IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 27, 2017, accepted March 21, 2017, date of publication April 24, 2017, date of current version June 7, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2696523

Transfer Learning Approaches to Improve Drug Sensitivity Prediction in Multiple Myeloma Patients

TURKI TURKI[1,2], ZHI WEI[2], (Member, IEEE), and JASON T. L. WANG[2], (Member, IEEE)
[1]Department of Computer Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2]Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102 USA

Corresponding author: Turki Turki (tturki@kau.edu.sa)

ABSTRACT Traditional machine learning approaches to drug sensitivity prediction assume that training data and test data must be in the same feature space and have the same underlying distribution. However, in real-world applications, this assumption does not hold. For example, we sometimes have limited training data for the task of drug sensitivity prediction in multiple myeloma patients (target task), but we have sufficient auxiliary data for the task of drug sensitivity prediction in patients with another cancer type (related task), where the auxiliary data for the related task are in a different feature space or have a different distribution. In such cases, transfer learning, if applied correctly, would improve the performance of prediction algorithms on the test data of the target task via leveraging the auxiliary data from the related task. In this paper, we present two transfer learning approaches that combine the auxiliary data from the related task with the training data of the target task to improve the prediction performance on the test data of the target task. We evaluate the performance of our transfer learning approaches exploiting three auxiliary data sets and compare them against baseline approaches using the area under the receiver operating characteristic curve on the test data of the target task. Experimental results demonstrate the good performance of our approaches and their superiority over the baseline approaches when auxiliary data are incorporated.

INDEX TERMS Machine learning, data mining, clinical informatics, precision medicine, cancer drug discovery.

I. INTRODUCTION

Cancer has a significant impact on public health worldwide and is the second leading cause of death in the US [1]. In 2016, the American Cancer Society predicts that 1,685,210 new cancer cases will be diagnosed, resulting in 595,690 deaths attributable to cancer in the US. Many of these cancer patients respond differently to the same cancer drug during chemotherapy. These response differences are attributable to not only environmental (i.e., external) factors such as tobacco, infectious organisms and an unhealthy diet, but also genetic (i.e., internal) factors such as inherited genetic mutations, hormones, immune conditions, and cancer cell heterogeneity, all of which make cancer drug discovery very difficult [2]–[6]. Because of the significant numbers of deaths associated with cancer, its study has attracted the attention of researchers from numerous domains including computational biology, machine learning, and data mining [7]–[11].

Traditional machine learning approaches to drug sensitivity prediction have been adopted to improve the performance of prediction algorithms. For example, Riddick *et al.* [12] presented an approach that employs random forests as a learning algorithm trained on gene expression signatures of selected cancer cell lines and corresponding drug IC_{50} values (i.e., labels), to induce (i.e., learn) a model. The learned model is then applied to gene expression signatures of cancer cell lines in the test set, to yield drug sensitivity predictions. Geeleher *et al.* [13] proposed an approach to drug sensitivity prediction that works as follows. The input data consisted of baseline expressions with drug IC_{50} values in cell lines and in vivo tumor gene expression. The raw microarray data for the cell lines and clinical trials were processed separately and then combined and homogenized. The homogenized expression data consisted of cell lines expression data (i.e., baseline gene expression levels in the cell lines) and clinical trial expression data (i.e., baseline tumor expression data from clinical trials). A learning algorithm was applied to the training set containing cell lines expression data along with the associated drug IC_{50} values for those cell lines, to

learn a model. The resulting model was applied to the clinical trial expression data in the test set, to yield drug sensitivity predictions.

Costello *et al.* [14] assessed the performance of 44 drug sensitivity prediction algorithms based on genomic, proteomic, and epigenomic profiling data for 53 breast cancer cell lines. The training set consisted of several profiling data for 35 cell lines, where each cell line was associated with responses of 28 drugs. The test set consisted of profiling data for 18 cell lines. The drug response data (also called the ground truth) were hidden for evaluation purposes. The goal of each prediction algorithm was to induce (i.e., learn) a model from the training set, and then perform predictions on the test set. The predicted drug responses corresponded to a ranked list of the most sensitive (to be ranked first) to the most resistant (to be ranked last) cell lines for each drug across all the 18 cell lines in the test set. The algorithms' predictions were evaluated against the ground truth using a weighted probabilistic c-index (wpc-index) to report final team rankings and resampled Spearman correlations for verifying the consistency between the team rankings [14]. The top-performing approach worked by integrating several profiling data with improved representation combined with a probabilistic nonlinear regression model [14]. The second-best performing approach employed random forest regression to learn a model from profiling data of the training set and perform predictions on the test set. The remaining prediction algorithms were not statistically different.

The previous approaches work well only under the common assumption: the training set and test set are in the same feature space and have the same distribution. However, this assumption does not hold in real-world applications [15]. As an example, consider the task of predicting drug sensitivity in multiple myeloma patients (referred to as the target task) where we have limited training data (called target training data). However, there exist an abundance of labeled auxiliary data for the task of predicting drug sensitivity in patients with another cancer type (referred to as the related task), where the auxiliary data are in a different feature space or have a different distribution. In addition, collecting additional training data to improve the accuracy of prediction algorithms for the target task requires larger infrastructures and is associated with higher costs of screening size [16]. Therefore, there is a need to create high-performance prediction algorithms trained with more easily obtained data from a related task. This methodology is referred to as transfer learning [15], [17], [18].

The key contributions of our paper are as follows: (1) we present two transfer learning approaches for the health informatics domain that combine auxiliary data from the related task with target training data, allowing a machine learning model to achieve high performance in the target task, and (2) we perform an experimental study on clinical trial data where we leverage three auxiliary datasets, combined one at a time with the target training set, to demonstrate the predictive power and the stability of our prediction algorithms that employ our proposed approaches against the prediction algorithms that employ baseline approaches.

The rest of this paper is organized as follows. Section 2 reviews notations and methods related to our work. Section 3 describes the details of our proposed approaches, including a transfer learning approach and a boosted transfer learning approach. Section 4 reports experimental results, including the comparison of our proposed approaches against the baseline approaches on clinical trial data pertaining to multiple myeloma patients. Section 5 presents an in-depth discussion of these results. Section 6 concludes the paper and points out some directions for future research. In the sequel, we use the terms ''sensitive'' (''resistant'', respectively) and ''responder'' (''non-responder'', respectively) interchangeably. The terms ''genes'' and ''features'' are also used interchangeably throughout the paper.

II. BACKGROUND

This section provides an introduction to the methods related to our work, namely synthetic minority over-sampling technique (SMOTE) [19] and CUR matrix decomposition [20]. We introduce each of them respectively after we present notations used in the paper.

A. NOTATIONS

To give a better understanding of the algorithms, we first summarize the notations used in the paper. Matrices are written as uppercase letters, e.g., matrix X. Vectors are denoted by lowercase letters, e.g., x. Vector elements are denoted by italic lowercase letters as scalars, e.g., y_i or x. A transpose of a matrix or a vector is indicated by T. So, for example, if x is a row vector, x^T is the corresponding column vector.

B. SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE (SMOTE)

SMOTE [19] is a popular and a powerful over-sampling method that has shown a great deal of success in many applications [21]–[23]. Here, we are given a dataset $\mathrm{D}_+ \cup \mathrm{D}_-$. $\mathrm{D}_+ \in \mathbb{R}^{m \times d}$ contains examples from the minority class, $\mathrm{D}_- \in \mathbb{R}^{n \times d}$ contains examples from the majority class, and $m \ll n$. For each example $\mathrm{x}_i \in \mathrm{D}_+$, SMOTE finds the k nearest neighbors $\mathrm{x}_i^1, \mathrm{x}_i^2, ..., \mathrm{x}_i^k$ of $\mathrm{x}_i \in \mathrm{D}_+$, where $\mathrm{x}_i^j \in \mathbb{R}^d$, $1 \leq j \leq k$, refers to the jth nearest neighbor of the ith example x_i in D_+. Then SMOTE generates synthetic examples $\mathrm{z}_i^1, \mathrm{z}_i^2, ..., \mathrm{z}_i^k$ along the lines between each minority example $\mathrm{x}_i \in \mathrm{D}_+$ and its k nearest neighbors in the minority class as follows:

1) for $i = 1$ to m
    1.1) for $j = 1$ to k
        1.1.1) $\mathrm{z}_i^j = \mathrm{x}_i + (\mathrm{x}_i^j - \mathrm{x}_i)\lambda$
        1.1.2) Store $[\mathrm{z}_i^j, +]$ in D_{++}
    1.2) end for
2) end for

where $\mathrm{z}_i^j \in \mathbb{R}^d$ refers to the jth synthetic example generated from the ith example $\mathrm{x}_i \in \mathrm{D}_+$, $\lambda \in (0, 1)$ is a random

number, and the + sign indicates that synthetic examples are labeled with the minority class label. A random subset $D'_{++} \subseteq D_{++}$ is then selected, where D'_{++} consists of $n - m$ synthetic examples. A learning algorithm could be called on the balanced dataset $D'_{++} \cup D_+ \cup D_-$, to induce a model and perform predictions on a given test set.

