A transfer-learning approach to feature extraction from cancer transcriptomes with deep autoencoders

Guillermo López-García*, José M. Jerez, Leonardo Franco, and Francisco J. Veredas

Universidad de Málaga, Departamento de Lenguajes y Ciencias de la Computación, ETSI Informática, Málaga (Spain)

Abstract. The diagnosis and prognosis of cancer are among the more challenging tasks that oncology medicine deals with. With the main aim of fitting the more appropriate treatments, current personalized medicine focuses on using data from heterogeneous sources to estimate the evolution of a given disease for the particular case of a certain patient. In recent years, next-generation sequencing data have boosted cancer prediction by supplying gene-expression information that has allowed diverse machine learning algorithms to supply valuable solutions to the problem of cancer subtype classification, which has surely contributed to better estimation of patient's response to diverse treatments. However, the efficacy of these models is seriously affected by the existing imbalance between the high dimensionality of the gene expression feature sets and the number of samples available for a particular cancer type, To counteract what is known as the curse of dimensionality, feature selection and extraction methods have been traditionally applied to reduce the number of input variables present in gene expression datasets. Although these techniques work by scaling down the input feature space, the prediction performance of traditional machine learning pipelines using these feature reduction strategies remains moderate. In this work, we propose the use of the Pan-Cancer dataset to pre-train deep autoencoder architectures on a subset composed of thousands of gene expression samples of very diverse tumor types. The resulting architectures are subsequently fine-tuned on a collection of specific breast cancer samples. This transfer-learning approach aims at combining supervised and unsupervised deep learning models with traditional machine learning classification algorithms to tackle the problem of breast tumor intrinsic-subtype classification. Our main goal is to investigate whether leveraging the information extracted from a large collection of gene expression data of diverse tumor types contributes to the extraction of useful latent features that ease solving a complex prediction task on a specific neoplasia.

Keywords: Next-generation sequencing; Deep Learning; Autocoders; Machine Learning; Transfer-learning; Predictive modelling

1 Introduction

Over the last decade, Next Generation Sequencing (NGS) techniques have transformed fields such as biochemistry, biology or medicine, generating an unprecedented vast

* Corresponding author (guilopgar@uma.es).

amount of data that is analyzed by the omics disciplines: genomics, transcriptomics, proteomics, metabolomics and epigenomics [1]. In particular, gene expression data analysis (transcriptomics) plays an increasingly important role in *P4 medicine*–which stands for predictive, preventive, personalized and participatory–, due to the advent of the high-throughput sequencing technology called RNA-Seq [2]. In areas such as oncology, gene expression data offers a new way of describing the molecular state of a patient. As cancer is considered to be a genetic disease, a gene expression sample from a patient–which describes the genetic changes responsible for the progression of the disease, such as the over-activity or the repression of genes–contains information of paramount importance for the prevention, diagnosis and treatment of this malignant disease.

Enormous potential exists for machine learning (ML) methods to analyze these data in order to solve many different cancer prediction tasks. In fact, numerous ML studies have been proposed to tackle the problem of cancer diagnosis and prediction using gene expression data [3, 4]. However, in clinical tasks such as cancer detection, the number (M) of available samples to solve a concrete problem is usually scarce (300-1K), while the number (N) of input features (genes or transcripts) is extremely large (10K-60K). This existing imbalance between both figures, seriously diminishes the performance of ML approaches when applied to gene expression data. To counteract the effects of what is known as the curse of dimensionality ($N \gg M$) [5], various traditional ML dimensionality reduction techniques, such as feature selection and extraction methods [6], have been applied to reduce the number of input variables. Although these techniques scale down the input feature space, the prediction performance of traditional ML methods remains moderate, as the features-samples imbalance problem is only partly solved. In this way, the reduced number of labeled samples used to train the ML models does not allow them to extract from the data the hidden patterns that contribute most to improve the performance of the predictive models.

With the aim of solving the problematic effects derived from the curse of dimensionality in a more effective way, a deep learning (DL) approach can be adopted. Nowadays, DL is the state-of-the-art technology in fields such as image recognition and natural language processing [7]. In particular, deep autoencoders (AEs) are specifically designed to exploit unlabeled data and learn high-level features, being widely employed as a feature extraction procedure [8]. In this work, we will apply different deep AE models to perform feature extraction on gene expression data, hence reducing the high number of initial features.

