Malik *et al. BMC Genomics* (2021) 22:214
https://doi.org/10.1186/s12864-021-07524-2

BMC Genomics

RESEARCH ARTICLE Open Access

Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer

Vidhi Malik, Yogesh Kalakoti and Durai Sundar*

Abstract

Background: Survival and drug response are two highly emphasized clinical outcomes in cancer research that directs the prognosis of a cancer patient. Here, we have proposed a late multi omics integrative framework that robustly quantifies survival and drug response for breast cancer patients with a focus on the relative predictive ability of available omics datatypes. Neighborhood component analysis (NCA), a supervised feature selection algorithm selected relevant features from multi-omics datasets retrieved from The Cancer Genome Atlas (TCGA) and Genomics of Drug Sensitivity in Cancer (GDSC) databases. A Neural network framework, fed with NCA selected features, was used to develop survival and drug response prediction models for breast cancer patients. The drug response framework used regression and unsupervised clustering (K-means) to segregate samples into responders and non-responders based on their predicted IC50 values (Z-score).

Results: The survival prediction framework was highly effective in categorizing patients into risk subtypes with an accuracy of 94%. Compared to single-omics and early integration approaches, our drug response prediction models performed significantly better and were able to predict IC50 values (Z-score) with a mean square error (MSE) of 1.154 and an overall regression value of 0.92, showing a linear relationship between predicted and actual IC50 values.

Conclusion: The proposed omics integration strategy provides an effective way of extracting critical information from diverse omics data types enabling estimation of prognostic indicators. Such integrative models with high predictive power would have a significant impact and utility in precision oncology.

Keywords: Multi-omics integration, Deep learning, Feature selection, Survival outcomes and drug response prediction

Background

Breast cancer has ranked among the most prevalent cancer type with a rate as high as 25.8 per 100,000 women in the Indian subcontinent [1]. Global and local studies have also reported a gradual increase in cancer-associated mortality in the region [2–4]. These metrics suggest an urgent need to devise robust knowledge-based prognostic systems that can generate phenotypic estimates for an individual. To address this issue, personalized medicine aims to provide the most effective treatment strategy based on the patient's medical history, genomic characteristics, and response to therapy [5, 6]. Substantial genomic characterization has been conducted in the past decade to support the idea, leading to clinically relevant molecular subtyping [7–9]. Still, out of all the pharmaceutical agents pitched in clinical setups, only about 15% demonstrate sufficient safety and potency to gain any sort of regulatory consent [10, 11].

* Correspondence: sundar@dbeb.iitd.ac.in
DAILAB, Department of Biochemical Engineering and Biotechnology, Indian Institute of Technology (IIT) Delhi, New Delhi, India

This implies the limitations in the current understanding of cancer complexity and the need for models that efficiently simulate the diversity of human tumor biology in a preclinical arrangement. With the advent of high-throughput data profiling technologies in the past decade, there is an opportunity for us to improve our understanding of the multi-layered molecular basis of cancer.

Large scale collaborative efforts such as The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium have led to numerous reports related to interim analyses of gene expression, somatic mutation, copy number variation (CNV) and protein expression data in the literature [12–16]. While it has allowed us access to a massive set of curated data, it is essential to address the long-standing bottleneck of omics integration to understand cancer prognosis and phenotype better. Multi-omics data integration has emerged as a promising approach for the prediction of clinical outcomes and identification of biomarkers in several cancer studies [17–20]. Modeling of survival and drug response clinical outcomes in cancer research can prove as steppingstones in the direction of personalized therapy. Omics integration allows us to analyze the human genome at multiple levels of complexity simultaneously and extract meaningful conclusions. Linear prediction models for such analysis often break down due to the steep dimensionality and heterogeneity associated with omics datasets. Hence, a refined integrative approach to handle these diverse datasets coherently is required.

Here, we address the challenge of building robust multi-omics integration based neural network models to predict clinical outcomes and response of an individual to a panel of 100 drugs. Neighbourhood component analysis (NCA) based feature selection algorithm was employed separately on each omics data to select high weighted features that were then fed into neural network-based classifier and regressor model to build multi-omics based integrative survival and drug response prediction models for breast cancer. These type of multi-omics integration based prediction models will not only help the physicians make rational chemotherapeutic decisions but also to understand the driving nodes in the cancer machinery.

Results

We trained breast cancer datasets from TCGA and GDSC to generate robust survival and drug response prediction models. We used 10-fold cross-validation for the survival prediction model and 5-fold cross-validation for drug response prediction model to better tune the hyperparameters. Ultimately, two neural network models were chosen to generate drug responses and survival estimates for the patients in validation sets. The corresponding performance metrics were calculated based on the losses incurred in the respective models.

Multi-omics integration improves survival prediction in BRCA patients

The NCA selected 246 six-omics feature set along with clinical features like age, gender, days to the last follow-up, pathologic stage, the number of affected lymph nodes, tumor stage, lymph node metastasis, metastatic stage and histological type were fed into neural network-based survival prediction model to classify the patients into two classes, i.e., high-risk class and low-risk class. The feed-forward neural network model was trained with two hidden layers of 7 nodes in each layer and an output layer of two neurons to classify patients into two survival classes. 10-fold cross-validation of neural-network along with optimization of regularization term and hidden layers architecture was performed using BayesOpt. The final layout of the neural network model consisted of two hidden layers (with seven nodes) and two output classes with a regularization term set to 0.9999. After multiple iterations of Bayesian optimization, 'trainscg' was selected as a training function that adopted a scaled conjugant gradient method to update weights and bias; cross-entropy was used as the performance evaluation function.

The survival prediction model was able to classify the patients into two survival classes – high-risk and low-risk, with a prediction accuracy of 94% (Fig. S1A). The prediction accuracies of training, validation and test dataset were 93.5, 93.7 and 98.1%, respectively. This clearly signified that the overfitting of the neural network model was successfully avoided here. AUROC (Area Under the Receiver Operating Characteristics) value of 0.98 was observed for both the classes, i.e., low-risk and high-risk, that showed the ability of prediction model to classify patients into two classes (Fig. S1B) efficiently. The performance of the model was also evaluated by calculating various other parameters like sensitivity, specificity, precision, false-positive rate, F1 Score, Matthews Correlation Coefficient and Kappa (Table 1). The value of all the parameters showed good ability of the prediction model to distinguish between two survival classes.

