


RESEARCH

Open Access



Prediction of chemoresistance trait of cancer cell lines using machine learning algorithms and systems biology analysis

Atousa Ataei¹, Niloufar Seyed Majidi², Javad Zahiri³, Mehrdad Rostami⁴, S. Shahriar Arab^{2*}  and Albert A. Rizvanov¹

*Correspondence:
sh.arab@modares.ac.ir
² Department of Biophysics,
Faculty of Biological Sciences,
Tarbiat Modares University,
Tehran, Iran
Full list of author information
is available at the end of the
article

Abstract

Most of the current cancer treatment approaches are invasive along with a broad spectrum of side effects. Furthermore, cancer drug resistance known as chemoresistance is a huge obstacle during treatment. This study aims to predict the resistance of several cancer cell-lines to a drug known as Cisplatin. In this papers the NCBI GEO database was used to obtain data and then the harvested data was normalized and its batch effects were corrected by the Combat software. In order to select the appropriate features for machine learning, the feature selection/reduction was performed based on the *Fisher Score* method. Six different algorithms were then used as machine learning algorithms to detect Cisplatin resistant and sensitive samples in cancer cell lines. Moreover, Differentially Expressed Genes (DEGs) between all the sensitive and resistance samples were harvested. The selected genes were enriched in biological pathways by the enrichr database. Topological analysis was then performed on the constructed networks using Cytoscape software. Finally, the biological description of the output genes from the performed analyses was investigated through literature review. Among the six classifiers which were trained to distinguish between cisplatin resistance samples and the sensitive ones, the KNN and the Naïve Bayes algorithms were proposed as the most convenient machines according to some calculated measures. Furthermore, the results of the systems biology analysis determined several potential chemoresistance genes among which *PTGER3*, *YWHAH*, *CTNNB1*, *ANKRD50*, *EDNRB*, *ACSL6*, *IFNG* and *CTNNB1* are topologically more important than others. These predictions pave the way for further experimental researches.

Keywords: Machine learning algorithms, Network topology, Cancer, Drug resistance, Classification, Feature selection

Introduction

Cancer is one of the most lethal and costly diseases around the world. There exist several common therapeutic procedures such as surgery, cytotoxic chemotherapy, targeted therapy, radiation therapy, endocrine therapy and immunotherapy which are used based on the level of cancer aggressiveness. The majority of the aforementioned approaches are invasive with a broad spectrum of side effects [1–4]. Furthermore, one important

challenge that clinical practice face is drug resistance, which results from tolerance of cancer cells to anti-cancer agents [5, 6].

The concept of drug resistance was first distinguished by observing bacterial resistance to antibiotics, but the phenomenon was also attributed to a wider range of disorders including cancers in no time [7]. Traditional chemotherapeutic agents destroy cancer cells by directly damaging DNA strand. Therefore, not only they are non-specific but also, they result in broad side effects. Furthermore, studies show that new drugs developed for targeting cancer cells are most effective in the beginning of the therapies, but as time passes, most patients show resistance to these drugs. Resistance to new targeted, chemotherapeutic agents is a big challenge in cancer therapies as these agents are responsible for the preponderance of relapses. The drug resistance phenomena root from different mechanisms which can be cancer-specific or not, such as drug efflux [7, 8].

So far, numerous researches have been done to distinguish and describe cancer drug resistance. *Housman et al.* categorized the mechanisms of drug resistance in cancer as drug inactivation, drug target alteration, drug efflux, DNA damage repair, cell death inhibition, and the epithelial-mesenchymal transition [7]. On the other hand, drug resistance was classified into intrinsic resistance that exists before drug treatment or acquired resistance which is induced after the therapy. The prediction of drug resistance can overcome the inevitable failure of targeted and chemical therapeutics in clinical anticancer treatment [9, 10].

Chemotherapy resistance prediction methods include cell culture-based chemo-sensitivity tests, DNA, RNA, and protein-based chemo-sensitivity tests and recently developed computational methods [11, 12]. Cell culture-based tests which have been used for more than 30 years have some technical weaknesses. The technique is time-consuming and the primary culture has a low potency to success [13, 14].