On the other hand, when having such a reduced number of samples, training a DL architecture from scratch would lead the model to serious over-fitting issues. Diverse strategies, such as data augmentation or transfer learning (TL) approaches are commonly used to prevent these issues. Namely, in a typical TL approach an initial DL model is pre-trained on a *base* dataset aimed at solving a *base* task. The pre-trained model is subsequently fine-tuned on a *target* dataset used to solve a *target* task, i.e. the final task (notice that *base* and *target* refer to different datasets and tasks). For the TL approach to work properly, the *base* dataset must contain a much greater number of samples than the *target* dataset. This technique has been successfully applied to many different domains, such as text classification, image processing or software error detection [9]. Here, we apply a TL approach to pre-train several deep AE models in an unsupervised manner using a large collection of unlabeled tumor samples. These pre-trained AE models are further fine-tuned on a smaller collection of labeled samples to solve a concrete supervised task for breast cancer (BRCA) subtype classification .

In fact, although the contributions of DL to cancer prediction using gene expression data are just starting to emerge and there are not yet numerous studies [10], [11], a few recent works have already successfully applied a TL strategy using AEs to solve different cancer classification tasks. In [12], the authors pre-trained a stacked sparse autoencoder (SSAE) using unlabeled samples from two different tumors, and then fine-tuned the architecture using labeled samples from a third tumor type to differentiate between normal and tumor samples. In [13], traditional ML classifiers were applied using the high-level features extracted by a SSAE model in order to separate samples from two distinct tumor types. However, the cancer prediction tasks tackled by these preliminary studies are very general and relatively simple, as they aimed at classifying gene expression samples into tumor or normal classes, or distinguishing between different tumor types, which are manageable task successfully tackled by traditional feature selection and ML methods. Furthermore, the number of samples used in these studies to pre-train the deep models could still be considered as scarce ($\sim$ 400-1.5K).

In this work, we use the Pan-Cancer dataset to pre-train deep AE architectures on 9K samples obtained from 32 different tumor types. The resulting architectures are then fine-tuned on a collection composed of $\sim$ 900 BRCA samples, aimed at solving a very specific cancer prediction task: breast tumor intrinsic subtypes classification. Our main goal is to investigate whether pre-training these DL models, by using a large collection of heterogeneous gene expression data from 32 distinct tumor types, contributes to the extraction of useful latent features that ease solving a complex cancer prediction task, such as the classification of BRCA subtypes. By means of a TL strategy, in this study we train and fine-tune different AE models and architectures to work as feature extractors, and use the extracted latent features as the input of three different ML classification algorithms that are analyzed in a comparative manner: logistic regression (LR), support vector machines (SVM) and shallow artificial neural networks (ANN). To evaluate the efficacy of the proposed TL approach, we compare the results obtained using the AE models with the performance achieved by those same three ML models when using four traditional dimensionality reduction methods: analysis of variance (ANOVA) feature selection, mutual information feature selection, chi-squared feature selection and principal component analysis (PCA).

The rest of the paper is organized as follows. Section 2 describes the gene expression datasets used within the analysis, as well as the different AE models, the transfer-learning strategy and the feature selection/extraction techniques used in combination with ML classifiers, which deal with the cancer prediction task being tackled in this study. The cross-validation strategy followed to assess the performance of the different approaches compared in this paper is also presented in that section. The results obtained from the analysis are given in Section 3 and, finally, some conclusions are provided in Section 4.

2 Materials and Methods

The work-flow of our TL approach is shown in Fig. 1, and the details of our method are discussed in the next subsections.