External validation of the multi-omics integration-based survival prediction model was performed by using single-omics and five-omics dataset of TCGA BRCA patients that were excluded for the training of model due to unavailability of all six-omics data (Table 2). The performance of the model with single-omics data or five-omics data as input for validation was not comparable to the performance of our model. It was observed that six-omics integrated data was able to predict both high-risk and low-risk individuals with good prediction accuracy.

Table 1 Performance of neural network-based classifier for survival prediction of BRCA patients

	Sensitivity	Specificity	Precision	False positive rate	F1 Score	Matthews Correlation Coefficient	Kappa
Parameters	0.95	0.92	0.93	0.07	0.94	0.87	0.87

However, when single-omics or five-omics data was given as input for external validation, the model was not able to predict high-risk individuals correctly due to class imbalance in dataset available for breast cancer. It was observed that single-omics input classified all individuals as low-risk class, therefore correctly predicting low-risk patients with 100% prediction accuracy, but failed to predict for high-risk class. Similarly, for five-omics input feed, the model was able to predict high-risk individuals correctly with prediction accuracy ranging from 0 to 10% only and that of low-risk individuals with prediction accuracies ranging from 83 to 100%. This showed that adding more layers of omics information would aid in better prediction. Integrating different omics data types improved the performance of the predictive models over the traditional single-omics approach as the highest accuracy was achieved with the model including all the omics-types.

Multi-omics signature predicts drug response in BRCA cell lines

The drug response prediction model was trained on BRCA cell lines for 212 drugs initially; however, some drugs were filtered out later due to poor performance of models for these drugs. The final regression model was trained for 42 cell lines and 100 drug molecules. The robustness of the regression model using the features optimally selected using NCA was demonstrated using various performance metrics. The optimal neural network regressor had two hidden layered architecture with 11 nodes in both the layers. Levenberg-Marquardt backpropagation, which is the fastest backpropagation algorithm, was used as a training function to propagate the losses incurred back to the network and reconfigure the weights. In addition to this, Bayesian optimization of the regularization term was performed with the final value set to 0.3743 with 5-fold cross-validation to avoid overfitting the model. Mean squared error (MSE) was used as a performance evaluation function of the neural network regression model. The drug response prediction regression model predicted IC50 values for each drug with MSE of 1.154 and an overall regression value of 0.92, which showed the linear relationship between predicted and actual IC50 values. This was followed by unsupervised clustering (K-means) of drug responses to segregate the samples into responders and non-responders based on their IC50 values. The clustered IC50 values for the first twenty drugs showed that a common threshold value for all of the drugs could not be used as each drug has its unique distribution of responses (Fig. S2). The best validation performance reported in terms of MSE as 0.66 is remarkable, considering the small number of datasets. Moreover, calculation of IC50 thresholds was also consistent among the two methods (K-means and waterfall) as quantified by a strong correlation of 0.91 (Fig. S3-B). However, the classification metrics lagged while using thresholds calculated by waterfall analysis (Fig. S4).

Drugs such as Dabrafenib, Mitomycin, Olaparib and Ruxolitinib performed exceptionally well on almost all the cell lines tested. Figure 1 shows the performance of drug response in terms of accuracy, specificity, and sensitivity corresponding to all the drugs as well as all the cell lines. It is evident from the results that most of the drugs performed at par or even outperform similar drug response prediction models [21]. These traditional methods employed Elastic Net and SVM models for drug response on GDSC datasets instead of Deep learning frameworks. Hence, their average sensitivity and specificity values were averaged around 0.75 and 0.78 respectively. Even with a large ensemble of tested drugs (100), the average sensitivity and specificity values reported here averaged around 0.80 (Fig. 1a and c). Individual drugs were analyzed for their contribution to the

Table 2 External validation prediction accuracy of our multi-omics integration-based survival prediction model for BRCA patients

External Validation with Single-omics data and clinical features as input to model			External validation with five-omics data and clinical features as input to model		
Datatype	**Samples**	**Prediction accuracy**	**Datatype**	**Samples**	**Prediction accuracy**
RNA	561	85.7%	five-omics data excluding Protein	59	78.0%
Protein	397	85.1%	five-omics data excluding Mutation	41	73.2%
Mutation	493	85.8%	five-omics data excluding miRNA	103	90.3%
miRNA	241	82.6%	five-omics data excluding Methylation	111	77.5%
CNV	548	85.8%	five-omics data excluding CNV	3	66.7%
Methylation	268	86.6%	five-omics data excluding RNA	0	

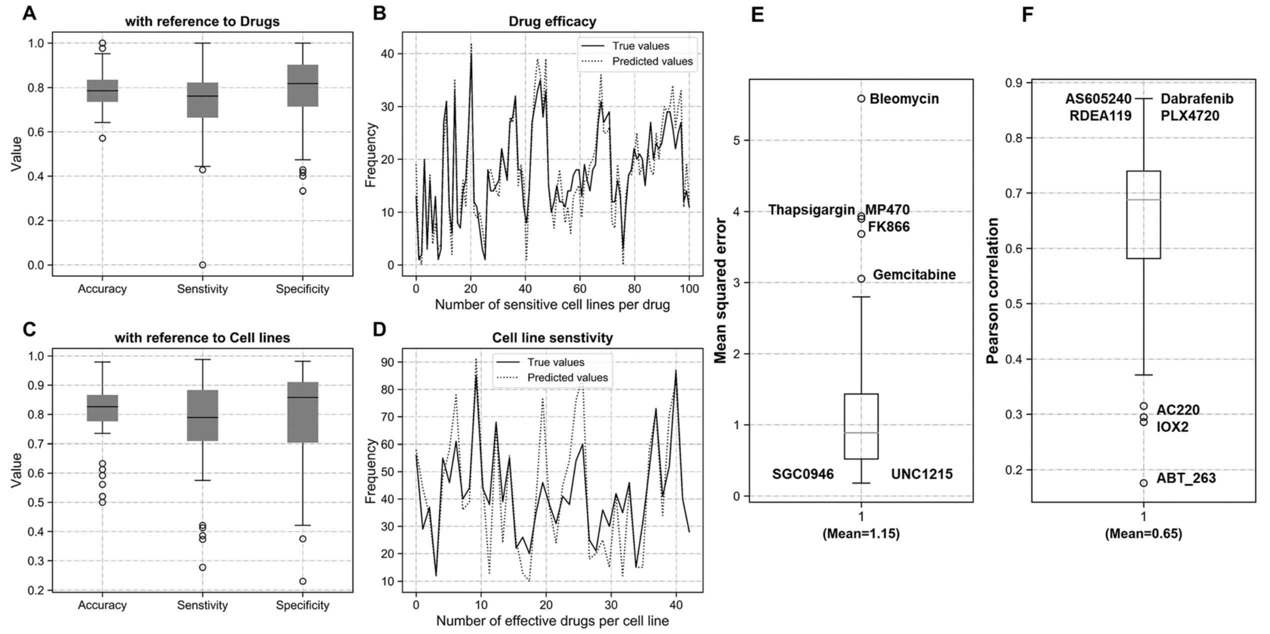

Fig. 1 Performance of drug response model for BRCA cell lines. Box plot showing accuracy, sensitivity and specificity of model for all drugs (**a**) and all cell lines (**c**). Scatter plot showing frequency of (**b**) sensitive cell lines per drug molecule and (**d**) effective drugs for each cell line. **e** Mean squared error and (**f**) Pearson's correlation for drug responses from the model for individual drugs