C. CUR MATRIX DECOMPOSITION

Suppose that we are given a dataset $F \in \mathbb{R}^{m \times p}$. Mahoney and Drineas [20] proposed CUR matrix decomposition as a dimensionality reduction paradigm that aims to obtain a low rank approximation of the matrix F, which is expressed in terms of some actual rows and columns of the original matrix F:

$$F \approx CUR \tag{1}$$

where C consists of a small number of the actual columns of F, R consists of a small number of the actual rows of F, and U is a constructed matrix that guarantees that CUR is close to F. Let v_j^{ξ} be the jth element of the ξth right singular vector of F. Let l be the rank of F. Then the normalized statistical leverage scores equal

$$\pi_j = \frac{1}{l} \sum_{\xi=1}^{l} (v_j^{\xi})^2 \tag{2}$$

for $j = 1, \ldots, p$ and $\sum_{j=1}^{p} \pi_j = 1$. C, U, and R matrices are constructed after calling the COLUMNSELECT algorithm of Mahoney et al., which takes the input matrix F, the rank parameter l, and an error parameter ε, and then performs the following steps:

1) Compute $v^1, v^2, ..., v^l$ (i.e., the top l right singular vectors of F) and the normalized statistical leverage scores in Equation (2).
2) Keep the jth column of F with probability $p_j = \min\{1, c\pi_j\}$ for $j = 1, \ldots, p$ where $c = O(l \log l/\varepsilon^2)$.
3) Return the matrix C consisting of the selected columns of F.

In step 1, the singular value decomposition (SVD) of F is computed, which decomposes F into $U\Sigma V^T$, where $U \in \mathbb{R}^{m \times l}$ is the orthogonal matrix containing the top l left singular vectors of F, $\Sigma \in \mathbb{R}^{l \times l}$ is the diagonal matrix containing singular values of F, $V^T \in \mathbb{R}^{l \times p}$ is the orthogonal matrix containing the top l right singular vectors of F, and l is the rank of F. The columns of U are pairwise orthogonal and normal (i.e., orthonormal), but its rows are not orthonormal as Euclidean norm is between 0 and 1. The rows of V^T are pairwise orthogonal and normal (i.e., orthonormal), but its columns are not orthonormal as Euclidean norm is between 0 and 1 [24]. The other matrices (i.e., R, and U) are constructed as follows:

1) Run COLUMNSELECT on F^T with $c = O(l \log l/\varepsilon^2)$ to choose rows of F (columns of F^T) and construct the matrix R.
2) The matrix U is defined as $U = C^+FR^+$, where C^+ and R^+ denote the Moore-Penrose generalized inverse of the matrices C and R, respectively.

Statistical leverage scores have been successfully used in data analysis to identify the most influential genes and outlier detection [20]. A high statistical leverage score for a given gene indicates that the gene is regarded as an important (i.e., influential) gene. A low statistical leverage score for a given gene indicates that the gene is regarded as an unimportant gene. To select the q most important genes from the matrix F, where $q < p$, we find the highest q statistical leverage scores used in computing the matrix C of F, which correspond to the q most influential (i.e., important) genes.

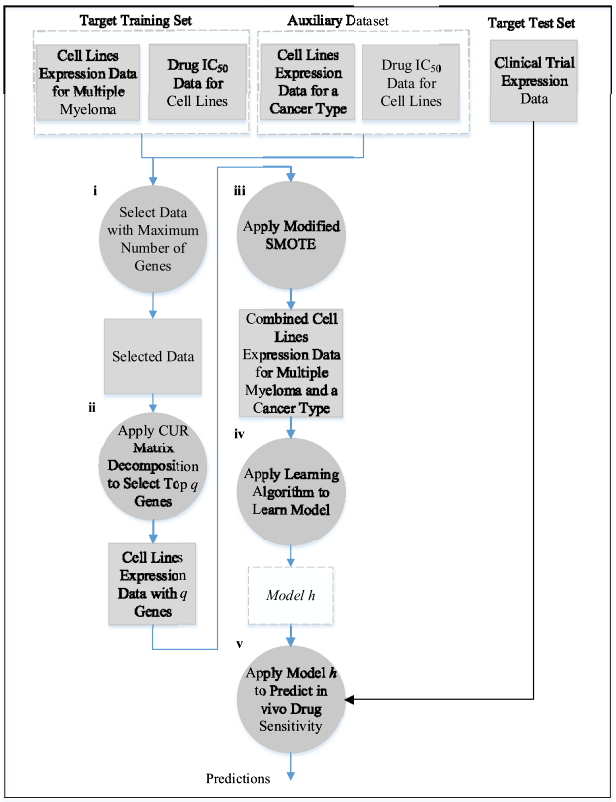

FIGURE 1. Flowchart of the proposed transfer learning approach to predicting in vivo drug sensitivity.

III. PROPOSED APPROACHES

A. THE TRANSFER LEARNING APPROACH

Figure 1 illustrates the proposed transfer learning approach, which works as follows. Suppose that we are given a target training set $F = \{(x_1, y_1), ..., (x_m, y_m)\}$ and a target test set $T = \{t_1, ..., t_r\}$. In the target training set, $x_i \in \mathbb{R}^p$ is the ith target training example with p genes (i.e., features), $y_i \in \mathbb{R}$ is the corresponding label of x_i, and $t_i \in \mathbb{R}^p$ is the ith target testing example with p genes. The target training set and target test set are disjoint, where m and r are the numbers of training and testing examples, respectively, in the target task. In addition, we have an auxiliary dataset $S = \{(s_1, u_1), ..., (s_f, u_f)\}$, where $s_i \in \mathbb{R}^n$ is the ith example (i.e., cell line of a cancer type) with

n genes (i.e., features), $u_i \in \mathbb{R}$ is the corresponding label of s_i, and n, the number of genes in the auxiliary data, is different from p, the number of genes in the target task. Our goal is to improve the prediction performance on the target test set T of the target task (i.e., prediction of bortezomib sensitivity in multiple myeloma patients) via learning an accurate model using the auxiliary dataset S and the target training set F. We summarize the problem definition in Table 1.

TABLE 1. Problem formulation.

Learning objective	Make predictions on the target test set of the target task
Target task	Prediction of bortezomib sensitivity in multiple myeloma patients
Related task	Prediction of specific drug sensitivity in patients with another cancer type
Target task data	Target training set: $F=\{(x_i,y_i)\}_{i=1}^{m}$ Target test set: $T=\{t_i\}_{i=1}^{r}$
Related task data	Auxiliary dataset: $S=\{(s_i,u_i)\}_{i=1}^{l}$ Cancer types: breast cancer, triple-negative breast cancer, and non-small cell lung cancer

To incorporate the auxiliary data into the target training set, we perform the following steps.

(i) If the number of genes p in the target training set F is greater than the number of genes n in the auxiliary dataset S, then we perform gene (i.e., feature) selection on F as explained in step (ii). Otherwise, we perform gene selection on S. Assume without loss of generality that $p > n$.

(ii) We select q genes from F based on their importance scores as defined in Equation (2), which depend on computing the matrices C, U, R of F and the input rank parameter l. (In this study, $q = n$ and we used the default parameter values for l, c, and ε in the CUR function [25].) Specifically, we store the indexes of the highest q leverage scores in I where $q < p$; these indexes correspond to the positions of the q most important genes in the matrix F. We then select the q genes from the target training set F based on the positions in I and store the target training examples with the q genes in $F' = \{(x_1', y_1'), ..., (x_m', y_m')\}$.

(iii) The following steps are based on a modified version of SMOTE [19] where each example in the auxiliary dataset S obtains a representation closer to the target training set F':

1) Select b examples from the auxiliary dataset S. (In the study presented here, $b = 100$.) For each example s_i, $1 \le i \le b$, selected from S, pick one of s_i's nearest target training examples from F', denoted x_i^*, such that the picked example is different from all the target training examples previously picked for s_j, $1 \le j < i$. More precisely, suppose s_i's k nearest target training examples are among the target training examples previously picked for s_j, $1 \le j < i$. Then x_i^* is s_i's $(k+1)$th nearest target training example from F'. Let y_i' be the corresponding label of x_i^*.
2) Change the representation of the examples selected from S using the following lines of code:

2.1) for $i = 1$ to b

2.1.1) $s_i^* = s_i + (x_j^* - s_i)\lambda$

2.1.2) Store $[s_i^*, (y_i - \alpha)]$ in S^+

2.2) end for

where $\lambda = 0.99$, $\alpha = 0.01$, and S^+ contains the new representations of the auxiliary data. Let $D = S^+ \cup F'$ contain the combined cell lines expression data, where $D \in \mathbb{R}^{m' \times n}$, and $m' = m + b$.

(iv) A learning algorithm is called on D to induce a model h.

(v) The n most important genes in the target test set T are selected based on the positions in I and stored in T'. The model h is applied to the target test set T' to perform predictions.

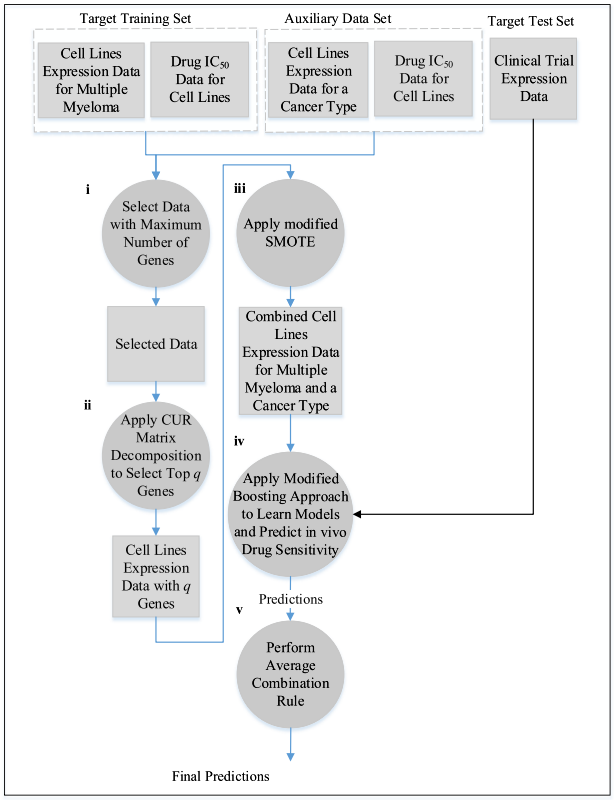

FIGURE 2. Flowchart of the proposed boosted transfer learning approach to predicting in vivo drug sensitivity.