2.1 Gene expression data and feature pre-selection

Tha Pan-Cancer dataset from The Cancer Genome Atlas (TCGA) project was used in this study [14], accessed from the UCSC Xena data browser [15]. This dataset consists of $\sim 11K$ RNA-Seq gene expression samples from 33 different tumor types, which have been previously pre-processed to take into account batch effects, using $log_2(TPM+0.001)$ transformed RSEM values. The initial number of features (transcripts) was 60498, which is an intractable number for any ML model. In order to perform an initial reduction of the feature space, we applied an unsupervised feature selection strategy. Firstly, using the standard deviation (SD) measure, the variables with constant expression values across all the samples were removed. Then, according to the meadian absolute deviation (MAD), the $\sim 9K$ most variably expressed genes across the samples were selected, having a final dataset of 10535 samples and 9076 features.

2.2 Dataset split

Rather than using the whole Pan-Cancer dataset, we split the data into two distinct subsets: one of the subsets contains only the BRCA tumor samples (BRCA dataset, 1212 samples), whereas the other one includes the remaining samples from the rest of the 32 tumor types (non-BRCA dataset, 9323 samples). The rationale for this split is that the labeled information relating to the cancer prediction task tackled in this work is only contained in the BRCA tumor samples (see section 2.4). For that reason, following a TL approach, as it is described in Fig. 1, the non-BRCA dataset is used during the unsupervised pre-training phase, while BRCA data containing the available labeled information is used to perform the supervised fine-tuning of the models.

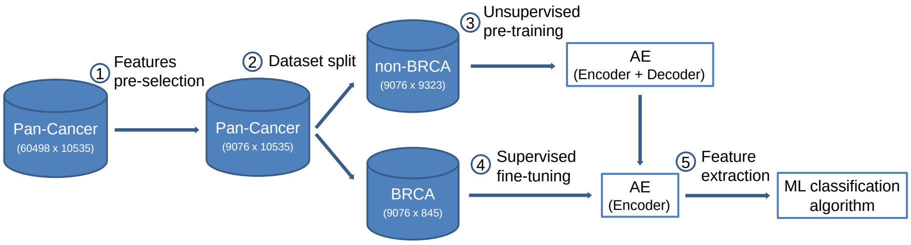

Fig. 1: A general overview of the TL strategy used in this work to perform BRCA instrinsic-subtypes classification.

2.3 Unsupervised pre-training of deep AE models

An AE, in its simplest (i.e. shallowest) form, is a feed-forward neural network with only three layers: an input, a hidden and an output layer (Fig. 2A). It is an unsupervised learning method for which the main aim is to reconstruct, at the output layer, a pattern given to the input layer, so that the reconstructed output pattern is as closely similar as possible to the original input pattern. This is done by training the network using the back-propagation algorithm to minimize the reconstruction error, a function that computes the difference between the input and the output vectors.

Given an input $x = \{x_1, x_2, ..., x_n\}$, an AE tries to learn a function $\hat{x} \approx x$ (Fig. 2A). The function that transforms the input into a hidden representation is called the encoder, and can be expressed as $h(x) = f(Wx + b)$, where f is the hidden activation function, W is the hidden weight matrix and b is the bias vector of the hidden layer. Given n the number of units in the input layer and k the number of hidden units, the matrix W is of dimensions $n \times k$. On the other hand, the function that takes the hidden representation and transforms it into the reconstructed input representation is called the decoder, and can be expressed as $\hat{x}(h) = g(W'h + b')$, where g is the output activation function, W' is the output weight matrix and b' is the bias vector of the output layer.

Having a hidden layer with fewer units than the input layer (i.e. $k < n$), forces the AE to compress the input vector into a lower dimensional representation, which can be reconstructed to its initial representation. In this case, the AE can be used as a dimensionality reduction method, in particular as a feature extraction procedure.

Constraining the network, such as using a small number of hidden units, has demonstrated to force the AE to extract more abstract and meaningful features in the hidden representations. In addition to reducing the number of hidden units, another popular way of constraining the model is using what is called a sparsity penalty [16]. This penalty creates sparse representations, in which hidden units tend to be inactive most of the time, favoring the units specialization. The sparsity constraint can be implemented using L1-regularization in the hidden layers, which is added to the reconstruction error function. In this way, for an input vector $x \in \Re^n$, if the mean squared reconstruction error as well as positive hidden activation functions are used, the overall loss function minimized during the learning procedure can be expressed as:

$$J_{sparse}(W, b) = \frac{1}{n} \sum_{i}^{n} (x_i - \hat{x}_i)^2 + \lambda \sum_{j}^{k} |h_j|$$

where n is the number of input and output units, k is the number of hidden units, h_j is the activation value of the j-th hidden unit and λ is the L1-regularizer penalty. The first term corresponds to the input reconstruction error, whereas the second term represents the L1-regularization, which tends to decrease the absolute values of the hidden activations towards zero, acting as a sparsity constraint.