To resolve the aforementioned issues, DNA, RNA, and protein-based chemosensitivity tests emerged [15, 16]. These methods include gene-based tools such as the Oncotype DX[®] assay which uses 21 genes to predict the recovery of breast cancer after treatment. The tool can also be used for other cancer types including colon and prostate cancers [11]. Another such tool is the MammaPrint[®] which employs 70 genes to predict the possibility of metastasis in breast cancer [17]. The principal challenge in these tests is the recognition of participating genes in the chemoresistance process. Moreover, due to the development of interdisciplinary techniques, computational and statistical methods are used for predicting chemotherapy responses [11].

So far, numerous computational approaches are developed to study drug resistance based on biological mechanisms. These computational techniques are generally divided into mechanism-based mechanistic modeling methods and data-driven prediction methods [18–20]. Molecular dynamics simulation, Kinetic models of signaling networks, Ordinary differential equation (ODE) model of cellular populations, Stochastic models, Partial differential equation models (PDEs), Agent-based and Pharmacokinetic–pharmacodynamic models are examples of the first class. The second class benefits from Omics data-based node biomarker screening, Static network biomarker prediction and Dynamic network biomarker prediction models. Linear models, support vector machines (SVMs), hierarchical clustering, principal components

analysis (*PCA*) and the formation of a scoring algorithm are other models of computational methods used for the prediction of cancer responses. These models which belong to a concept known as “machine learning algorithms (ML)” are being used to predict resistance of cancer cells to chemotherapy [20, 21]. Huang et al. represent ML as a part of artificial intelligence that can find correlations in the cancer-relevant datasets [21]. Conventional analytical approaches for determining treatment are very expensive and are also limited due to innate technical issues. ML algorithms are considered as cost–benefit, time saving strategies which can evaluate multiple cells line simultaneously [22, 23]. Yet another challenge which ML can overcome compared to conventional methods is the ability to determine biological information which are concealed by tumor cell heterogeneity [24].

This study aims to predict the resistance of several cancer cell-lines to Cisplatin. In this study, different algorithms including Naïve Bayes, K-Nearest Neighbors (with $k=3$), Decision tree, Random Forest and Neural network are used to classify Cisplatin sensitive and resistance samples. Moreover, the results of our systems biology analysis indicated several potential chemoresistance genes among which *PTGER3*, *YWHAH*, *CTNNB1*, *ANKRD50*, *EDNRB*, *ACSL6*, *IFNG* and *CTNNB1* are topologically more important than others. Our results have been validated against different databases such as UniProt, Enrichr and DIANA mirPath v.3 and the papers extracted from the literature. Furthermore, the results of the systems biology analysis determined several potential chemoresistance genes among which *PTGER3*, *YWHAH*, *CTNNB1*, *ANKRD50*, *EDNRB*, *ACSL6*, *IFNG* and *CTNNB1* are topologically more important than others.

The remainder of this paper is formed as below: “**Materials and methods**” section reviews the Materials and Methods. The results are presented in “**Results**” section. The discussion is reported in “**Discussion**” section and finally, “**Conclusion**” section present the conclusion of the overall work.

Materials and methods

Data collection

NCBI GEO database was used to obtain data from datasets [25]. Five different platforms of microarray including GPL13667, GPL6947, GPL6480, GPL6104, and GPL6244 have been used. A total of 85 samples of gene expression datasets of microarray data were selected from different platforms. The selected datasets were related to various cell lines such as ovarian, pancreatic and non-small cell lung cancer (NSCLC) both resistance and sensitive to Cisplatin drug. Sample numbers as well as cancer types are indicated in Table 1.

The data has been normalized using the LIMA package in the R software [26]. Average has been taken between the expressed values of repetitive probes in each dataset to obtain a unique expression value for each probe. A total 7621 genes were harvested after the isolation of common genes between platforms. The Combat software was then used to eliminate batch effects between different platforms and experiments [27]. Also, an average has been taken between replicates of each platform (sensitive and resistance separately) which reduces the total number of samples from 85 to 14 sensitive and 14 resistances to Cisplatin samples.