overall performance metrics that led to the discovery of certain outliers like Bleomycin, Gemcitabine, Thapsigargin, MP470 and FK866 (Fig. 1e-f). While these drugs negatively affected the model performance, drugs such as Dabrafenib, AS605240, RDEA119 and PLX4720 depicted exceptional correlation with the actual drug-responses across the test set (Fig. 1f and 2).

Proposed model performs better than similar approaches

The proposed breast cancer survival and drug response prediction models were compared with one survival prediction method and two drug response prediction methods (Table 3). For survival prediction, a similar study on BRCA patients reported accuracy and AUC values of 0.73 and 0.79 respectively [22]. As a direct comparison, our proposed model performed significantly better for the same metrics with prediction accuracy of 0.94 and AUC value of 0.98.

On the other hand, SVM-based and late-integration based models have been extensively used to predict drug responses in cancer patients [23]. On similar lines, an SVM model was built *in-house* using NCA selected features for comparative analysis. SVM parameters were optimized using grid search on a range of cost and gamma that were adapted from a similar SVM based study [23]. A value of 10 for cost and 0.5 for gamma was found to be optimal for predicting drug responses. Similarly, MOLI was employed to predict drug responses for our datasets (https://github.com/hosseinshn/MOLI) [19]. However, only a subset of the drugs (Docetaxel and Gemcitabine) could be compared as MOLI was limited to only a few drugs. The proposed method was able to outperform the competition on both the instances, reinforcing the effectiveness of the proposed method (Table 3).

Moreover, to gauge the effectiveness of the proposed drug response model, a measure of external validation was necessary. Drug response data for TCGA breast cancer (BRCA) patients was available from a similar study [24]. TCGA identifiers and drug responses for four drugs (Vinblastine, Gemcitabine, Tamoxifen, Docetaxel) were extracted from the dataset. mRNAseq, methylation, CNV and miRNAseq data for the selected TCGA identifiers was processed and passed through the saved neural network. The predicted drug responses, binarized using previously calculated drug thresholds, were fairly accurate with about 0.79 accuracy for Docetaxel (24 patients) and 0.5 for Tamoxifen (11 patients). For Vinblastine and Gemcitabine, the dataset of single patient for each drug was available to compare predictions of developed drug response prediction model. The developed model was able to predict drug response for Vinblastine and Gemcitabine correctly. Therefore, considering that the initial model was trained on cell lines, the overall external validation accuracy of 0.73 is

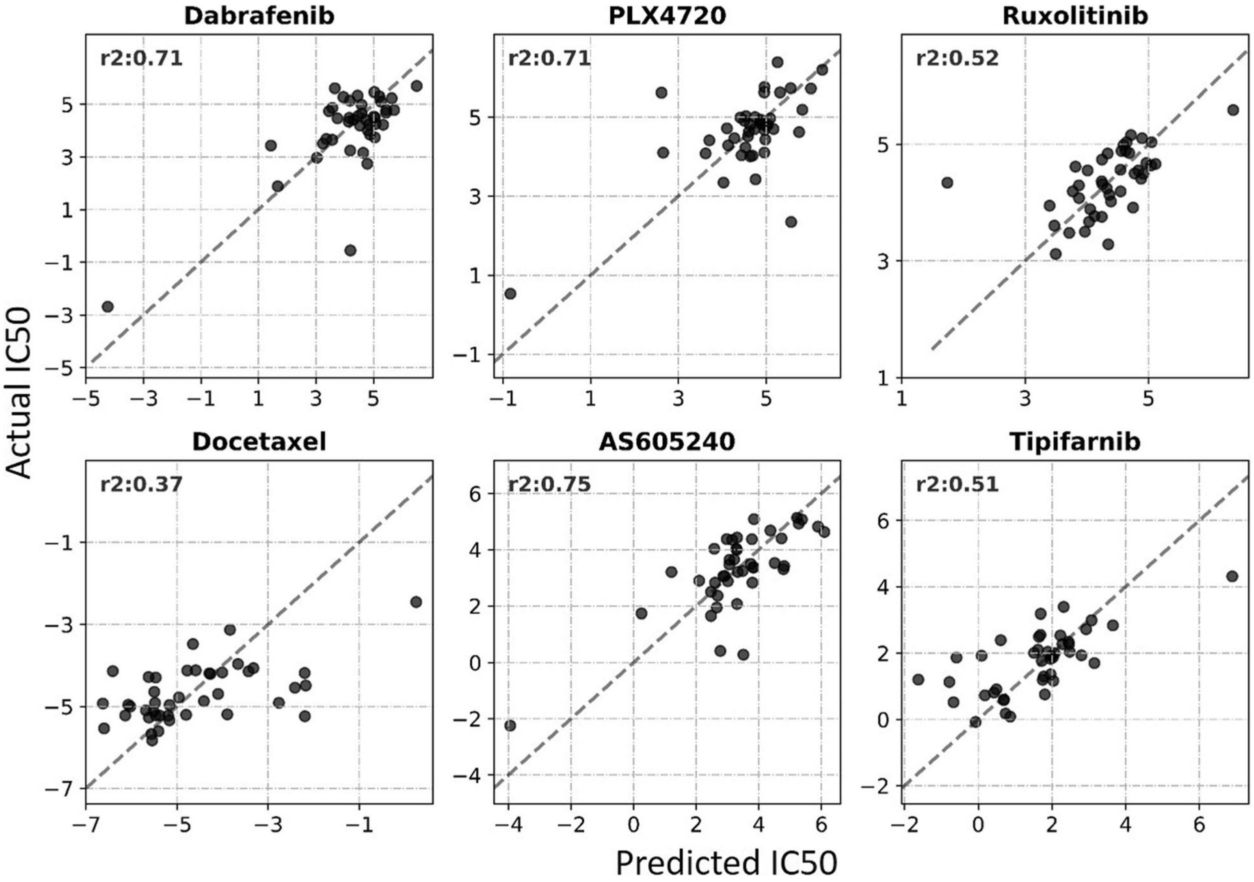

Fig. 2 The drug response model, trained on the multi-omics profile of cell lines has the capacity to predict response for 100 drugs. The figure evaluates the performance for some of the prominent cancer drugs in terms of R squared performance measure (Note: For an ideal case, all the points would lie on a straight line y = x (dashed) with *r*2 = 1)

consistent with internal validation and reinforces the effectiveness of the proposed method.