Finally, another widely used approach to constrain the network is known as denoising AE [17]. During training, noise is added to the input data, and the difference between the input reconstruction and the original noiseless data is minimized using back-propagation. Thus, the goal of the network is to obtain a hidden representation robust to the introduction of noise at the input layer. In order to be able to reconstruct

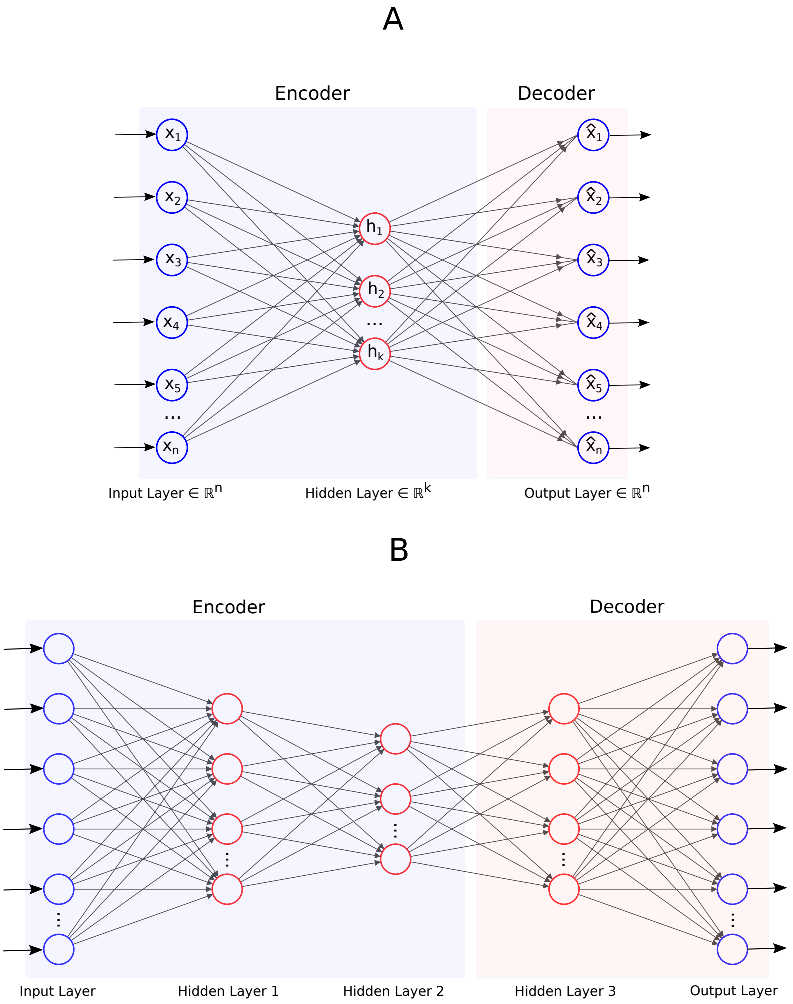

Fig. 2: Different AE architectures. **A** The architecture of a basic AE, where $\{x_1, x_2, ..., x_n\}$ are the units of the input layer, $\{h_1, h_2, ..., h_k\}$ are the hidden units and $\{\hat{x}_1, \hat{x}_2, ..., \hat{x}_n\}$ represent the output neurons. **B** The architecture of a deep AE with 3 hidden layers.

the input correctly, the corruption of the input data forces the network to extract more abstract and meaningful features in the hidden representation. This can be easily implemented using dropout in the input layer of the AE network [18].