B. THE BOOSTED TRANSFER LEARNING APPROACH

Figure 2 illustrates the proposed boosted transfer learning approach. Here, steps (i), (ii), and (iii) are the same as steps (i), (ii), and (iii) of the transfer learning approach.

(iv) We employ a modified AdaBoost algorithm [26]–[29], which works as follows. Initially, each training example $(x_i, y_i) \in D$ is assigned a weight $w_i = 1$ for $i = 1, \ldots, m'$. The probability for selecting the ith training example (x_i, y_i) in the training set D is

$$p_i = \frac{w_i}{\sum_{i=1}^{m'} w_i} \tag{3}$$

where $\sum_{i=1}^{m'} p_i = 1$. Select m' training examples (without replacement) from D to form the training set D'. A learning

algorithm is called on D' to learn a model h and perform predictions on D, where the predictions are then stored in $y' = (y'_1, ..., y'_{m'})$. Select the n most important genes in the target test set $T = \{t_1, ..., t_r\}$ using the positions in I, and store the target testing examples with the n genes in T'. Apply the model h to the target test set T' to yield predictions, which are stored as the first row vector in a matrix G. Repeat the following steps j times by executing the while loop below. (In this study, $j = 6$.) Initially $a = 1$.

While $a \leq j$

1) Update the weights: $w_i = (y_i - y'_i)^2$ for $i = 1, \ldots, m'$.
2) Calculate probabilities $p = (p_1, p_2, ..., p_{m'})$ of the training examples in D, where p_i, $1 \leq i \leq m'$, is as defined in Equation (3).
3) Calculate the median of the probabilities $p_1, p_2, \ldots, p_{m'}$ and store the median in v.
4) Select training examples from D where the weight of each selected example must be greater than or equal to v. Store the selected training examples in D'. Let p* contain the probabilities corresponding to the selected training examples.
5) Select m' training examples (with replacement) from D' according to the probabilities in p* and store the selected training examples in D''. The higher probability a training example is associated with, the more likely this training example will be included in D''.
6) A learning algorithm is called on D'' to learn a model h and perform predictions on D.
7) Store the predictions performed on D in q.
8) Let $y' = y' + q$, which corresponds to the cumulative predictions on the training set D.
9) Apply the learned model h to the target test set T' and store the predictions as the $(a+1)$th row vector in G.
10) $a = a + 1$.

(v) Output the final predictions as

$$Q = e^T G \tag{4}$$

where G is a $(j+1) \times r$ matrix of predictions on the target test set T', the ith row vector of G corresponds to the predictions made in the $(i-1)$th iteration in step (iv), $e = (\frac{1}{j+1}, ..., \frac{1}{j+1})^T$ is a $(j+1) \times 1$ column vector, and Q is a $1 \times r$ row vector where the ith element in Q is the average of the values in the ith column of G.

IV. EXPERIMENTS

A. DATASETS

1) DATA PERTAINING TO MULTIPLE MYELOMA PATIENTS

The target training set $F \in \mathbb{R}^{280 \times 9115}$ contains 280 target training examples (i.e., cancer cell lines), 9,114 genes, and drug IC_{50} values that correspond to a 280-dimensional column vector. The target test set $T \in \mathbb{R}^{188 \times 9114}$ is composed of 188 samples of multiple myeloma patients and 9,114 genes. The drug IC_{50} values for bortezomib [30], [31] were downloaded from (http://genemed.uchicago.edu/~pgeeleher/cgpPrediction/) [13], and the data for the cancer cell lines were downloaded from the ArrayExpress repository (the accession number is E-MTAB-783 or available at https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-783/?query=EMTAB783) [32]–[34]. The clinical trial data corresponding to the target test set were downloaded from the Gene Expression Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo/) with the accession number GSE9782. The data were downloaded, processed and mapped according to Geeleher *et al.* [13].

2) DATA PERTAINING TO BREAST CANCER PATIENTS

The auxiliary data correspond to a 482×6539 matrix containing 482 examples and 6,538 genes plus labels, i.e., drug IC_{50} values, for breast cancer patients. The drug IC_{50} values for docetaxel [35], [36] (a chemotherapy drug) were downloaded from (http://genemed.uchicago.edu/~pgeeleher/cgpPrediction/) [13]. The cell lines expression data were downloaded from the ArrayExpress repository (with the accession number being E-MTAB-783, available at https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-783/?query=EMTAB783) [32]–[34]. All the data were downloaded and processed according to the approach developed by Geeleher *et al.* [13].

3) DATA PERTAINING TO PATIENTS OF TRIPLE-NEGATIVE BREAST CANCER AND NON-SMALL CELL LUNG CANCER

The auxiliary data correspond to a 497×9621 matrix containing 497 examples and 9,620 genes plus labels and a 258×9508 matrix containing 258 examples and 9,507 genes plus labels for triple-negative breast cancer patients and non-small cell lung cancer patients, respectively. The data were downloaded from (http://genemed.uchicago.edu/~pgeeleher/cgpPrediction/).

B. EVALUATION AND BASELINE APPROACHES

We compared our proposed transfer learning approaches with two different baseline approaches, described below.

1) FIRST BASELINE (B1)

This baseline employs the approach developed by Geeleher *et al.* [13].

2) SECOND BASELINE (B2)

In this baseline, we apply CUR matrix decomposition to F. We then store the indexes of the largest n statistical leverage scores of F in I, as in our proposed approaches. The n most important genes from the target training examples in F are selected using the positions in I. A learning algorithm is called on the auxiliary data with n genes combined with the target training examples with the n most important genes, to learn a model h. Then, the n most important genes in the target test set are selected using the positions in I. The model h is applied to the target testing examples with the n most important genes, to yield drug sensitivity predictions. Thus, this baseline differs from our proposed approaches in that it does not have a transfer learning mechanism (cf. step (iii) in Section 3.1).

TABLE 2. Summary of the twelve drug sensitivity prediction algorithms studied in this paper.

Abbreviation	Prediction Algorithm
T+SVR+L	The transfer learning approach using support vector regression with a linear kernel
T+SVR+S	The transfer learning approach using support vector regression with a sigmoid kernel
T+RR	The transfer learning approach using ridge regression
BT+SVR+L	The boosted transfer learning approach using support vector regression with a linear kernel
BT+SVR+S	The boosted transfer learning approach using support vector regression with a sigmoid kernel
BT+RR	The boosted transfer learning approach using ridge regression
B1+SVR+L	The first baseline approach using support vector regression with a linear kernel
B1+SVR+S	The first baseline approach using support vector regression with a sigmoid kernel
B1+RR	The first baseline approach using ridge regression
B2+SVR+L	The second baseline approach using support vector regression with a linear kernel
B2+SVR+S	The second baseline approach using support vector regression with a sigmoid kernel
B2+RR	The second baseline approach using ridge regression

The proposed transfer learning approaches and the baseline approaches employ two machine learning algorithms, namely support vector regression (SVR) and ridge regression (RR). Table 2 summarizes the twelve prediction algorithms studied in this paper.

Each prediction algorithm was trained on a training set, whose labels were continuous, to yield a model. Then, each model was applied to the target test set to yield predictions (i.e., predicted labels), which were also continuous values. The target test set consists of patients' clinical trial expression data, which are baseline tumor expression data from primary tumor biopsies before treatment with a cancer drug (e.g., bortezomib or docetaxel).

The true labels of the target test set are categorical, which are either ''sensitive'' or ''resistant''. These true labels were clinically evaluated by the degree of reduction in tumor size to a cancer drug. A cancer patient is categorized as sensitive to the cancer drug treatment if the cancer patient exhibits less than 25% residual tumor. A cancer patient is categorized as resistant to the cancer drug treatment if the cancer patient exhibits greater than or equal to 25% residual tumor [13].

Using in vitro drug sensitivity of the training data to predict in vivo drug sensitivity of the target test set is a challenging task and a main goal in precision medicine, which corresponds to predicting the clinical outcome that is crucial for the life of the human being [37]. If the clinical drug response (i.e., clinical response to a cancer drug) is incorrectly predicted, the tumor size of a cancer patient would increase significantly over the time, which causes sequelae that lead to death. If the clinical drug response is correctly predicted, the tumor size would decrease significantly over the time and that would save the patient. By predicting clinical outcomes in the target test set correctly, clinicians would benefit from understanding the relationship between in vivo and in vitro drug sensitivity, which leads to better personalized treatment.

Ten-fold cross validation is not suitable in this study as labels of the target test set are categorical while labels of the corresponding target training set are real numbers. Hence, to evaluate whether the proposed approaches exhibit stable performance as sample sizes change, we randomly reduced the sample size for the target training set by 1% in each run, until the reduction reached 4%. In other words, we performed 5 runs with sample sizes of 280, 278, 275, 272, 269, respectively.

The accuracy of the prediction algorithms was measured using the area under the receiver operating characteristic (ROC) curve (AUC), as described in [13]. The higher the AUC score an algorithm achieves, the better its performance is. We used MAUC to denote the mean of the AUC values averaged over the five runs of experiments. Each run here includes the predictions of a learned model on the target test set in which the model was learned from a training set whose size is varied to assess the stability of the prediction algorithms. A stable prediction algorithm is one whose prediction accuracy on the target test set does not change dramatically owing to small changes of the training set size [38], [39]. This type of assessment is important in biological systems, where the best prediction algorithm is the one that outperforms the other algorithms many times on conducted experiments. The statistical significance of each prediction algorithm was calculated.