Biological significance of identified signature

Feature selection using NCA provided us with a set of genes that were weighted highly for their predictive potency. Therefore, Gene Set Enrichment Analysis (GSEA) was employed to calculate gene enrichment scores corresponding to every entity. Reactome knowledge database was used to carry out the analysis [25, 26]. Gene set screened from mRNA dataset for the survival prediction module revealed pathways and reactions that are critical for the patient's survival (Table S2). TP53 dependent transcription regulation, gene expression and DNA damage response were among the most significantly enriched pathways among all data types. The identified signature of survival and drug response prediction was also combined and mapped onto KEGG pathways using DAVID functional annotation tool [27, 28]. The identified pathway mainly consisted of cancer pathways and all major pathways whose dysregulation is well reported in cancer (Table S3).

Table 3 Comparison of the proposed survival and drug response prediction model with similar methods

Survival Prediction	**Accuracy**	**AUC**
C. Wang et al. [22]	0.79	0.93
Proposed method	**0.94**	**0.98**
Drug response prediction	**Docetaxel (AUC)**	**Gemcitabine (AUC)**
MOLI	0.67	0.71
SVM	0.63	0.69
Proposed method	**0.83**	**0.78**

Discussion

Robust classification of cancer patients into risk groups and having prior information about the possible drug responses will identify novel screening methods, prognostic factors, methods and perhaps guide the next steps in personalized therapies. In this study, the high prognostic accuracy of neural networks has been demonstrated owing to their capacity to model complex relationships among variables [29, 30].

Table 4 Biological significance of gene set that aid in prediction of BRCA survival and drug response

GENE	NAME	Reported Biomarker function
EFHD1	EF-Hand Domain Family Member D1	Part of digital RNA resistance signature to predict response to breast cancer therapy [31]
CDH1	Cadherin 1	CDH1 structural alterations as novel prognostic biomarker in gastric cancer patients [32] CDH1 gene as a prognostic biomarker in hepatocellular carcinoma [33]
PIK3CA	Phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha	PIK3CA is a predictive biomarker for use of alpelisib and fulvestrant in BRCA patients [34]
TP53		mutant p53 as a possible therapeutic target and biomarker for breast cancer [35]

For identification of probable prognostic biomarkers among the screened gene-pool, a ranking criterion was devised among the genes. The screening methodology (NCA) enabled us to rank the associated genes based on their predictive ability. Four genes, EFHD1, CDH1, PIK3CA and TP53, were identified by our feature selection algorithm that aid in prediction of both survival and drug response prediction of breast cancer patients. The role of these genes, to serve as prognostic/predictive biomarkers has already established in many cancer types (Table 4). EF-hand Domain Family Member D1 (EFHD1) is shown to be overexpressed in breast cancer and is reported to serve as a potent breast cancer-specific RNA signature [36]. Similarly, genetic and epigenetic alterations in E-Cadherin (CDH1) relates to aberrant expression and microsatellite instabilities in breast cancer patients have also been related to the incidence of breast cancer [37, 38]. Besides, Phosphatidylinositol 3-kinase (PIK3CA) and Tumor protein 53 (TP53) genes, which are two of the most mutated genes in breast cancer, were also shortlisted by the workflow [39, 40].

The drug response model captured the relationship between the patient's multi-omics profile and well-known breast cancer drugs such as Dabrafenib ($r2 = 0.71$), Gemcitabine ($r2 = 0.59$) and (AS605240) PI3K inhibitor ($r2 = 0.75$) among others with a high degree of confidence (Fig. 2). In addition to the omics types included in the study, the approach can be theoretically scaled for the integration of other omics types such as proteomics. Ambiguous data remains to be a hurdle in the way of these models being clinically acceptable. For example, patients who die of an unrelated cause or have a sparse follow-up will have to be incorporated accordingly into the model. A few alternatives to mitigate this issue is reported in the literature, but none of them have yet been successful [41, 42].

Conclusions

Survival statistics are one of the most important prognostic factors in breast cancer. However, it can be debated whether a response to therapy is also as detrimental to the patient's ultimate treatment routine. Probing the potential of cumulative analysis of survival prediction and response to therapy could open doors for practical solutions in improving therapy in cancer. Global genomic profiling of cancer cell line panels and patient-derived samples have contributed a lot in building risk-classification models and suggesting novel therapeutic measures. However, a large pool of drug compounds has not been assessed over the potential of available genomics data. With an increase in biological resources that capture disease characteristics such as genotype, phenotype and their associations, novel strategies are required to efficiently process this information and reveal critical insights for the disease. Here, we employed late integrative deep learning frameworks for building survival and drug response prediction models that performed at par with existing individual solutions.

We conclude that an artificial deep neural network, which is trained on the multi-omics signature of an individual, in tandem with its clinico-pathological factors, can not only segregate individual into low-risk and high-risk subgroups but also assist in screening a pool of drugs based on the sensitivity values corresponding to the patient under observation. The results reinforce the idea that an integrative approach can make more accurate and personalized decisions for drug administration and general treatment strategy.

Methods

General workflow

This workflow was designed to predict the survival outcome and drug response for a given BRCA patient, characterized by its multi-omics signature. The underlying assumption is data being independent and identically distributed. The workflow followed multiple feed-forward networks and dimensionality-reduction measures corresponding to every omics type. The features learned were clubbed together that served as an input to a regression and classification network for drug-response and survival prediction, respectively (Fig. 3).