In this work, with the purpose of extracting complex non-linear patterns from the high-dimensional gene-expression data, two different deep AE approaches have been implemented and analyzed: a deep sparse model with 3 hidden layers (see Fig. 2B), and a deep sparse denoising AE with 5 hidden layers. In both cases, the sparsity constraint has been implemented using an L1-regularization penalty for all the hidden layers in the encoder sub-networks. On the other hand, dropout has been used to introduce the noise necessary for the deep sparse denoising AE to work as expected. In addition, with the aim of reducing the initial number of features (9076) in an incremental way, the deep sparse AE uses $5K$ units in hidden layer 1 and 500 unit in hidden layer 2. For its part, the deep sparse denoising network counts on $5K$ nodes in hidden layer 1, $2K$ nodes in hidden layer 2 and 500 nodes in hidden layer 3. In terms of the number of units of each layer, both architectures are symmetric with respect to the central hidden layer, i.e. hidden layer 2 in case of the deep sparse AE and hidden layer 3 in the deep sparse denoising model. Additionally, for comparison reasons, we have also implemented a shallow sparse architecture (see Fig. 2A), which uses a single hidden layer of 500 units in which L1-regularization is employed as the sparsity constraint.

Finally, with the purpose of training the AEs using a large collection of unlabeled gene expression samples from 32 different tumor types, the models are pre-trained using the non-BRCA dataset in an unsupervised way. Before pre-training the AEs, the non-BRCA gene expression dataset is normalized using zero-one scaling. The activation function of the output layer of all three AE models is a sigmoid.

2.4 Supervised fine-tuning

Once pre-trained, the resulting AEs are fine-tuned using the BRCA dataset. The cancer prediction task tackled in this work is breast tumor instrinsic-subtypes classification. Hence, the variable to be predicted is the PAM50 intrinsic subtype, included among the clinical output variables contained in the BRCA samples from the Pan-Cancer dataset. PAM50 is a widely used 50-gene BRCA intrinsic subtype predictor [19], which groups the samples into four main subtypes: Luminal A, Luminal B, Basal-like and Her-2 enriched. From the 1212 samples contained in the BRCA dataset (see Section 2.2), we only select the samples for which PAM50 subtypes information is known, giving a final BRCA dataset composed of 845 labeled samples (see Fig. 1). Since each sample in this dataset is labeled with one of the 4 possible PAM50 subtypes, the classification task becomes a multi-class prediction problem with 4 different classes.

To perform the fine-tuning of the AEs using the BRCA samples labeled with the PAM50 subtype labels, the unsupervised AE models have to be transformed into supervised models. To do so, the decoder part of the networks (see Fig. 2) is replaced by a softmax output layer with 4 units (one for each BRCA intrinsic subtype). Finally, the resulting network architectures are fine-tuned in an supervised manner, using back-propagation to minimize the categorical cross-entropy loss function.

2.5 Autoencoders for feature extraction

After fine-tuning the models, we eliminate the softmax output layer from the AEs, so that only the encoder part of the networks remains available. In this way, the resulting fine-tuned encoders are used as feature extraction mechanisms. Thus the encoders work by propagating forward the high-dimensional patterns given as their input, so that they are transformed, layer by layer, to get a final latent representation with fewer number of variables than the original gene-expression data. These extracted features are then used as inputs that are fed into three different ML classifiers, namely LR, SVM and ANN, which are both trained in a supervised manner and evaluated using the PAM50 subtyes information.

2.6 Comparison to other dimensionality reduction methods

With the aim of evaluating the efficacy of the deep AEs as feature extraction methods, we compare them to other classical feature extraction and selection algorithms when they are used in combination with the same three ML supervised models (i.e. LR, SVM and shallow NN) to tackle the PAM50-subtypes prediction task. Namely, we compare the AE feature-extraction networks to three different feature selection methods, ANOVA, mutual information and chi-squared feature selection, as well as a feature extraction procedure, PCA. Like the encoders obtained from the pre-trained and fine-tuned AEs, these algorithms are also applied to reduce the number of features contained in the labeled BRCA dataset. Again, the selected/extracted variables given by these methods are used as the input for three ML classifiers (LR, SVM, ANN), which are trained and evaluated using the PAM50 intrinsic subtypes labels.