The software used in this work included support vector regression with linear and sigmoid kernels (with their default parameter values) in the LIBSVM package [40], ridge regression [13], gene selection using CUR and topLeverage functions in the rCUR package [25], and R code for processing the datasets and performance evaluation [13]. We used R to write code for the prediction algorithms and to perform the experiments.

C. EXPERIMENTAL RESULTS

We evaluate the relative performance of our proposed approaches compared to the baseline approaches. Each time we use the target training set of multiple myeloma along with one of the auxiliary datasets pertaining to breast cancer, triple-negative breast cancer, and non-small cell lung cancer respectively to train the approaches described in the paper (except the B1 approach that uses only the target training set), to yield prediction models and perform predictions on the target test set.

1) EXPLOITING AUXILIARY DATA OF BREAST CANCER PATIENTS

Table 3 shows details of the target training set and auxiliary dataset pertaining to breast cancer patients used by each prediction algorithm. The target training set is obtained from the target task (i.e., prediction of bortezomib sensitivity in multiple myeloma patients) and the auxiliary data are acquired from the related task (i.e., prediction of docetaxel sensitivity in breast cancer patients). Row ''m/l'' shows the number of examples or cell lines in the target training set/auxiliary dataset used in each run. Row ''p/n'' shows the number of genes or features in the target training set/auxiliary dataset used in each run. Row ''$p \cap n$'' shows the number of overlapped (i.e., intersected) genes between the target training set and the auxiliary dataset in each run. Rows ''$\mathrm{P}_{m/l}$'' and ''$\mathrm{P}_{p/n}$'' show the number of selected examples in the target training set/auxiliary dataset and the number of selected genes in the target training set/auxiliary dataset, respectively, that were used by the prediction algorithms employing our approaches during the training stage to learn models. Rows ''$\mathrm{B1}_m$'' and ''$\mathrm{B1}_p$'' show the number of selected exam-

TABLE 3. Details of the target training set and auxiliary dataset pertaining to multiple myeloma patients and breast cancer patients, respectively, used by each prediction algorithm.

Run	1	2	3	4	5
m/l	280/482	278/482	275/482	272/482	269/482
p/n	9114/6538	9114/6538	9114/6538	9114/6538	9114/6538
$p \cap n$	5478	5478	5478	5478	5478
$P_{m/l}$	280/100	278/100	275/100	272/100	269/100
$P_{p/n}$	6538/6538	6538/6538	6538/6538	6538/6538	6538/6538
$B1_m$	280	278	275	272	269
$B1_p$	9114	9114	9114	9114	9114
$B2_{m/l}$	280/482	278/482	275/482	272/482	269/482
$B2_{p/n}$	6538/6538	6538/6538	6538/6538	6538/6538	6538/6538

ples and genes, respectively, in the target training set that were used during the training stage by the prediction algorithms employing the first baseline approach (B1). Rows "$B2_{m/l}$" and "$B2_{p/n}$" show the number of selected examples and genes, respectively, in the target training set/auxiliary dataset that were used by the prediction algorithms employing the second baseline approach (B2). In each run we change the size of the target training set and train all the prediction algorithms employing the approaches described in Sections 3 and 4.2 to yield models.

TABLE 4. AUC scores of the twelve prediction algorithms on the target test set of multiple myeloma patients where the target training set and auxiliary dataset pertaining to multiple myeloma patients and breast cancer patients, respectively, are used. In each run, the highest AUC is shown in bold. Std is the standard deviation of the AUC values obtained from the five runs.

Run	1	2	3	4	5	Std
T+SVR+L	0.644	0.636	0.640	0.647	0.653	0.0065
T+SVR+S	0.643	0.642	0.654	0.659	0.666	0.0103
T+RR	0.653	0.655	0.652	0.657	0.652	0.0021
BT+SVR+L	0.658	0.657	0.678	0.665	0.675	0.0096
BT+SVR+S	**0.682**	0.682	**0.687**	**0.695**	**0.703**	0.0090
BT+RR	0.670	**0.693**	0.668	0.681	0.659	0.0131
B1+SVR+L	0.613	0.609	0.622	0.628	0.632	0.0097
B1+SVR+S	0.602	0.600	0.601	0.605	0.598	0.0025
B1+RR	0.614	0.611	0.603	0.607	0.606	0.0043
B2+SVR+L	0.440	0.491	0.466	0.501	0.549	0.0408
B2+SVR+S	0.449	0.485	0.487	0.506	0.500	0.0221
B2+RR	0.499	0.505	0.511	0.511	0.516	0.0065

Table 4 shows the AUCs of the twelve prediction algorithms on the target test set of multiple myeloma patients. As shown in Table 4, BT+SVR+S performs better than the baseline prediction algorithms (i.e., B2+SVR+L, B2+SVR+S, B2+RR, B1+SVR+L, B1+SVR+S and B1+RR). In particular, BT+SVR+S achieves the highest AUC in 4 out of 5 runs. The BT+SVR+S results were consistently good compared to the other prediction algorithms in terms of AUC on the target test set as we changed the target training set size. These results indicate that the performance of BT+SVR+S is stable.

Table 5 shows the P-values of a two-sample t-test on the target test set for each run, as in [13]. For each prediction algorithm, its highly statistically significant results are shown in red ($P < 0.001$); its statistically significant results are shown in blue ($0.001 \leq P < 0.05$). As shown in Table 5, our proposed prediction algorithms yield highly statistically

TABLE 5. P-values of a two-sample t-test for the twelve prediction algorithms on the target test set where the target training set and auxiliary dataset pertaining to multiple myeloma patients and breast cancer patients, respectively, are used. For each prediction algorithm, its results with $P < 0.001$ are considered highly statistically significant and colored in red; its results with $P < 0.05$ are considered statistically significant and colored in blue.

Run	1	2	3	4	5
T+SVR+L	6088×10^{-6}	906×10^{-6}	674×10^{-6}	442×10^{-6}	372×10^{-6}
T+SVR+S	765×10^{-6}	815×10^{-6}	498×10^{-6}	273×10^{-6}	219×10^{-6}
T+RR	188×10^{-6}	218×10^{-6}	244×10^{-6}	194×10^{-6}	313×10^{-6}
BT+SVR+L	218×10^{-6}	245×10^{-6}	2396×10^{-6}	132×10^{-6}	6888×10^{-8}
BT+SVR+S	4×10^{-5}	2172×10^{-8}	1506×10^{-8}	1378×10^{-8}	527×10^{-8}
BT+RR	3382×10^{-8}	9221×10^{-9}	4307×10^{-8}	1169×10^{-8}	6069×10^{-8}
B1+SVR+L	6×10^{-3}	7×10^{-3}	4×10^{-3}	2×10^{-3}	2×10^{-3}
B1+SVR+S	8×10^{-3}	1×10^{-2}	11×10^{-3}	9×10^{-3}	14×10^{-3}
B1+RR	261×10^{-5}	4×10^{-3}	6×10^{-3}	4×10^{-3}	5×10^{-3}
B2+SVR+L	881×10^{-3}	665×10^{-3}	489×10^{-3}	724×10^{-3}	613×10^{-3}
B2+SVR+S	784×10^{-3}	722×10^{-3}	708×10^{-3}	607×10^{-3}	633×10^{-3}
B2+RR	4992×10^{-4}	466×10^{-3}	398×10^{-3}	386×10^{-3}	342×10^{-3}

significant results; these highly statistically significant results reflect the superior performance of our proposed prediction algorithms.

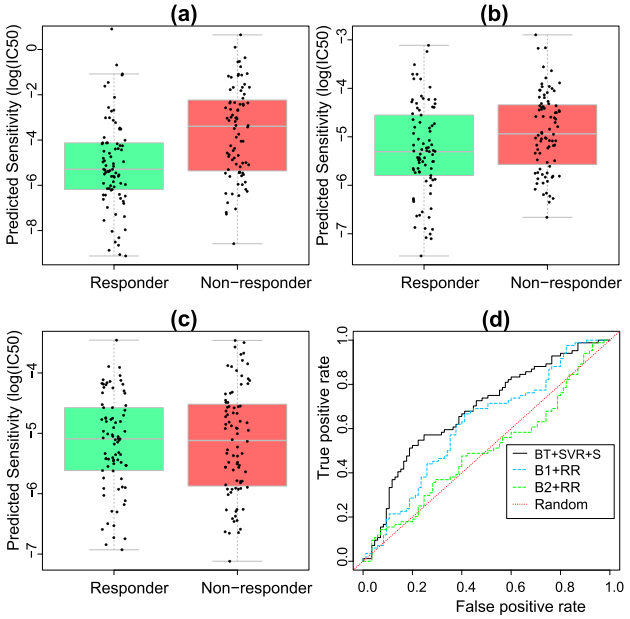

FIGURE 3. Predictions of bortezomib sensitivity on the target test set of multiple myeloma patients where the target training set and auxiliary dataset pertaining to multiple myeloma patients and breast cancer patients, respectively, are used. Strip charts and boxplots in (a), (b), and (c) show the differences in predicted drug sensitivity to bortezomib treatment between the responder (i.e., sensitive) group and non-responder (i.e., resistant) group using BT+SVR+S, B1+RR, and B2+RR, respectively. (d) shows the ROC curves of the prediction algorithms, which reveal the proportion of true positives compared to the proportion of false positives. ROC = receiver operating characteristic.