Table 1 Samples collected from NCBI GEO database

Series number	Platform	Cancer Type	Number of Samples
GSE26465	GPL6104	Ovarian	6(2 s,4r)
GSE33482	GPL6480	Ovarian	12(6 s,6r)
GSE21656	GPL6244	Lung	6(3 s,3r)
GSE84146	GPL6480	Lung (2 cell lines), Ovarian (2 cell lines)	16(2 s,2r&2 s,2r&2 s,2r&2 s,2r)
GSE73935	GPL13667	Ovarian (2 cell lines)	15(3 s,6r&3 s,3r)
GSE58470	GPL6947	Ovarian	6(3 s,3r)
GSE45553	GPL6244	Ovarian	8(4 s,4r)
GSE73978	GPL6244	Pancreatic Cancer (2 cell lines)	12(3 s,3r&3 s,3r)
GSE51683	GPL6244	Ovarian	4(2 s,2r)

Data processing

As the data was collected from various sources, it was necessary to somehow remove the discrepancies known as “Batch effects” between different samples. The batch effects of 85 different samples, each containing 7620 genes, were corrected by the SVA package. The SVA package comprises functions to remove the batch factors and other undesirable conversion in high-throughput examination. Specifically, the SVA package comprises functions to identify and build surrogate variables for high-dimensional datasets. Surrogate variables are covariates built directly from high-dimensional data (such as gene expression and RNA sequencing data) that can be utilized in subsequent analyses to adjust for unknown, unmodeled, or latent sources of noise. Moreover, the t-SNE algorithm was then used in MATLAB software to make the data presentable before and after the batch effect correction [28].

Feature selection

High dimensional DNA microarray has presented serious challenges to the existing machine learning and classification methods. In other words, in many of medical and microarray datasets, it is possible that many genes are irrelevant or redundant for machine learning algorithm [29–32]. Feature selection or gene selection is a popular and powerful approach in medical datasets to overcome this shortcoming [33–35]. In gene selection, to decrease the microarray data dimensions, by eliminating the irrelevant and similar genes, only a subset of relevant and dissimilar genes that are strongly related to the objective function are selected [36]. This is a powerful strategy to increase the efficiency of the machine learning algorithm, reduce time complexity, build more general classification algorithm, and reduce storage requirements [37, 38]. Gene selection approaches have been successfully employed in many medical applications including; gene expression [39], cancer classification [40], medical diagnosis [32], etc. In other words, a fundamental problem in machine learning algorithms is the high dimensional datasets, in which the size of the feature subset is much higher than the size of the patterns. Therefore, the classification accuracy is significantly reduced. As a result, it is necessary to reduce the initial features using dimension reduction techniques [41, 42]. One efficient way to reduce the dimension is feature selection (gene selection in DNA microarray datasets). In feature selection, an attempt is made to choose a set of initial features that satisfy

two targets simultaneously: the minimum similarity between the selected genes and the maximum relevance of these genes with the target class [43, 44]. The main goal of this step is to select appropriate and important feature from original features [45]. To do this purpose, Fisher Score (FS) feature selection algorithm is used to select the features that are most relevant to the target class. Fisher Score is a supervised filter method that selects a feature subset such that the samples in the specific class are most similar to each other and the samples in the different classes are less similar. As a result, this measure Scores higher value to genes with higher separation characteristic. The FS of gene f_i is defined as follows:

$$FS(f_k) = \frac{\sum_{v \in V} n_v (\bar{f}_k^v - \bar{f}_k)^2}{\sum_{v \in \wedge V} (\sigma_v \wedge V(f_k))^2} \quad (1)$$

where \bar{f}_k is the mean value of all the samples regarding the feature f_k , V is a set of all classes in a dataset, n_v is the number of pattern on the class v , and $\sigma(f_k)$ and \bar{f}_k^v shows the variance and average of feature f_k on class v , correspondingly.

The Fisher Score method was performed for 14 folds and in each step, 100 features that were most consistent with the normal distribution were selected so that a total of 1400 features were obtained. In the next step, in order to reduce the sample size, the *PCA* [46] method was performed for 14 folds on the output features of the *Feature Score* method and reduced set of features were selected for machine training.

Machine learning

In order to measure the flexibility of the developed method on different classifiers, in the designed experiments, the efficiency of the various approaches on three widely used six machine learning algorithms including *Naïve Bayes* [47], *SVM* [7], *KNN* [48], *Decision tree* [49], *Random Forest* [50] and *Neural Network* [51] algorithms is examined for machine training to detect Cisplatin sensitive and resistant samples in cancer patients undergoing chemotherapy. These machine learning algorithms are one of the most well-known and widely used machine learning algorithms that are used by researchers in various prediction and classification problems.

In these experiments the developed approach and other compared methods implemented using Python language programming on an Intel Core-i9 CPU with 16 GB of RAM.