Note that, on the one hand, the main difference between the TL approach (feature extraction via AE + ML classifier) and the traditional ML pipeline (classical feature selection/extraction algorithm + ML classifier) used in this study is the strategy employed to reduce the dimensionality of the gene expression data, as the same classification algorithms are used by both methods to perform the PAM50 subtypes prediction task. On the other hand, while the TL strategy makes use of both the non-BRCA—for unsupervised learning—and the BRCA dataset—for supervised learning—, only the BRCA dataset is used in a supervised manner in the traditional ML pipeline followed in this work for comparison purposes.

2.7 Validation scheme

In this work, a 10-fold cross-validation (CV) scheme is used to estimate the predictive performance of each model using the labeled BRCA dataset. The average accuracy (ACC) calculated across the 10 test folds is used as the evaluation measure. Regarding the optimization of models' hyper-parameters, Random Search [20] with 20 iterations was performed using 5-fold CV within each of the 10 train folds, thus carrying out a nested CV procedure, using the inner 5-fold CV for model selection and the outer 10-fold CV for model evaluation. In case of the TL approach, both the fine-tuning hyper-parameters of the deep AE models (such as dropout, learning rate, momentum and number of epochs) and the hyper-parameters of the ML classifiers (such as kernel function, C and gamma for SVM and the hidden layer size, learning rate and momentum for ANN) are tuned, whereas in case of the traditional ML pipeline, only the hyper-parameters of the ML supervised models are optimized.

3 Results

Table 1 shows the average accuracy (ACC) and standard deviation from the 10-fold cross-validation obtained by each combination of feature selection/extraction method and ML model, when predicting PAM50 intrinsic subtypes. The rows in the table represent the different methods used to reduce the dimensionality of the gene-expression data, whereas the columns stand for the classification algorithms used to perform the PAM50-subtype prediction task. While the first four rows in Table 1 correspond to the classical feature selection/extraction procedures analyzed in this study, the last three represent the distinct AE architectures used within the TL approach for feature extraction proposed in this paper. Additionally, Fig. 3 contains a box-plot that depicts the 10-fold CV ACC test values distribution obtained by each ML classifier when using the selected/extracted features given by the different dimensionality reduction procedures.

Table 1: Average and standard deviation from 10-fold CV ACC test results.

Feature selection/ extraction method	**ML classifier**		
	LR	**SVM**	**ANN**
ANOVA	90.76 ± 3.03	90.99 ± 2.87	91.24 ± 3.37
Mutual Information	90.99 ± 1.94	91.35 ± 1.90	90.75 ± 1.74
Chi-Squared	88.07 ± 3.04	86.35 ± 3.86	86.96 ± 3.62
PCA	90.62 ± 2.71	90.50 ± 3.72	90.62 ± 3.37
Sparse AE	87.10 ± 2.84	88.08 ± 3.83	88.21 ± 4.53
Deep Sparse AE	88.31 ± 2.95	88.68 ± 3.10	89.91 ± 3.97
Deep Sparse Denosing AE	89.29 ± 4.13	89.88 ± 2.77	90.26 ± 2.85

In terms of the average test ACC, if we compare only the results obtained by the approaches using the different AE models as feature extraction methods, for all ML classifiers (i.e. LR, SVM and ANN), the deep sparse architecture performs better (88.31, 88.68 and 89.91, respectively) than the shallow sparse network (87.10, 88.08 and 88.21). Moreover, the deep sparse denoising AE obtains better performance rates for the three ML classifiers (89.29, 89.88 and 90.26) than the deep sparse model. Thus, we can conclude that the deeper the AE architecture is, the better results are achieved, showing the great potential of this sort of DL models to extract complex patterns from high-dimensional data useful for classification purposes. On the other hand, if we focus separately on the analysis of the predictive capacity of each of the three ML classification algorithms, when using the features extracted by the AEs (Sparse, Deep Sparse and Deep Sparse Denoising), SVM (88.08, 88.68 and 89.88) performs better than LR (87.10, 88.31 and 89.29), whereas shallow ANN (88.21, 89.91 and 90.26) obtains better results than SVM. Since the shallow ANN is a connectionist model, it takes more advantage of the features extracted by the AEs—which are also feed-forward NNs—to perform the final classification task.