Figures 3(a), 3(b), and 3(c) show the predictions of BT+SVR+S, B1+RR, and B2+RR, respectively, on the target test set in the first run. The result of BT+SVR+S shown in Figure 3(a) was highly statistically significant ($P = 4 \times 10^{-5}$ from a two-sample t-test). The result of B1+RR shown in Figure 3(b) was statistically significant with $P = 261 \times 10^{-5}$ from a two-sample t-test. The result of B2+RR shown in Figure 3(c) was not statistically

significant with $P = 49920 \times 10^{-5}$ from a two-sample t-test. In Figure 3(d), the ROC curves reveal AUC values of 0.682, 0.614, and 0.499 for BT+SVR+S, B1+RR, and B2+RR, respectively.

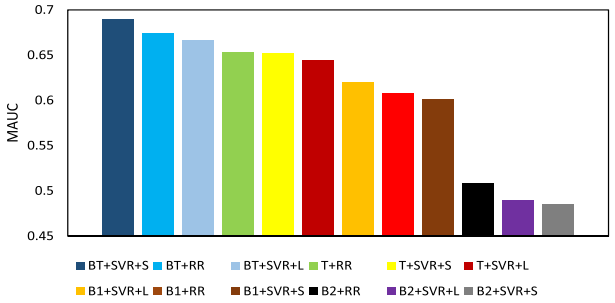

FIGURE 4. The Mean AUC (MAUC) values of the twelve bortezomib sensitivity prediction algorithms for multiple myeloma patients where the target training set and auxiliary dataset pertaining to multiple myeloma patients and breast cancer patients, respectively, are used. The algorithms are ranked from left to right where the leftmost algorithm has the highest MAUC and the rightmost algorithm has the lowest MAUC.

Figure 4 shows the ranking of the twelve prediction algorithms based on their MAUC values. The MAUC of an algorithm is calculated by taking the mean of the AUC values the algorithm receives from the 5 runs of experiments. As shown in Figure 4, our prediction algorithms outperform the baseline prediction algorithms with respect to the MAUC.

TABLE 6. Details of the target training set and auxiliary dataset pertaining to multiple myeloma patients and triple-negative breast cancer patients, respectively, used by each prediction algorithm.

Run	1	2	3	4	5
m/l	280/497	278/497	275/497	272/497	269/497
p/n	9114/9620	9114/9620	9114/9620	9114/9620	9114/9620
$p \cap n$	7911	7911	7911	7911	7911
$P_{m/l}$	280/100	278/100	275/100	272/100	269/100
$P_{p/n}$	9114/9114	9114/9114	9114/9114	9114/9114	9114/9114
$B1_m$	280	278	275	272	269
$B1_p$	9114	9114	9114	9114	9114
$B2_{m/l}$	280/497	278/497	275/497	272/497	269/497
$B2_{p/n}$	9114/9114	9114/9114	9114/9114	9114/9114	9114/9114

2) EXPLOITING AUXILIARY DATA OF TRIPLE-NEGATIVE BREAST CANCER PATIENTS

Table 6 shows details of the target training set and auxiliary dataset pertaining to triple-negative breast cancer patients used by each prediction algorithm. The target training set is obtained from the target task (i.e., prediction of bortezomib sensitivity in multiple myeloma patients) and the auxiliary dataset is obtained from the related task (i.e., prediction of cisplatin sensitivity in triple-negative breast cancer patients). The only difference between Table 3 and Table 6 is that Table 6 has different auxiliary data pertaining to triple-negative breast cancer patients, while the target test set remains the same.

Table 7 shows the AUCs of the twelve prediction algorithms on the target test set of multiple myeloma patients. As shown in Table 7, our prediction algorithms employing the boosted transfer learning (BT) approach perform

TABLE 7. AUC scores of the twelve prediction algorithms on the target test set of multiple myeloma patients where the target training set and auxiliary dataset pertaining to multiple myeloma patients and triple-negative breast cancer patients, respectively, are used. In each run, the highest AUC is shown in bold. Std is the standard deviation of the AUC values obtained from the five runs.

Run	1	2	3	4	5	Std
T+SVR+L	0.601	0.599	0.604	0.617	0.618	0.0090
T+SVR+S	0.620	0.621	0.629	0.634	0.632	0.0063
T+RR	0.615	0.615	0.614	0.620	0.623	0.0039
BT+SVR+L	0.628	0.635	0.665	0.627	**0.669**	0.0205
BT+SVR+S	**0.683**	0.641	0.659	**0.695**	0.667	0.0209
BT+RR	0.638	**0.654**	**0.675**	0.677	0.654	0.0163
B1+SVR+L	0.613	0.609	0.622	0.628	0.632	0.0097
B1+SVR+S	0.602	0.600	0.601	0.605	0.598	0.0025
B1+RR	0.614	0.611	0.603	0.607	0.606	0.0043
B2+SVR+L	0.528	0.528	0.538	0.536	0.523	0.0062
B2+SVR+S	0.472	0.466	0.473	0.477	0.471	0.0039
B2+RR	0.464	0.460	0.466	0.472	0.476	0.0063

better than the baseline prediction algorithms. Specifically, BT+SVR+S and BT+RR yielded the highest AUC in 4 out of 5 runs. These results indicate that our prediction algorithms employing the BT approach achieve high performance in terms of AUC on the target test set. The results also show the stability of the prediction algorithms employing the BT approach.

TABLE 8. *P*-values of a two-sample t-test for the twelve prediction algorithms on the target test set where the target training set and auxiliary dataset pertaining to multiple myeloma patients and triple-negative breast cancer patients, respectively, are used. For each prediction algorithm, its results with $P < 0.001$ are considered highly statistically significant and colored in red; its results with $P < 0.05$ are considered statistically significant and colored in blue.

Run	1	2	3	4	5
T+SVR+L	8701×10^{-6}	9608×10^{-6}	7425×10^{-6}	4303×10^{-6}	4126×10^{-6}
T+SVR+S	3877×10^{-6}	3499×10^{-6}	2547×10^{-6}	1708×10^{-6}	1897×10^{-6}
T+RR	3908×10^{-6}	4911×10^{-6}	5189×10^{-6}	363×10^{-5}	3932×10^{-6}
BT+SVR+L	2698×10^{-6}	1171×10^{-6}	188×10^{-6}	2163×10^{-6}	132×10^{-6}
BT+SVR+S	2716×10^{-8}	719×10^{-6}	185×10^{-6}	1185×10^{-8}	104×10^{-6}
BT+RR	382×10^{-6}	111×10^{-6}	2692×10^{-8}	2774×10^{-8}	115×10^{-6}
B1+SVR+L	6×10^{-3}	7×10^{-3}	4×10^{-3}	2×10^{-3}	2×10^{-3}
B1+SVR+S	8×10^{-3}	1×10^{-2}	11×10^{-3}	9×10^{-3}	14×10^{-3}
B1+RR	261×10^{-5}	4×10^{-3}	6×10^{-3}	4×10^{-3}	5×10^{-3}
B2+SVR+L	4249×10^{-4}	426×10^{-3}	3987×10^{-4}	3964×10^{-4}	4338×10^{-4}
B2+SVR+S	8505×10^{-4}	8672×10^{-4}	8341×10^{-4}	8028×10^{-4}	8263×10^{-4}
B2+RR	6622×10^{-4}	694×10^{-3}	6596×10^{-4}	6066×10^{-4}	57×10^{-2}

Table 8 shows the P-values of a two-sample t-test on the target test set for each run. Our prediction algorithms BT+SVR+S and BT+RR yield highly statistically significant results in each run (see the results colored in red in Table 8). These highly statistically significant results show the good performance of BT+SVR+S and BT+RR algorithms (cf. Table 7).

Figures 5(a), 5(b), and 5(c) show the predictions of BT+SVR+S, B1+RR, and B2+RR, respectively, on the target test set in the first run. BT+SVR+S in Figure 5(a) achieved a highly statistically significant result ($P = 2716 \times 10^{-8}$ from a two-sample t-test). The result of B1+RR in Figure 5(b) was statistically significant with $P = 261 \times 10^{-5}$ from a two-sample t-test. The result of B2+RR in Figure 5(c) was not statistically significant ($P = 6622 \times 10^{-4}$ from a two-sample t-test). In Figure 5(d), the ROC curves reveal AUC values of 0.683, 0.614, and 0.464 for BT+SVR+S, B1+RR, and B2+RR, respectively.

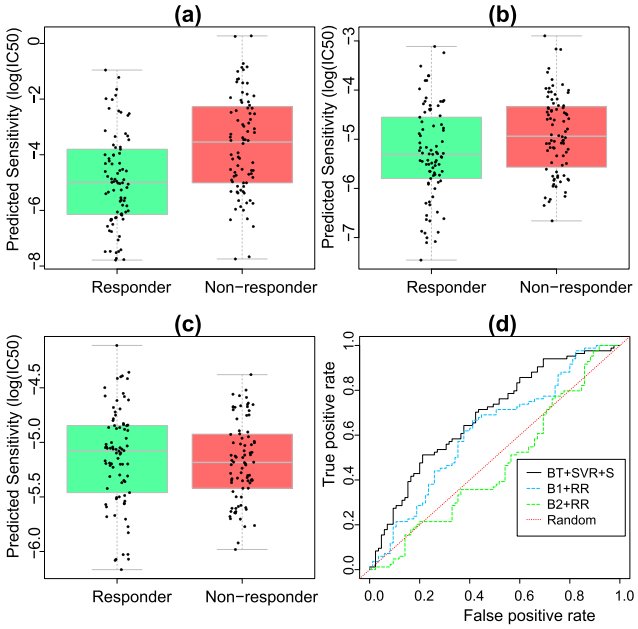

FIGURE 5. Predictions of bortezomib sensitivity on the target test set of multiple myeloma patients where the target training set and auxiliary dataset pertaining to multiple myeloma patients and triple-negative breast cancer patients, respectively, are used. Strip charts and boxplots in (a), (b), and (c) show the differences in predicted drug sensitivity to bortezomib treatment between the responder (i.e., sensitive) group and non-responder (i.e., resistant) group using BT+SVR+S, B1+RR, and B2+RR, respectively. (d) shows the ROC curves of the prediction algorithms, which reveal the proportion of true positives compared to the proportion of false positives. ROC = receiver operating characteristic.