To train these algorithms and due to the small number of available samples (14 pairs of samples including 14 sensitive and 14 resistant samples), the *Leave one out method* (LOO) was used [52]. In this method, at each step of the machine training, 13 pairs of samples were used as training data and one pair was used as test data. The training process continued until all pairs of samples were once used as the test data. The training was performed twice, once by considering 1400 features extracted from *Fisher Score* method and once by considering 210 features extracted by *PCA* method as the training samples. Finally, the performance of the trained machines was evaluated according to *Accuracy* [53], *Specificity* [54], *Sensitivity* [54], *Precision* [53], *MCC* [55] and *F1 scores* [55].

Biological system evaluation

73 genes were selected using the CountIF filter in Excel (Refer to online resource 1). These genes were repeated in 50% or more of the 14 pairs sample extracted from the GEO. The LIMA package was used to calculate the DEGs between all sensitive and resistance samples. The result indicated that 34 of 73 genes which were chosen in the feature selection phase were down-regulated and the rest were up-regulated. The genes were enriched in biological pathways using the Enrichr database [56]. To obtain the mirs related to the 73 harvested genes, miRCancer [57] and miRDB [58] databases were used and the common mirs between the harvested mirs from the two aforementioned databases were selected based on the specific cancer cell line as well as the direction of the regulation. mirs were enriched using DIANA mirPath v.3 database [59] (refer to online resource 2) and pathways related to chemoresistance were selected among the obtained pathways. Furthermore, the transcription factors which regulate the obtained DE genes were harvested using the TRRUST v2 database [60] and were enriched in pathways related to chemoresistance by Enricher database.

Networks

Using the obtained transcription factors and mirs, two types of network including a TF-gene network and a mir-gene network were constructed using Cytoscape software [61]. Topological analyses were then performed using degree and betweenness centralities to report the networks hub genes.

Biological description

The obtained DEGs were studied through literature review. All of the 73 genes were checked in articles and are indicated in a table based on the mechanism of chemo resistance. Some of these genes were not mentioned in the related articles as chemo resistance agents and therefore, they were nominated for further resistance studies which indicated that these genes have main roles in cells and are included in important biological pathways.

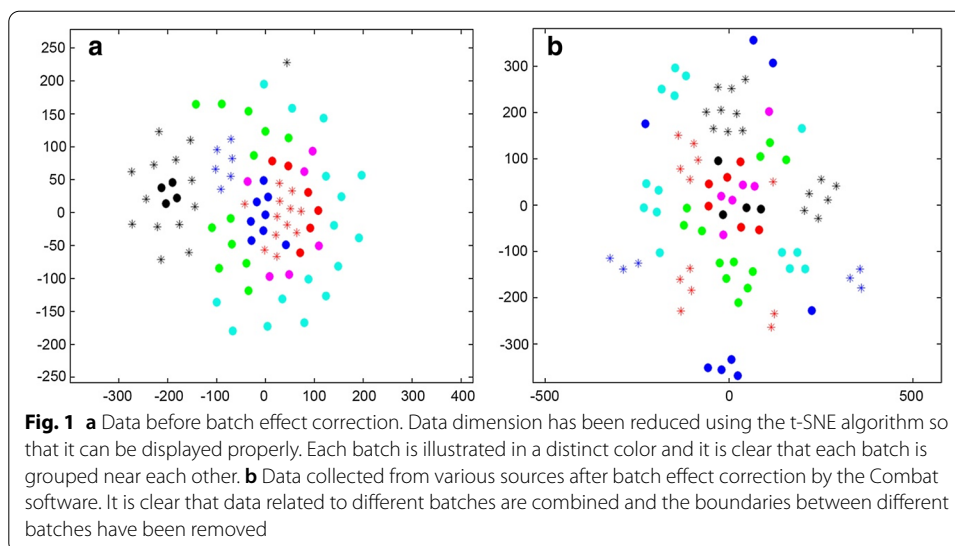
Results

Data processing

To correct batch effects between different samples, t-SNE algorithm was used (Fig. 1). This algorithm reduces data dimension and makes the data easier to picture and therefore, easier to comprehend.

Features selected and reduced by Fisher Score and PCA algorithms

In order to select appropriate features for machine learning algorithms, feature selection was performed for 14 folds using the Fisher Score method. In each fold, 100 genes which were most similar to the normal distribution were selected out of 7620 genes. Furthermore, to reduce the dimension of the selected features, the PCA algorithm



was performed for 14 folds on the extracted features and several genes were selected in each fold (Table 2).