However, when comparing the traditional ML pipeline with the TL approach, the classical feature selection/extraction algorithms contribute to undoubtedly better performance than the AE models do. In terms of ACC, ANOVA and Mutual Information feature selection, as well as PCA feature extraction, outperform any of the AE models, and

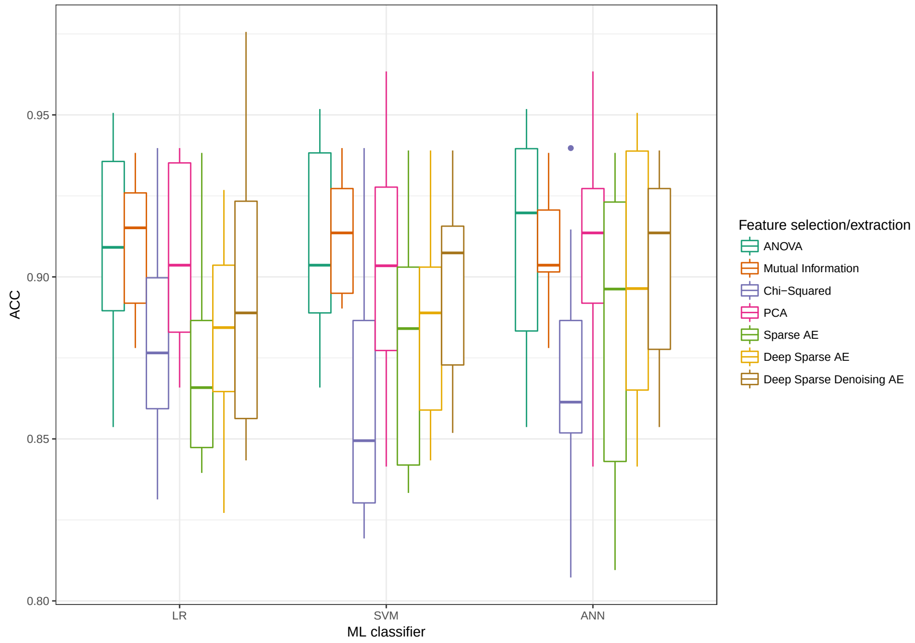

Fig. 3: Box-plot describing the 10-fold CV ACC test values distribution obtained by each combination of feature selection/extraction method and ML classification algorithm.

only the Chi-Squared feature selection procedure is surpassed by our TL approach with AEs. Among the traditional dimensionality reduction algorithms, ANOVA and Mutual Information contribute to the best predictive performances of the ML classifiers, and the highest average ACC (91.35) among all models is obtained when combining Mutual Information with SVM classifier.

The TL approach proposed in this work aims to apply deep AE models in combination with ML classification algorithms to tackle the problem of breast tumor intrinsic-subtypes classification using a scarce (845 samples) gene expression dataset. By pre-training the model with a large collection of unlabeled samples—from 32 tumors different from BRCA—and fine-tuning the resulting architecture using the reduced collection of BRCA samples, the deep AEs are able to make use of the knowledge extracted from data of other tumors to solve a particular cancer prediction task. However, the efficacy of this strategy has been shown to be limited. Thus, in terms of accuracy, the ML classifiers analyzed in this work achieve better predictive performance rates when they are preceded by classical dimensionality reduction algorithms as feature selection/extraction methods, in contrast to slightly lower efficacy rates given by the AE models used with this same purpose. This may be due to the fact that different types of cancer are actually different kinds of diseases, thus leveraging information from a large collection of gene expression samples from a wide variety of tumors does not contribute to a great extent to solve a complex cancer prediction task such as the BRCA intrinsic-subtypes classification.

4 Conclusions

In this paper, we have presented a TL approach that, in combination with diverse supervised ML algorithms, aims at tackling the problem of breast cancer intrinsic-subtype classification. This approach makes use of deep AE models to propose a solution to the adverse effects derived from the curse of dimensionality, that arises when dealing with gene expression data. The Pan-Cancer dataset has been employed to pre-train three different AE architectures on a heterogeneous dataset composed of thousands of gene expression samples obtained from tenths of different cancer types. Once pre-trained in a unsupervised manner, the resulting AEs have been fine-tuned in a supervised way by using a reduced dataset composed of hundreds of breast tumor labeled samples. The final purpose of this TL strategy was to reduce the dimensionality of the gene expression data by extracting valuable features to be subsequently used by ML classifiers to predict BRCA intrinsic-subtypes. Finally, with the aim of assessing the effectiveness of AE models as feature extraction mechanisms, we have analysed the contribution of three different AE architectures to the performance of several ML classifiers, and compared it to the efficacy achieved by these same ML models when preceded by four different traditional feature selection/extraction algorithms in a classical ML pipeline.