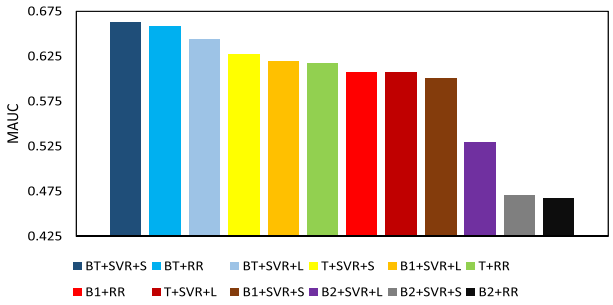

FIGURE 6. The Mean AUC (MAUC) values of the twelve bortezomib sensitivity prediction algorithms for multiple myeloma patients where the target training set and auxiliary dataset pertaining to multiple myeloma patients and triple-negative breast cancer patients, respectively, are used. The algorithms are ranked from left to right where the leftmost algorithm has the highest MAUC and the rightmost algorithm has the lowest MAUC.

Figure 6 shows that our prediction algorithms BT+SVR+S, BT+RR, BT+SVR+L, and T+SVR+S outperform the baseline prediction algorithms with respect to the MAUC.

3) EXPLOITING AUXILIARY DATA OF NON-SMALL CELL LUNG CANCER PATIENTS

Table 9 shows details of the target training set and auxiliary dataset pertaining to non-small cell lung cancer patients used by each prediction algorithm. The target training set is obtained from the target task (i.e., prediction of bortezomib sensitivity in multiple myeloma patients) and the auxiliary dataset is obtained from the related task (i.e., prediction of erlotinib sensitivity in non-small cell lung cancer patients). Here, we use a different auxiliary dataset, which pertains to non-small cell lung cancer patients, while the target test set remains the same.

TABLE 9. Details of the target training set and auxiliary dataset pertaining to multiple myeloma patients and non-small cell lung cancer patients, respectively, used by each prediction algorithm.

Run	1	2	3	4	5
m/l	280/258	278/258	275/258	272/258	269/258
p/n	9114/9507	9114/9507	9114/9507	9114/9507	9114/9507
$p \cap n$	7855	7855	7855	7855	7855
$P_{m/l}$	280/100	278/100	275/100	272/100	269/100
$P_{p/n}$	9114/9114	9114/9114	9114/9114	9114/9114	9114/9114
$B1_m$	280	278	275	272	269
$B1_p$	9114	9114	9114	9114	9114
$B2_{m/l}$	280/258	278/258	275/258	272/258	269/258
$B2_{p/n}$	9114/9114	9114/9114	9114/9114	9114/9114	9114/9114

TABLE 10. AUC scores of the twelve prediction algorithms on the target test set of multiple myeloma patients where the target training set and auxiliary dataset pertaining to multiple myeloma patients and non-small cell lung cancer patients, respectively, are used. In each run, the highest AUC is shown in bold. Std is the standard deviation of the AUC values obtained from the five runs.

Run	1	2	3	4	5	Std
T+SVR+L	0.600	0.598	0.604	0.617	0.617	0.0092
T+SVR+S	0.621	0.619	0.630	0.635	0.634	0.0073
T+RR	0.614	0.615	0.610	0.615	0.624	0.0051
BT+SVR+L	0.670	0.653	0.623	**0.663**	0.663	0.0185
BT+SVR+S	**0.673**	**0.653**	**0.651**	0.657	**0.678**	0.0122
BT+RR	0.658	0.639	0.619	0.594	0.627	0.0237
B1+SVR+L	0.613	0.609	0.622	0.628	0.632	0.0097
B1+SVR+S	0.602	0.600	0.601	0.605	0.598	0.0025
B1+RR	0.614	0.611	0.603	0.607	0.606	0.0043
B2+SVR+L	0.403	0.409	0.436	0.422	0.421	0.0127
B2+SVR+S	0.643	0.644	0.644	0.646	0.648	0.0020
B2+RR	0.641	0.641	0.642	0.641	0.642	0.0005

Table 10 shows the AUCs of the twelve prediction algorithms on the target test set of multiple myeloma patients. Our prediction algorithm BT+SVR+S achieved the highest AUC scores in 4 out of 5 runs. The high performance results indicate the stability and superiority of the proposed BT approach combined with SVR+S.

TABLE 11. *P*-values of a two-sample *t*-test for the twelve prediction algorithms on the target test set where the target training set and auxiliary dataset pertaining to multiple myeloma patients and non-small cell lung cancer patients, respectively, are used. For each prediction algorithm, its results with $P < 0.001$ are considered highly statistically significant and colored in red; its results with $P < 0.05$ are considered statistically significant and colored in blue.

Run	1	2	3	4	5
T+SVR+L	8887×10^{-6}	9892×10^{-6}	7654×10^{-6}	4282×10^{-6}	4214×10^{-6}
T+SVR+S	3883×10^{-6}	3735×10^{-6}	2486×10^{-6}	1606×10^{-6}	1833×10^{-6}
T+RR	3887×10^{-6}	49×10^{-4}	6343×10^{-6}	4458×10^{-6}	3918×10^{-6}
BT+SVR+L	135×10^{-6}	368×10^{-6}	3893×10^{-6}	212×10^{-6}	18×10^{-5}
BT+SVR+S	1195×10^{-7}	261×10^{-6}	394×10^{-6}	383×10^{-6}	4695×10^{-8}
BT+RR	3887×10^{-6}	49×10^{-4}	6343×10^{-6}	4458×10^{-6}	3918×10^{-6}
B1+SVR+L	6×10^{-3}	7×10^{-3}	4×10^{-3}	2×10^{-3}	2×10^{-3}
B1+SVR+S	8×10^{-3}	1×10^{-2}	11×10^{-3}	9×10^{-3}	14×10^{-3}
B1+RR	261×10^{-5}	4×10^{-3}	6×10^{-3}	4×10^{-3}	5×10^{-3}
B2+SVR+L	939×10^{-3}	9358×10^{-4}	9073×10^{-4}	9173×10^{-4}	926×10^{-3}
B2+SVR+S	5556×10^{-6}	5455×10^{-6}	5414×10^{-6}	5493×10^{-6}	5196×10^{-6}
B2+RR	3881×10^{-6}	3855×10^{-6}	3888×10^{-6}	3905×10^{-6}	3822×10^{-6}

Table 11 shows the *P*-values of a two-sample *t*-test on the target test set for each run. Our prediction algorithm BT+SVR+S yields a highly statistically significant result

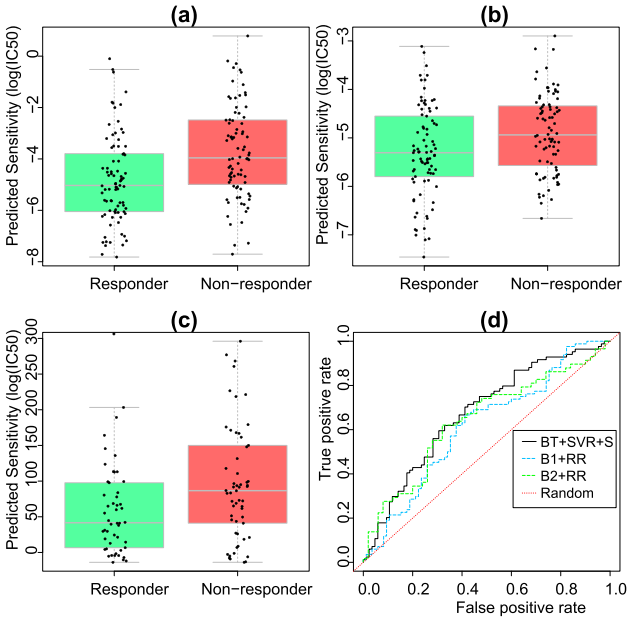

FIGURE 7. Predictions of bortezomib sensitivity on the target test set of multiple myeloma patients where the target training set and auxiliary dataset pertaining to multiple myeloma patients and non-small cell lung cancer patients, respectively, are used. Strip charts and boxplots in (a), (b), and (c) show the differences in predicted drug sensitivity to bortezomib treatment between the responder (i.e., sensitive) group and non-responder (i.e., resistant) group using BT+SVR+S, B1+RR, and B2+RR, respectively. (d) shows the ROC curves of the prediction algorithms, which reveal the proportion of true positives compared to the proportion of false positives. ROC = receiver operating characteristic.

in each run (see the results colored in red in Table 11). These highly statistically significant results show the good performance of the BT+SVR+S algorithm (cf. Table 10).