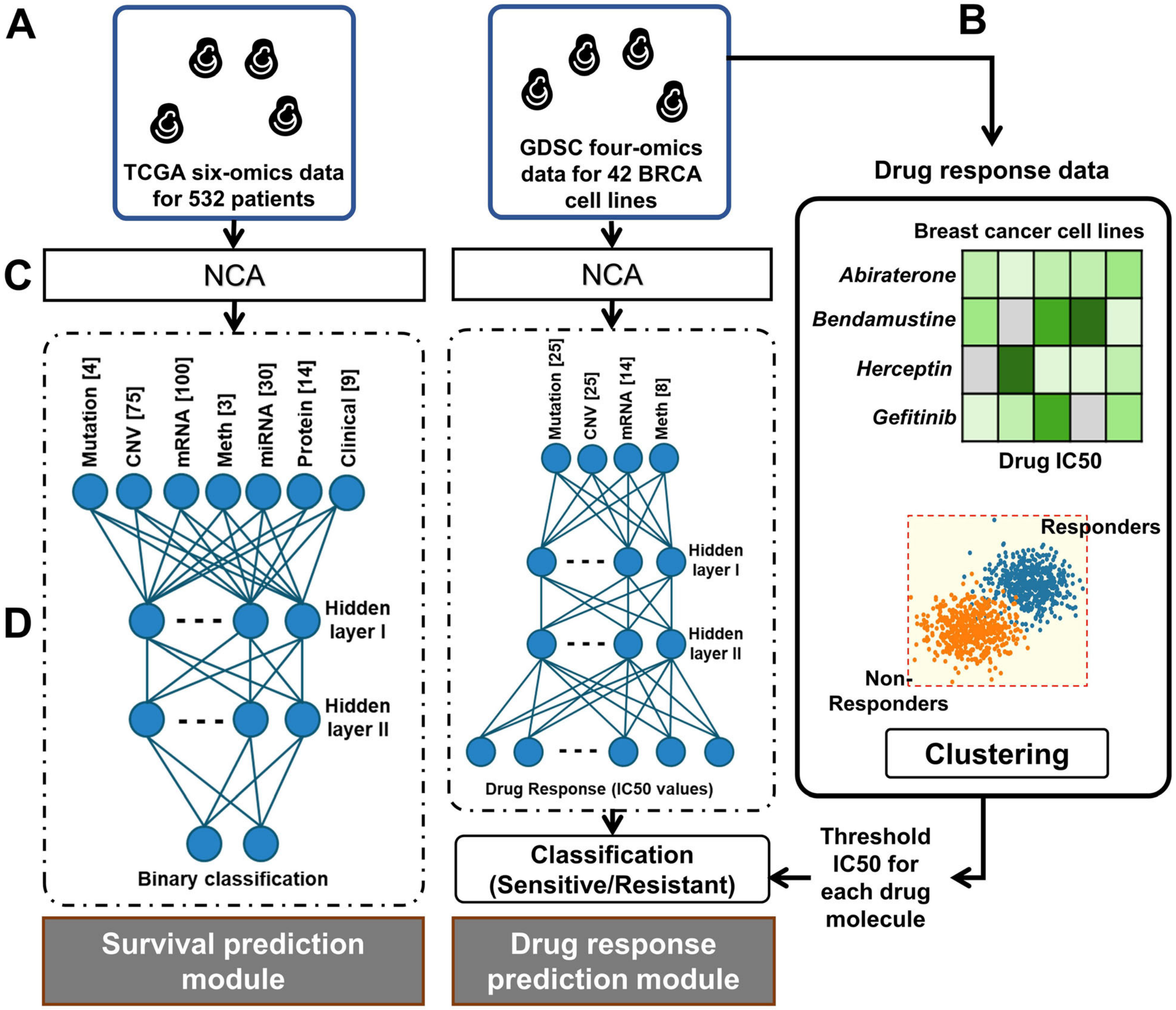

Fig. 3 Schematic of the general pipeline followed for survival and drug response prediction task. The flowchart depicting various steps followed during prediction models development including (**a**) data retrieval, (**b**) drug response processing, (**c**) omics data processing and (**d**) training/optimizing deep neural networks

Datasets

Two major resources were used for the analysis. Datasets for breast invasive carcinoma (BRCA) patients were retrieved from TCGA, whereas GDSC was used to source multi-omics as well as drug-response datasets for BRCA cell lines [43]. GDSC was preferred among other sources due to its broad spectrum of screened drugs.

Preprocessing TCGA breast cancer patient's data

TCGA BRCA multi-omics datasets, along with their clinical information was available for more than 1000 patients, including 1089, 977, 1097, 1078, 1093 and 887 patient's GISTIC2 CNV, mutation, methylation, miRNA, RNA and protein expression data respectively. The pre-processed TCGA dataset was obtained using FireBrowse utility (http://firebrowse.org). For RNA, z-scaled RSEM values of RNA expression were used and for miRNA log2-RPM values were retrieved. Protein expression and methylation data (β values) obtained from database were already scaled. Binary data was obtained for mutation of genes and GISTIC2 calculated CNV data was obtained directly from FireBrowse. The dataset was screened by filtering patients and features with more than 20% missing values. Further missing values in the omics dataset were imputed using R package impute [44]. An overlapping set of 314 patients was obtained for which all six-omics datasets along with their clinical information was available. The final processed data was observed to be class imbalanced. Therefore, an oversampling technique called Synthetic Minority Oversampling TEchnique (SMOTE) [45] was employed to balance the data that increases our sample set from 314 to 532.

Preprocessing breast cancer cell line data obtained from GDSC

The breast cancer cell lines omics data and drug response data were retrieved from the GDSC database. Already preprocessed data for all cell lines was obtained from the GDSC database followed by filtration step to filter out other cancer cell lines data and only breast cancer cell lines data was retained for the analysis. Binary data was obtained for gene mutations and already pre-processed β-values were obtained for methylation of CpG islands. RMA normalized basal expression level was obtained for RNA data and copy number values were obtained ranging from − 1 to 1, where 0 indicates normal copy number, − 1 and + 1 indicates loss and gain of copy of genes respectively. The overlapping set of 43 cell lines were selected for those drug responses for which at least 80% of drugs and all four omics datasets, namely CNV, methylation, mutation and mRNA, were available. Similar preprocessing was done to remove cell lines and features having more than 20% missing values. This reduced our sample set to 42 cell lines only. Remaining missing values in omics and drug response data were imputed using the impute R package. The next filter for genes was applied to screen out genes for which omics data was not available in the TCGA BRCA dataset.