A machine learning approach to detect Cisplatin sensitive and resistant samples in cancer cell lines

Performance of the machines trained by using reduced features extracted from the PCA algorithm (Table 2) are shown in Fig. 2. Among the developed machines, the Decision Tree algorithm with the average Accuracy of 50% has the weakest performance in terms of accuracy. On the other hand, KNN shows the highest accuracy with an average of 67%. The best performance based on accuracy criteria also belong to KNN and Decision Tree algorithms according to the obtained box plots (Fig. 2c).

Among the developed machines, the Naïve Bayes algorithm is the weakest machine in terms of negative sample detection with a 50% Specificity criteria. Decision Tree algorithm, on the other hand, has the highest average Specificity criterion of 69%, followed by Random Forest and KNN algorithms. The KNN algorithm has the best performance based on the specificity criterion based on the extracted box plots (Fig. 2c).

Based on the results, the weakest performance in terms of Sensitivity criteria is related to the Naïve bayes algorithm, which has not correctly detected any positive samples. On the other hand, KNN and Decision Tree algorithms have the highest Sensitivity criteria with an average of 78 and 67%, respectively (Fig. 2). The Decision Tree algorithm has the best performance in terms of sensitivity criteria based on the obtained box plots (Fig. 2C).

Among these algorithms, Decision Tree and Random Forest algorithms with an average precision of 71% have the highest average precision. The Decision Tree algorithm has the best performance in terms of precision according to the extracted box plots (Fig. 2C).

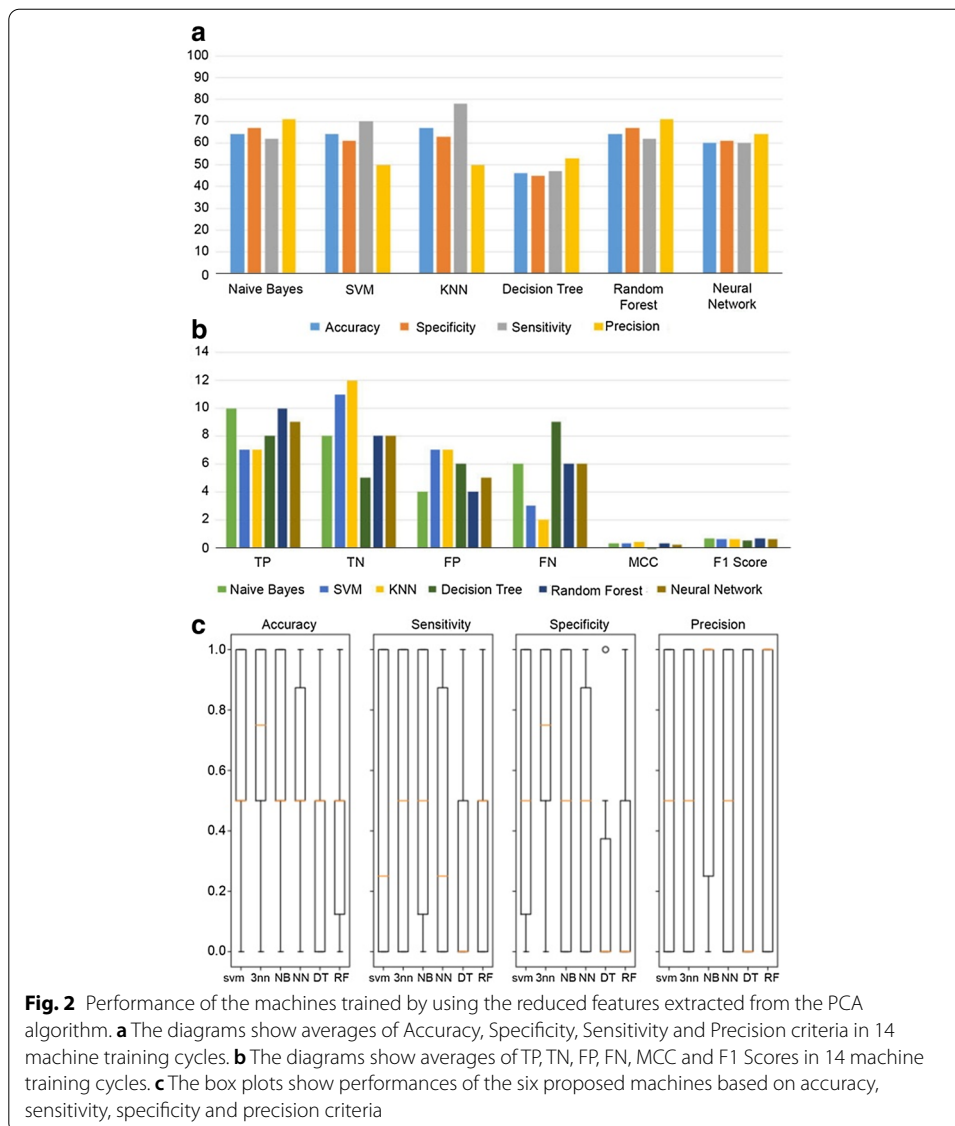
Similarly, a new set of six machines was trained only this time, 1400 features extracted from the Fisher Score algorithm were used in the training process. Performance results of these machines are shown in Fig. 3. In the new set, the KNN algorithm with an