Figures 7(a), 7(b), and 7(c) show the predictions of BT+SVR+S, B1+RR, and B2+RR, respectively, on the target test set in the first run. BT+SVR+S in Figure 7(a) yielded a highly statistically significant result ($P = 1195 \times 10^{-7}$ from a two-sample t-test). The result of B1+RR in Figure 7(b) was statistically significant with $P = 261 \times 10^{-5}$ from a two-sample t-test. B2+RR in Figure 7(c) yielded a statistically significant result ($P = 3381 \times 10^{-6}$ from a two-sample t-test). In Figure 7(d), the ROC curves reveal AUC values of 0.673, 0.614, and 0.641 for BT+SVR+S, B1+RR, and B2+RR, respectively.

Figure 8 shows that our prediction algorithms BT+SVR+S and BT+SVR+L outperform the baseline prediction algorithms with respect to the MAUC.

V. DISCUSSION

Our experimental results show that our proposed approaches significantly outperform the existing approach [13]. Furthermore, our proposed approaches are well-suited for a wide range of tasks, such as integration of different types of omics data to increase the accuracy of inferring gene regulatory networks (GRN) [41]–[45], and integration of different cancer data to enhance the performance of drug sensitivity prediction.

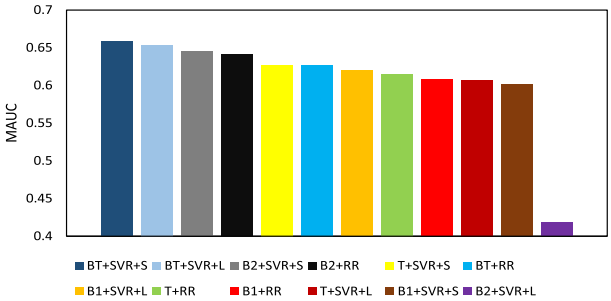

FIGURE 8. The Mean AUC (MAUC) values of the twelve bortezomib sensitivity prediction algorithms for multiple myeloma patients where the target training set and auxiliary dataset pertaining to multiple myeloma patients and non-small cell lung cancer patients, respectively, are used. The algorithms are ranked from left to right where the leftmost algorithm has the highest MAUC and the rightmost algorithm has the lowest MAUC.

In our work, the labels of training data are continuous values and the predicted labels (i.e., predictions) of target testing examples are also continuous values. However, the true labels of the target testing examples are categorical, which are either "sensitive" or "resistant." As in [13], the mapping of the predicted continuous values to the true categorical labels was performed using the ROCR package [46]. The details of this mapping algorithm can be found in [47]. In a nutshell, the mapping algorithm sorts the predicted continuous values obtained from a prediction algorithm in increasing order. The mapping algorithm works iteratively by examining one value at a time, from the smallest to the largest value. When examining a particular value v, the mapping algorithm labels v and all the values greater than or equal to v as "resistant" (i.e., positive) and all the values smaller than v as "sensitive" (i.e., negative). The mapping algorithm compares these "resistant" and "sensitive" labels with the corresponding true labels in the target test set to build a confusion matrix. The true positive rate (TPR) and false positive rate (FPR) with respect to the value v are then calculated and plotted. After all the predicted continuous values are examined, multiple points are plotted, where the x-coordinate of a point is a FPR and the y-coordinate of the point is a TPR. These points constitute the ROC curve of the prediction algorithm.

The biological rationale behind the good results of our approaches is that combining cancer drugs is often used to achieve enhanced therapeutic efficacy in a treatment [48]. For example, docetaxel (a chemotherapy drug) is used to treat breast cancer in combination with other specific chemotherapy drugs [35], [49]. The bortezomib and docetaxel combination has been used as a therapy for breast cancer [50], [51]. Hence, the task of predicting bortezomib sensitivity in multiple myeloma patients is closely related to the task of predicting docetaxel sensitivity in breast cancer patients, where closeness plays an important role in machine learning. For example, suppose we are given an unseen example (i.e., a testing example). If the unseen example has an expression profile closer to a training example with the corresponding response (i.e., drug IC_{50} value), then the unseen example is most likely to have a response closer to the response

associated with the training example. The same holds for combining bortezomib and cisplatin, which clinically led to synergistic killing of head and neck squamous cell carcinoma (HNSCC) cells [52]. In addition, erlotinib plus bortezomib showed a synergistic antitumor activity against the H460 non-small cell lung cancer (NSCLC) cell line [53].

In the proposed approaches, we assumed that the number of features (i.e., genes) in the target training set is greater than the number of features in an auxiliary dataset. Then, the top q (or n) features in the target test set are selected using the highest q statistical leverage scores computed on the target training set. However, if the number of features in the auxiliary dataset is greater than the number of features in the target training set (like the cases for triple-negative breast cancer patients and non-small cell lung cancer patients), then we select the top q features from the auxiliary dataset using the highest q statistical leverage scores computed from the auxiliary dataset, where q equals the number of features in the target training set, and no further feature selection is performed on the target training and target test sets.

In this work, differences in distributions between the target training data and auxiliary data have contributed to the degraded performance on the target test set for the prediction algorithms employing the second baseline approach (B2), which does not have a transfer learning mechanism. It is worth mentioning that we also assessed the performance of other machine learning algorithms, including random forests [54], support vector regression with a polynomial kernel of degree 2, and support vector regression with a Gaussian kernel. However, they exhibited poor performance; consequently, their results are not included in this paper.

VI. CONCLUSIONS

In this paper, we present two approaches to improve drug sensitivity prediction, namely a transfer learning approach and a boosted transfer learning approach. The transfer learning approach works by (1) performing feature selection to balance the number of features; (2) changing the representation of auxiliary data of a related task to a new representation that is closer to target training data; and (3) combining the target training data with the auxiliary data, and using the combined result as input to a standard machine learning algorithm. The boosted transfer learning approach boosts the performance of the transfer learning approach using a modified version of AdaBoost.

The proposed approaches employ two machine learning algorithms, namely support vector regression and ridge regression. Our experimental results demonstrate the stability of the proposed transfer learning approaches. Our approaches outperform the baseline approaches including an existing approach [13] as measured by their higher and statistically significant AUC scores.

In future work we plan to (1) extend the transfer learning approaches proposed here to handle auxiliary data from multiple related tasks simultaneously; (2) collaborate with domain experts, where we leverage signaling pathways to improve the prediction performance on a drug sensitivity prediction task; and (3) adopt new feature representation methods to improve the proposed transfer learning approaches for other drug sensitivity prediction tasks.

REFERENCES

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," *CA, Cancer J. Clinicians*, vol. 66, no. 1, pp. 7–30, 2016.

[2] A. Kamb, S. Wee, and C. Lengauer, "Why is cancer drug discovery so difficult?" *Nature Rev. Drug Discovery*, vol. 6, no. 2, pp. 115–120, 2007.

[3] V. Marx, "Cancer: A most exceptional response," *Nature*, vol. 520, no. 7547, pp. 389–393, 2015.

[4] N. C. Turner and J. S. Reis-Filho, "Genetic heterogeneity and cancer drug resistance," *Lancet Oncol.*, vol. 13, no. 4, pp. e178–e185, 2012.

[5] D. M. Roden and A. L. George, Jr., "The genetic basis of variability in drug responses," *Nature Rev. Drug Discovery*, vol. 1, no. 1, pp. 37–44, 2002.

[6] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nature Rev. Genet.*, vol. 16, no. 6, pp. 321–332, 2015.

[7] T. Turki and Z. Wei, "A noise-filtering approach for cancer drug sensitivity prediction," *CoRR*, Dec. 2016. [Online]. Available: http://arxiv.org/abs/1612.00525

[8] F. Sanchez-Garcia *et al.*, "Integration of genomic data enables selective discovery of breast cancer drivers," *Cell*, vol. 159, no. 6, pp. 1461–1475, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0092867414013750

[9] P. Zhang and V. Brusic, "Mathematical modeling for novel cancer drug discovery and development," *Expert Opinion Drug Discovery*, vol. 9, no. 10, pp. 1133–1150, 2014.

[10] W. Du and O. Elemento, "Cancer systems biology: Embracing complexity to develop better anticancer therapeutic strategies," *Oncogene*, vol. 34, no. 25, pp. 3215–3225, 2015.

[11] T. Turki and Z. Wei, "Learning approaches to improve prediction of drug sensitivity in breast cancer patients," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Orlando, FL, USA, Aug. 2016, pp. 3314–3320. [Online]. Available: http://dx.doi.org/10.1109/EMBC.2016.7591437

[12] G. Riddick *et al.*, "Predicting *in vitro* drug sensitivity using random forests," *Bioinformatics*, vol. 27, no. 2, pp. 220–224, 2011. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/27/2/220.abstract

[13] P. Geeleher, N. J. Cox, and R. S. Huang, "Clinical drug response can be predicted using baseline gene expression levels and *in vitro* drug sensitivity in cell lines," *Genome Biol.*, vol. 15, no. 3, p. R47, 2014.

[14] J. C. Costello *et al.*, "A community effort to assess and improve drug sensitivity prediction algorithms," *Nature Biotechnol.*, vol. 32, no. 12, pp. 1202–1212, 2014.

[15] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[16] M. P. Menden *et al.*, "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties," *PLoS ONE*, vol. 8, no. 4, pp. e61318-1–e61318-7, 04 2013. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0061318

[17] H. S. Bhatt, R. Singh, M. Vatsa, and N. K. Ratha, "Improving cross-resolution face matching using ensemble-based co-transfer learning," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5654–5669, Dec. 2014.