Constructing representative gene sets

The high dimensionality of omics datasets remains a significant bottleneck in generating robust prediction models that are clinically relevant [46]. The goal of feature engineering here is to find an effective low-dimensional manifold of a given high dimensional dataset. Fortunately, biological processes are highly correlated and can be represented in a lower-dimensional sub-space [47, 48]. Many approaches, like Principal component analysis, Correspondence Analysis, Partial Triadic Analysis and Multiple co-inertia analysis, which are based on variance, correlation, inertia, eigenvalue among others have utilized this fact. They have been quite successful in this effort [49–51]. However, none of these commonly used approaches considers the effect of labels corresponding to the datasets. Due to this particular reason, we opted for Neighborhood Component Analysis (NCA), which is a supervised dimensionally reduction method for learning Mahal Nobis distance measure for k-nearest Neighbors [52].

Given an omics data set, $X = x_1, x_2, x_3, \ldots, x_n \in R^P$ and corresponding class labels $c_1, c_2, c_3, \ldots, c_n$, which is generally an $n \times p$ matrix with n observations (patients) and p variables (genes) corresponding to the measurements of mRNA and other omics datasets. NCA reduces the dimensions by restricting the quadratic distant metric to be low rank. The underlying distance metric can be defined as follows.

$$d(x, y) = (x - y)^T (z - y) = (Ax - Ay)^T (Ax - Ay) \quad (1)$$

Also, Leave One Out performance is utilized as test data is not available during training under the following objective function.

$$f(A) = \sum_i \sum_{j \in C_i} p_{ij} = \sum_i p_i \quad (2)$$

The ultimate features extracted corresponding to each omics type for building survival and drug response models are summarized in Table S1.

Multi-omics feature integration

A single multi-omics representation of different omics datasets was generated by employing a late integration approach, where the learned features for each of the omics types were concatenated before being fed into the neural networks. For instance, three single-omics input with three $m \times n$ feature matrices will result in a single $m \times 3n$ representation matrix after integration.

Survival prediction

A two-layer neural network model was constructed for the binary classification (survival) task using MATLAB ver. R2019b. Cross-entropy loss was employed for the optimization of the objective function. Losses were propagated back using scaled conjugate gradient backpropagation and the hyperparameters were optimized using Bayesian hyperparameter optimization (BayesOpt) [53]. Hyperbolic tangent, a symmetric activation function that provides mean-zero initial weights, was used in the hidden layers, followed by sigmoid activation at the output layer [41]. The output of the network that ranged from zero to one was used to infer the risk group as a categorical variable.

$$CE = \begin{cases} -\ \log(f(s_1))\ if\ t_1 = 1 \\ -\ \log(1 - (f(s_1))\ if\ t_1 = 0 \end{cases} \quad (3)$$

where $t_1 = 1$ denotes the assignment of $C_1 = C_i$ for the sample. The entire network and its parameters were optimized using grid search and Bayes-opt optimizer.

Drug response prediction

Similar to the architecture of the survival prediction model, the drug response prediction neural network also had two hidden layers, followed by an output regression layer. The drug response model was trained on data points from 42 cell lines to predict drug responses of 212 drugs initially using MATLAB ver. R2020a (data not shown). However, with a limited model capacity due to small dataset, modelling large number of drug responses had an inverse effect on the performance, reflected in below-par metrics for many drugs. Therefore, another model was built for limited number of drugs that is reported in this study. This was done by eliminating drugs depicting an accuracy of less than 0.5. A total of 100 drugs fulfilled the criteria and the network architecture was modified to predict their drug responses.

The neural network was modelled as a regression problem to predict IC50 values. However, to binarize predicted IC50 to responses as sensitive or resistant, the original IC50 values were clustered into two classes using K-means clustering. It tries to make the inter-cluster points as similar as possible while trying to keep the clusters as far as possible under the objective function defined in eq. (4).

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \left|\left| x^i - \mu_k \right|\right|^2 \qquad (4)$$

The threshold IC50 value between the two classes was saved for each drug and later used to compare and test the effectiveness of the drug response prediction model. In addition to K-means as a method to calculate thresholds, waterfall analysis was also performed. Our implementation of the waterfall analysis was similar to a previous approach [54, 55]. For each of the 100 drugs, IC50 values were sorted to generate a waterfall distribution (Fig. S5). If the distribution is non-linear (Pearson correlation coefficient to linear fit ≤0.95), the inflection point was calculated by first smoothening the curve with a gaussian filter, followed by analysing the differential. In case of a linear distribution (Pearson correlation coefficient to linear fit > 0.95), median IC50 was used instead. 43 drugs had a non-linear waterfall distribution (Fig. S3-A). Inflection points and medians were used as thresholds to segregate among sensitive and resistant cell lines.

Hyperparameter optimization

BayesOpt and grid search was employed for tuning the parameters of the classification and regression neural network models. BayesOpt builds a probability model of the objective function to screen the best parameters to evaluate the model objective function [56]. For both drug response regression and survival classification tasks, hyperparameters corresponding to the objective functions were optimized using BayesOpt. The basic formulation is represented in eq. (5).

$$x^* = \arg\ min_{x \in \mathcal{X}} P(score \mid x) \qquad (5)$$

where $P(y| x)$ is the surrogate objective function (Mean Square Error or cross-entropy) and x^* is the set of hyper-parameters with the best model performance. It works by finding the parameters that correspond to the best performing surrogate function and using them on the actual objective function iteratively.

Abbreviations

AUROC: Area Under the Receiver Operating Characteristics; BRCA: Breast invasive carcinoma; CNV: Copy number variation; GDSC: Genomics of Drug Sensitivity in Cancer; GSEA: Gene Set Enrichment Analysis; MSE: Mean square error; NCA: Neighborhood component analysis; SMOTE: Synthetic Minority Oversampling TEchnique; TCGA: The Cancer Genome Atlas

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07524-2.

Additional file 1: Table S1. Description of the total number of features for each dataset that were used for multi-omics data integration type for building survival and drug response models. **Table S2** REACTOME pathways mapped onto the screened genes that outlines the critical reactions and modules that modulates a patient's survival. **Table S3** KEGG pathways mapped by identified signature of aberration for survival and drug response prediction model.

Additional file 2: Figure S1. Performance of multi-omics integration based neural network survival prediction model of BRCA patients. (A) Confusion matrix and (B) ROC plot of neural network prediction model.

Additional file 3: Figure S2. Violin plot showing the distribution of IC50 values of sensitive (blue) and resistant (orange) cell lines for the first 21 drugs.

