Conf. on Machine learning*, pp. 759–766. ACM, 2007. ISBN 978-1-59593-793-3.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J.P., Poggio, T., Gerald, W., Loda, M., Lander, E.S., and Golub, T.R. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. of the National Academy of Sciences of the United States of America*, 98(26):15149–15154, Dec. 2001.

Sharma, A., Imoto, S., and Miyano, S. A top-r feature selection algorithm for microarray gene expression data. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 9:754–764, May 2012. ISSN 1545-5963.

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., and Richie, J.P. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209, Mar. 2002.

Tan, Aik C.C. and Gilbert, D. Ensemble machine learning on gene expression data for cancer classification. *Applied bioinformatics*, 2, 2003. ISSN 1175-5636.

Verhaak, R.G.W., Wouters, B.J., Erpelinck, C.A.J., Abbas, S., Beverloo, H.B., Lugthart, S., Lwenberg, B., Delwel, H.R., and Valk, P.J.M. Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica*, 94:131–134, Jan. 2009.

Wang, Y., Klijn, J.G.M., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J., Jatkoe, T., Berns, E.M.J.J., Atkins, D., and Foekens, J.A. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460):671–679, Feb. 2005a.

Wang, Y., Tetko, I.V., Hall, M.A., Frank, E., Facius, A., Mayer, K.F. X., and Mewes, H.W. Gene selection from microarray data for cancer classification-a machine learning approach. *Comput. Biol. Chem.*, 29(1):37–46, Feb. 2005b. ISSN 1476-9271.

Woodward, W.A., Krishnamurthy, S., Yamauchi, H., El-Zein, R., Ogura, D., Kitadai, E., Niwa, S., Cristofanilli, M., Vermeulen, P., Dirix, L., Viens, P., Laere, S., Bertucci, F., Reuben, J.M., and Ueno, N.T. Genomic and expression analysis of microdissected inflammatory breast cancer. *Breast Cancer Research and Treatment*, pp. 1–12, 2013. ISSN 0167-6806.

Yagi, T., Morimoto, A., Eguchi, M., Hibi, S., Sako, M., Ishii, E., Mizutani, S., Imashuku, S., Ohki, M., and Ichikawa, H. Identification of a gene expression signature associated with pediatric aml prognosis. *Blood*, 102:1849–56, 2003. ISSN 0006-4971.