[18] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, p. 9, 2016. [Online]. Available: http://dx.doi.org/10.1186/s40537-016-0043-6

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002. [Online]. Available: http://dl.acm.org/citation.cfm?id=1622407.1622416

[20] M. W. Mahoney and P. Drineas, "CUR matrix decompositions for improved data analysis," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 3, pp. 697–702, 2009. [Online]. Available: http://www.pnas.org/content/106/3/697.abstract

[21] R. Batuwita and V. Palade, "*microPred*: Effective classification of pre-miRNAs for human miRNA gene prediction," *Bioinformatics*, vol. 25, no. 8, pp. 989–995, 2009.
[22] Y. B. Marques, A. de Paiva Oliveira, A. T. R. Vasconcelos, and F. R. Cerqueira, "Mirnacle: Machine learning with SMOTE and random forest for improving selectivity in pre-miRNA ab initio prediction," *BMC Bioinformatics*, vol. 17, no. 18, p. 53, 2016.
[23] M. Nakamura, Y. Kajiwara, A. Otsuka, and H. Kimura, "LVQ-SMOTE—Learning vector quantization based synthetic minority over–sampling technique for biomedical data," *BioData Mining*, vol. 6, no. 1, p. 16, 2013.
[24] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, "Relative-error *CUR* matrix decompositions," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 2, pp. 844–881, 2008.
[25] A. Bodor, I. Csabai, M. W. Mahoney, and N. Solymosi, "rCUR: An R package for CUR matrix decomposition," *BMC Bioinform.*, vol. 13, p. 103, May 2012. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-13-103
[26] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997. [Online]. Available: http://dx.doi.org/10.1006/jcss.1997.1504
[27] H. Drucker, "Improving regressors using boosting techniques," in *Proc. 14th Int. Conf. Mach. Learn. (ICML)*, San Francisco, CA, USA, 1997, pp. 107–115. [Online]. Available: http://dl.acm.org/citation.cfm?id=645526.657132
[28] L. Zhong, J. T. L. Wang, D. Wen, V. Aris, P. Soteropoulos, and B. A. Shapiro, "Effective classification of microrna precursors using feature mining and adaboost algorithms," *Omics: J. Integrative Biol.*, vol. 17, no. 9, pp. 486–493, 2013.
[29] T. Turki, M. Ihsan, N. Turki, J. Zhang, U. Roshan, and Z. Wei, *Top-k Parametrized Boost*. Cham, Switzerland: Springer, 2014, pp. 91–98. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-13817-6_10
[30] K. Neubert *et al.*, "The proteasome inhibitor bortezomib depletes plasma cells and protects mice with lupus-like disease from nephritis," *Nature Med.*, vol. 14, no. 7, pp. 748–755, 2008.
[31] A. Paramore and S. Frantz, "Bortezomib," *Nature Rev. Drug Discovery*, vol. 2, no. 8, pp. 611–612, 2003.
[32] A. Brazma *et al.*, "ArrayExpress—A public repository for microarray gene expression data at the EBI," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 68–71, 2003.
[33] L. Venkova *et al.*, "Combinatorial high-throughput experimental and bioinformatic approach identifies molecular pathways linked with the sensitivity to anticancer target drugs," *Oncotarget*, vol. 6, no. 29, pp. 27227–27238, 2015.
[34] M. J. Garnett *et al.*, "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature*, vol. 483, no. 7391, pp. 570–575, 2012.
[35] H. Joensuu *et al.*, "Adjuvant docetaxel or vinorelbine with or without trastuzumab for breast cancer," *New England J. Med.*, vol. 354, no. 8, pp. 809–820, Feb. 2006.
[36] M. Aujla, "Chemotherapy: Treating older breast cancer patients," *Nature Rev. Clin. Oncol.*, vol. 6, no. 6, p. 302, Jun. 2009.
[37] L. C. Wienkers and T. G. Heath, "Predicting *in vivo* drug interactions from *in vitro* drug discovery data," *Nature Rev. Drug Discovery*, vol. 4, no. 10, pp. 825–833, 2005.
[38] O. Bousquet and A. Elisseeff, "Stability and generalization," *J. Mach. Learn. Res.*, vol. 2, pp. 499–526, Mar. 2002.
[39] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi, "General conditions for predictivity in learning theory," *Nature*, vol. 428, no. 6981, pp. 419–422, 2004.
[40] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011. [Online]. Available: http://doi.acm.org/10.1145/1961189.1961199
[41] F. Petralia, P. Wang, J. Yang, and Z. Tu, "Integrative random forest for gene regulatory network inference," *Bioinformatics*, vol. 31, no. 12, pp. i197–i205, 2015.
[42] T. Turki and J. T. L. Wang, "A new approach to link prediction in gene regulatory networks," in *Proc. Int. Conf. Intell. Data Eng. Autom. Learn.*, 2015, pp. 404–415.
[43] T. Turki, W. Bassett, and J. T. L. Wang, "A learning framework to improve unsupervised gene network inference," in *Machine Learning and Data Mining in Pattern Recognition*. Cham, Switzerland: Springer, 2016, pp. 28–42.
[44] Y. Abduallah, T. Turki, K. Byron, Z. Du, M. Cervantes-Cervantes, and J. T. L. Wang, "Mapreduce algorithms for inferring gene regulatory networks from time-series microarray data using an information-theoretic approach," *BioMed Res. Int.*, vol. 2017, Jan. 2017, Art. no. 6261802.
[45] T. Turki, J. T. L. Wang, and I. Rajikhan, "Inferring gene regulatory networks by combining supervised and unsupervised methods," in *Proc. 15th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2016, pp. 140–145.
[46] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "ROCR: Visualizing classifier performance in R," *Bioinformatics*, vol. 21, no. 20, pp. 3940–3941, 2005.
[47] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, 1st ed. Boston, MA, USA: Pearson, 2005.
[48] P. Nowak-Sliwinska *et al.*, "Optimization of drug combinations using feedback system control," *Nature Protocols*, vol. 11, no. 2, pp. 302–315, 2016.
[49] M. Marty *et al.*, "on, A. Lluch, "Randomized phase II trial of the efficacy and safety of trastuzumab combined with docetaxel in patients with human epidermal growth factor receptor 2–positive metastatic breast cancer administered as first-line treatment: The M77001 study group," *J. Clin. Oncol.*, vol. 23, no. 19, pp. 4265–4274, 2005.
[50] K.-F. Chen *et al.*, "CIP2A mediates effects of bortezomib on phospho-Akt and apoptosis in hepatocellular carcinoma cells," *Oncogene*, vol. 29, no. 47, pp. 6257–6266, 2010.
[51] A. Awada *et al.*, "Bortezomib/docetaxel combination therapy in patients with anthracycline-pretreated advanced/metastatic breast cancer: A phase I/II dose-escalation study," *Brit. J. Cancer*, vol. 98, no. 9, pp. 1500–1507, 2008.
[52] C. Li, R. Li, J. R. Grandis, and D. E. Johnson, "Bortezomib induces apoptosis via Bim and Bik up-regulation and synergizes with cisplatin in the killing of head and neck squamous cell carcinoma cells," *Molecular Cancer Therapeutics*, vol. 7, no. 6, pp. 1647–1655, 2008.
[53] T. J. Lynch *et al.*, "A randomized phase 2 study of erlotinib alone and in combination with bortezomib in previously treated advanced non-small cell lung cancer," *J. Thoracic Oncol.*, vol. 4, no. 8, pp. 1002–1009, 2009.
[54] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://dx.doi.org/10.1023/A:1010933404324

TURKI TURKI received the B.S. degree in computer science from King Abdulaziz University, the M.S. degree in computer science from NYU.POLY, and the Ph.D. degree in computer science from the New Jersey Institute of Technology. He is currently a Faculty Member with the Department of Computer Science, King Abdulaziz University, Saudi Arabia. He is also working with Prof. Jason T. L. Wang and Prof. Zhi Wei on several biomedicine related projects. He has authored several papers in these areas. His research interests include algorithms, machine learning, data mining, big data analytics, health informatics, bioinformatics, computational biology, and social networks. He received a scholarship from King Abdulaziz University. He is supported by King Abdulaziz University and Saudi Arabian Cultural Mission.

ZHI WEI received the M.S. degree in computer science from Rutgers University, New Brunswick, NJ, USA, in 2004 and the Ph.D. degree in bioinformatics from the University of Pennsylvania, Philadelphia, PA, USA, in 2008. He is currently an Associate Professor of Computer Science and an Associate Professor of Statistics (joint appointment) with New Jersey Institute of Technology. His works have been published in prestigious journals including *Nature*, *Nature Medicine*, *Journal of the American Statistical Association*, *Biometrika*, *AOAS*, *American Journal of Human Genetics*, *PLoS Genetics*, *Bioinformatics*, and *Biostatistics*, and top data mining and machine learning conferences including NIPS, KDD, and ICDM. His research focuses on data science and advanced analytics using statistical and machine learning, with application to a broad range of fields including biology, genetics, medicine, digital marketing, social media, and real estate. He is also an Editorial Board Member of *BMC Bioinformatics, BMC Genomics, PLoS ONE, Frontiers in Bioinformatics and Computational Biology*, and *Frontiers in Applied Genetic Epidemiology*.

JASON T. L. WANG received the B.S. degree in mathematics from National Taiwan University, Taipei, Taiwan, and the Ph.D. degree in computer science from Courant Institute of Mathematical Sciences, New York University in 1991. He is currently a Professor of Computer Science and Bioinformatics with New Jersey Institute of Technology and the Director of the university's Data and Knowledge Engineering Laboratory. He has authored 9 books and over 150 papers in these and related areas. His research interests include data mining, machine learning, big data, computational biomedicine, and smart cities. He has served on the program committees of over 200 national and international conferences and the editorial boards of several journals including *Knowledge and Information Systems* (Springer), *Intelligent Data Analysis* (IOS Press, Amsterdam), and *Information Systems* (Elsevier).