Additional file 4: Figure S3. (A) Histogram depicting drugs with linear and non-linear correlation to a linear fit. (B) Correlation of IC50 thresholds calculated from the two methods (K-means and waterfall) shows that the two methods have consistent results.

Additional file 5: Figure S4. Performance of drug response model using thresholds from waterfall analysis. Box plot showing accuracy, sensitivity, and specificity of model for all drugs (A) and all cell lines (C). Scatter plot showing frequency of (B) sensitive cell lines per drug molecule and (D) effective drugs for each cell line.

Additional file 6: Figure S5. Waterfall distribution for 100 drugs under considerations. Blue lines depict IC50 thresholds as calculated by inflection point and median (threshold IC50 value is also depicted on each plot).

Acknowledgements

Not applicable.

Authors' contributions

D.S., and V.M. have conceived and designed the computational pipeline. V.M. has contributed to this study for the design, computational pipeline, in executing experiments and in writing manuscript. Y.K. supported in executing experiments, compiling the data and writing manuscript. The author(s) read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

All the codes and data used for the study can be accessed at https://github.com/TeamSundar/BRCA_multiomics.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no potential conflicts of interests.

Received: 28 November 2020 Accepted: 9 March 2021
Published online: 24 March 2021

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015;136(5):E359–86. https://doi.org/10.1002/ijc.29210.
2. Porter PL. Global trends in breast cancer incidence and mortality. Salud Publica Mex. 2009;51(Suppl 2):s141–6. https://doi.org/10.1590/S0036-36342009000800003.
3. Ali I, Wani WA, Saleem K. Cancer Scenario in India with Future Perspectives, vol. 8; 2011.
4. Babu GR, Lakshmi SB, Thiyagarajan JA. Epidemiological correlates of breast cancer in South India. Asian Pac J Cancer Prev. 2013;14(9):5077–83. https://doi.org/10.7314/APJCP.2013.14.9.5077.
5. Wang C, Machiraju R, Huang K. Breast cancer patient stratification using a molecular regularized consensus clustering method. Methods. 2014;67(3):304–12. https://doi.org/10.1016/j.ymeth.2014.03.005.
6. Chen X, Shachter RD, Kurian AW, Rubin DL. Dynamic strategy for personalized medicine: An application to metastatic breast cancer. J Biomed Inform. 2017;68:50–7. https://doi.org/10.1016/j.jbi.2017.02.012.
7. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, Weinstein JN. A gene expression database for the molecular pharmacology of cancer. Nat Genet. 2000;24(3):236–44. https://doi.org/10.1038/73439.
8. Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, Scherf U, Lee JK, Reinhold WO, Weinstein JN, Mesirov JP, Lander ES, Golub TR. Chemosensitivity prediction by transcriptional profiling. Proc Natl Acad Sci U S A. 2001;98(19):10787–92. https://doi.org/10.1073/pnas.191368598.
9. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, Olson JA Jr, Marks JR, Dressman HK, West M, Nevins JR. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature. 2006;439(7074):353–7. https://doi.org/10.1038/nature04296.
10. Arrowsmith J. Trial watch: phase II failures: 2008-2010. Nat Rev Drug Discov. 2011;10(5):328–9. https://doi.org/10.1038/nrd3439.
11. DiMasi JA, Reichert JM, Feldman L, Malins A. Clinical approval success rates for investigational cancer drugs. Clin Pharmacol Ther. 2013;94(3):329–35. https://doi.org/10.1038/clpt.2013.117.
12. Cancer Genome Atlas Research N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(7216):1061–8. https://doi.org/10.1038/nature07385.
13. Cancer Genome Atlas Research N. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474(7353):609–15. https://doi.org/10.1038/nature10166.
14. Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487(7407):330–7.
15. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70.
16. Cancer Genome Atlas Research N. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489(7417):519–25. https://doi.org/10.1038/nature11404.
17. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning-based multi-Omics integration robustly predicts survival in liver Cancer. Clin Cancer Res. 2018;24(6):1248–59. https://doi.org/10.1158/1078-0432.CCR-17-0853.
18. Zhang L, Lv C, Jin Y, Cheng G, Fu Y, Yuan D, Tao Y, Guo Y, Ni X, Shi T. Deep learning-based multi-Omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. Front Genet. 2018;9:477. https://doi.org/10.3389/fgene.2018.00477.
19. Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. Bioinformatics. 2019;35(14):i501–9. https://doi.org/10.1093/bioinformatics/btz318.
20. Malik V, Dutta S, Kalakoti Y, Sundar D. Multi-omics Integration based Predictive Model for Survival Prediction of Lung Adenocarcinaoma. In: Grace Hopper Celebration India (GHCI): 2019: IEEE Xplore: 9071831; 2019. p. 1–5.
21. Ding MQ, Chen L, Cooper GF, Young JD, Lu X. Precision oncology beyond targeted therapy: combining Omics data with machine learning matches the majority of Cancer cells to effective therapeutics. Mol Cancer Res. 2018;16(2):269–78. https://doi.org/10.1158/1541-7786.MCR-17-0378.
22. Wang C, Guo J, Zhao N, Liu Y, Liu X, Liu G, Guo M. A Cancer survival prediction method based on graph convolutional network. IEEE Trans NanoBioscience. 2020;19(1):117–26. https://doi.org/10.1109/TNB.2019.2936398.
23. Dong Z, Zhang N, Li C, Wang H, Fang Y, Wang J, Zheng X. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. BMC Cancer. 2015;15(1):489. https://doi.org/10.1186/s12885-015-1492-6.
24. Ding Z, Zu S, Gu J. Evaluating the molecule-based prediction of clinical drug responses in cancer. Bioinformatics. 2016;32(19):2891–5. https://doi.org/10.1093/bioinformatics/btw344.
25. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, Milacic M, Roca CD, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Viteri G, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P. The Reactome pathway knowledgebase. Nucleic Acids Res. 2018;46(D1):D649–55. https://doi.org/10.1093/nar/gkx1132.
26. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, Sidiropoulos K, Cook J, Gillespie M, Haw R, Loney F, May B, Milacic M, Rothfels K, Sevilla C, Shamovsky V, Shorser S, Varusai T, Weiser J, Wu G, Stein L, Hermjakob H, D'Eustachio P: The reactome pathway knowledgebase. Nucleic Acids Res 2020, 48(D1):D498-D503, DOI: https://doi.org/10.1093/nar/gkz1031.
27. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2008;37(1):1–13.
28. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30. https://doi.org/10.1093/nar/28.1.27.
29. Baxt WG. Complexity, chaos and human physiology: the justification for non-linear neural computational analysis. Cancer Lett. 1994;77(2–3):85–93. https://doi.org/10.1016/0304-3835(94)90090-6.
30. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol. 1996;49(11):1225–31. https://doi.org/10.1016/S0895-4356(96)00002-9.
31. Kwan TT, Bardia A, Spring LM, Giobbie-Hurder A, Kalinich M, Dubash T, Sundaresan T, Hong X, LiCausi JA, Ho U, et al. A digital RNA signature of circulating tumor cells predicting early therapeutic response in localized and metastatic Breast Cancer. Cancer Discovery. 2018;8(10):1286–99. https://doi.org/10.1158/2159-8290.CD-18-0432.
32. Corso G, Pascale V, Marrelli D, Pinheiro H, Carvalho J, Garosi L, Seruca R, Oliveira C, Roviello F. CDH1 structural alterations as novel prognostic biomarker in gastric cancer patients. J Clin Oncol. 2011;29(4_suppl):42.
33. El-Araby RE, Khalifa MA, Zoheiry MM, Zahran MY, Rady MI, Ibrahim RA, El-Talkawy MD, Essawy FM. CDH1 gene as a prognostic biomarker in HCV (genotype 4) induced hepatocellular carcinoma in the Egyptian patients. Gene Reports. 2019;16:100452. https://doi.org/10.1016/j.genrep.2019.100452.
34. Stirrups R. Ibrutinib and rituximab for chronic lymphocytic leukaemia. Lancet Oncol. 2019;20(9):e471. https://doi.org/10.1016/S1470-2045(19)30528-5.
35. Duffy MJ, Synnott NC, Crown J. Mutant p53 in breast cancer: potential as a therapeutic target and biomarker. Breast Cancer Res Treat. 2018;170(2):213–9. https://doi.org/10.1007/s10549-018-4753-7.
36. Davidson B, Stavnes HT, Risberg B, Nesland JM, Wohlschlaeger J, Yang Y, Shih le M, Wang TL. Gene expression signatures differentiate adenocarcinoma of lung and breast origin in effusions. Hum Pathol. 2012;43(5):684–94. https://doi.org/10.1016/j.humpath.2011.06.015.
37. Pharoah PDP, Guilford P, Caldas C. Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. Gastroenterology. 2001;121(6):1348–53. https://doi.org/10.1053/gast.2001.29611.
38. Sarrió D, Moreno-Bueno G, Hardisson D, Sánchez-Estévez C, Guo M, Herman JG, Gamallo C, Esteller M, Palacios J. Epigenetic and genetic alterations of APC and CDH1 genes in lobular breast cancer: relationships with abnormal E-cadherin and catenin expression and microsatellite instability. Int J Cancer. 2003;106(2):208–15. https://doi.org/10.1002/ijc.11197.
39. Bachman KE, Argani P, Samuels Y, Silliman N, Ptak J, Szabo S, Konishi H, Karakas B, Blair BG, Lin C, Peters BA, Velculescu VE, Park BH. The PIK3CA gene is mutated with high frequency in human breast cancers. Cancer Biol Therapy. 2004;3(8):772–5. https://doi.org/10.4161/cbt.3.8.994.
40. Olivier M, Langer A, Carrieri P, Bergh J, Klaar S, Eyfjord J, Theillet C, Rodriguez C, Lidereau R, Bi I, et al. The clinical value of somatic TP53 gene