The results of the analysis showed that, on the one hand, the deep AE architectures extracted more useful features for classification purposes than the shallow AE model. On the other hand, the features selected/extracted by the traditional methods, led the ML classifiers to achieve slightly better predictive performance rates than the AE models. Hence, leveraging information from many cancer types does not seems to contribute to solve a more complex and specific cancer classification task such as prediction of breast tumor intrinsic subtypes. This findings support the hypothesis stating that different types of cancer are merely different types of diseases, all of them called cancer.

In future work, authors will continue to explore the adaptation of different DL approaches to be applied to the biomedical/bioinformatics domain, in particular to gene expression data. Special attention will be paid to model interpretability, as though many efforts have already been made in this particular field, most of DL models are still considered as "black-boxes". In areas such as oncology, if these algorithms aim to become a benchmark, interpretability must not be a lacking quality, but a main characteristic.

Acknowledgments. This work was partially supported by the project TIN2017-88728-C2-1-R, MINECO, Plan Nacional de I+D+I, and I Plan Propio de Investigación y Transferencia of the Universidad de Málaga.

References

1. Schuster, S.C.: Next-generation sequencing transforms today's biology. Nature methods **5**(1) (2007) 16

2. Wang, Z., Gerstein, M., Snyder, M.: Rna-seq: a revolutionary tool for transcriptomics. Nature reviews genetics **10**(1) (2009) 57
3. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I.: Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. **13** (2015) 8–17
4. Bashiri, A., Ghazisaeedi, M., Safdari, R., Shahmoradi, L., Ehtesham, H.: Improving the prediction of survival in cancer patients by using machine learning techniques: Experience of gene expression data: A narrative review. Iran. J. Public Health **46**(2) (February 2017) 165–172
5. Guyon, I.: An introduction to variable and feature selection. Journal of Machine Learning Research **3** (2003) 1157–1182
6. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics **23**(19) (2007) 2507–2517
7. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553) (2015) 436
8. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786) (2006) 504–507
9. Pan, S.J., Yang, Q., et al.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10) (2010) 1345–1359
10. Fakoor, R., Ladhak, F., Nazi, A., Huber, M.: Using deep learning to enhance cancer diagnosis and classification. In: Proceedings of the International Conference on Machine Learning. Volume 28., ACM New York, USA (2013)
11. Danaee, P., Ghaeini, R., Hendrix, D.A.: A deep learning approach for cancer detection and relevant gene identification. In: Pacific Symposium on Biocomputing, World Scientific (2017) 219–229
12. Xiao, Y., Wu, J., Lin, Z., Zhao, X.: A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data. Computer Methods and Programs in Biomedicine **166** (2018) 99–105
13. Sevakula, R.K., Singh, V., Verma, N.K., Kumar, C., Cui, Y.: Transfer learning for molecular cancer classification using deep neural networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2018)
14. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., et al.: The cancer genome atlas pan-cancer analysis project. Nature genetics **45**(10) (2013) 1113–1120
15. Goldman, M., Craft, B., Kamath, A., Brooks, A.N., Zhu, J., Haussler, D.: The ucsc xena platform for cancer genomics data visualization and interpretation. bioRxiv (2018)
16. Poultney, C., Chopra, S., Cun, Y.L., et al.: Efficient learning of sparse representations with an energy-based model. In: Advances in neural information processing systems. (2007) 1137–1144
17. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th international conference on Machine learning, ACM (2008) 1096–1103
18. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research **15**(1) (2014) 1929–1958
19. Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al.: Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology **27**(8) (2009) 1160
20. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. Journal of Machine Learning Research **13**(Feb) (2012) 281–305