mutations in 1,794 patients with breast cancer. Clin Cancer Res. 2006;12(4): 1157–67. https://doi.org/10.1158/1078-0432.CCR-05-1029.

41. Brown M, An PE, Harris CJ, Wang H. How Biased is Your Multi-Layered Perceptron? In: Proc World Congress on Neural Networks (01/01/93); 1993. p. 507–11.
42. Faraggi D, Simon R: A neural network model for survival data. 1995, 14(1): 73–82.
43. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, Aben N, Goncalves E, Barthorpe S, Lightfoot H, et al. A landscape of Pharmacogenomic interactions in Cancer. Cell. 2016;166(3):740–54. https://doi.org/10.1016/j.cell.2016.06.017.
44. Trevor Hastie RT. Balasubramanian Narasimhan and Gilbert Chu impute: impute: Imputation for microarray data; 2018.
45. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics. 2013;14(1):106. https://doi.org/10.1186/1471-2105-14-106.
46. Mirza B, Wang W, Wang J, Choi H, Chung NC, Ping P. Machine Learning and Integrative Analysis of Biomedical Big Data. Genes. 2019;10(2).
47. Huang E, Ishida S, Pittman J, Dressman H, Bild A, Kloos M, D'Amico M, Pestell RG, West M, Nevins JR. Gene expression phenotypic models that predict the activity of oncogenic pathways. Nat Genet. 2003;34(2):226–30. https://doi.org/10.1038/ng1167.
48. Li L, Li H. Dimension reduction methods for microarrays with application to censored survival data. Bioinformatics. 2004;20(18):3406–12. https://doi.org/10.1093/bioinformatics/bth415.
49. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25(22):2906–12. https://doi.org/10.1093/bioinformatics/btp543.
50. Louhimo R, Hautaniemi S. CNAmet: an R package for integrating copy number, methylation and expression data. Bioinformatics. 2011;27(6):887–8. https://doi.org/10.1093/bioinformatics/btr019.
51. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. Bioinformatics. 2012;28(24):3290–7. https://doi.org/10.1093/bioinformatics/bts595.
52. Goldberger J, Hinton GE, Roweis ST, Salakhutdinov RR. Neighbourhood components analysis. Adv Neural Inf Proces Syst. 2005;2005:513–20.
53. Møller MF. A scaled conjugate gradient algorithm for fast supervised learning: Aarhus University, Computer Science Department; 1990.
54. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, et al. The Cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603–7. https://doi.org/10.1038/nature11003.
55. Haibe-Kains B, El-Hachem N, Birkbak NJ, Jin AC, Beck AH, Aerts HJ, Quackenbush J. Inconsistency in large pharmacogenomic studies. Nature. 2013;504(7480):389–93. https://doi.org/10.1038/nature12831.
56. Martinez-Cantin R. Bayesopt: a bayesian optimization library for nonlinear optimization, experimental design and bandits. J Mach Learn Res. 2014; 15(1):3735–9.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.