Table 2 Features selected for machine learning purposes. Selected features are obtained using Fisher Score and PCA algorithms, respectively

1	2	3	4	5	6	7	8	9
actn4	abcb7	abcb7	abca5	acad8	abcb7	abcc13	actn3	abcb1
abhd12b	acot9	acaca	abcc3	abcb7	abcc9	acox1	abca9	abcb11
Acad1	abcd1	abcc2	acot9	abcc3	actn1	Acads	abca13	abca4
abcc5	actn3	acot2	acacb	acot9	acot9	actn4	abcb1	acot9
acaa1	acaa2	abtb2	actn4	abtb2	acaa2	abcc5	acot7	acot11
abcb7	abca5	actn4	abtb2	actn4	abca4	Acaca	abtb2	acaa2
abca12	acaa1	abca5	abca4	abca5	actn4	abca5	abca12	abca13
Acan	actn4	abcc5	abcc8	abcd1	acaa1	Acd	actr1a	actr1a
Abra	aars2	aadat	aars2	aars2	Aadat	Aadat	aadat	aadat
Aadat	abra	ablim1	abo	Abr	Abra	abtb2	abra	abtb2
acsl4	abhd10	abcc3	abcc5	abcg8	abhd10	abcg5	abhd12b	abhd14a
abcc3	accs	abcg4	abcg8	acbd5	acbd5	Ache	accs	acd
Accs	acsl4	acan	acbd5	abcc8	acsl1	abcc3	abcc8	abcc9
abcg4	abcc9	acads	acads	Acan	abcc8	acsl5	acsl1	acsl4
Acd	acbd5	acsf2	acsl4	acsl4	Acan	ace2	acbd5	accs
10	11	12	13	14	15	16	17	18
acox3	abcc13	abcd1	abcc2	abcb7	abca4	abcc5	abca6	abce1
abcc8	acan	abcc2	abcd1	acox1	abcb7	acsbg2	abcc5	abra
acbd5	abcc8	abcc8	abcc8	abcc3	acox1	abcg5	acads	abcg4
actn4	aass	abcb8	abcb8	Abr	abcc5	acad8	actn4	acot12
acad8	actn4	abcb7	abcb7	ace2	Abra	abca5	acaca	acaa2
abca5	acad8	acot9	acot7	acsl6	abcg1	abcc2	abcc2	abcf2
abcc3	abca6	abca6	abca6	abcg1	acsl6	actn4	abca5	abl2
acaa2	abcd3	actn4	actn4	Acd	ace2	Acd	aars2	abcf3
aars2	aars2	Aadat	aadat	Acaca	Acacb	Aadat	abcd1	acss2
abcd3	acaa2	Abra	abr	aco1	aco1	acaa2	abcc9	abcg5
abcd1	abcd1	abhd10	abhd10	acsl5	acsl5	abhd12b	acaa2	Acads
abhd14a	abhd14b	Accs	accs	Acadm	Acads	Acmsd	abhd11	abcc3
Acly	ache	acsl1	acsf2	abcg5	abcg5	acsl6	ache	abca8
Ache	ace2	abce1	abce1	abca13	abca12	abcc9	ace2	acot2
acsl6	acsm3	acbd5	acbd5	abcc9	abcd1	Acly	acsl5	acsl4

average of 67% accuracy has the highest percentage of correct sample detection compared to other algorithms. Random Forest, Naïve Bayes and SVM algorithms come afterward with an average accuracy of 64%. The KNN algorithm is also the best machine in terms of accuracy based on extracted box plots (Fig. 3c).

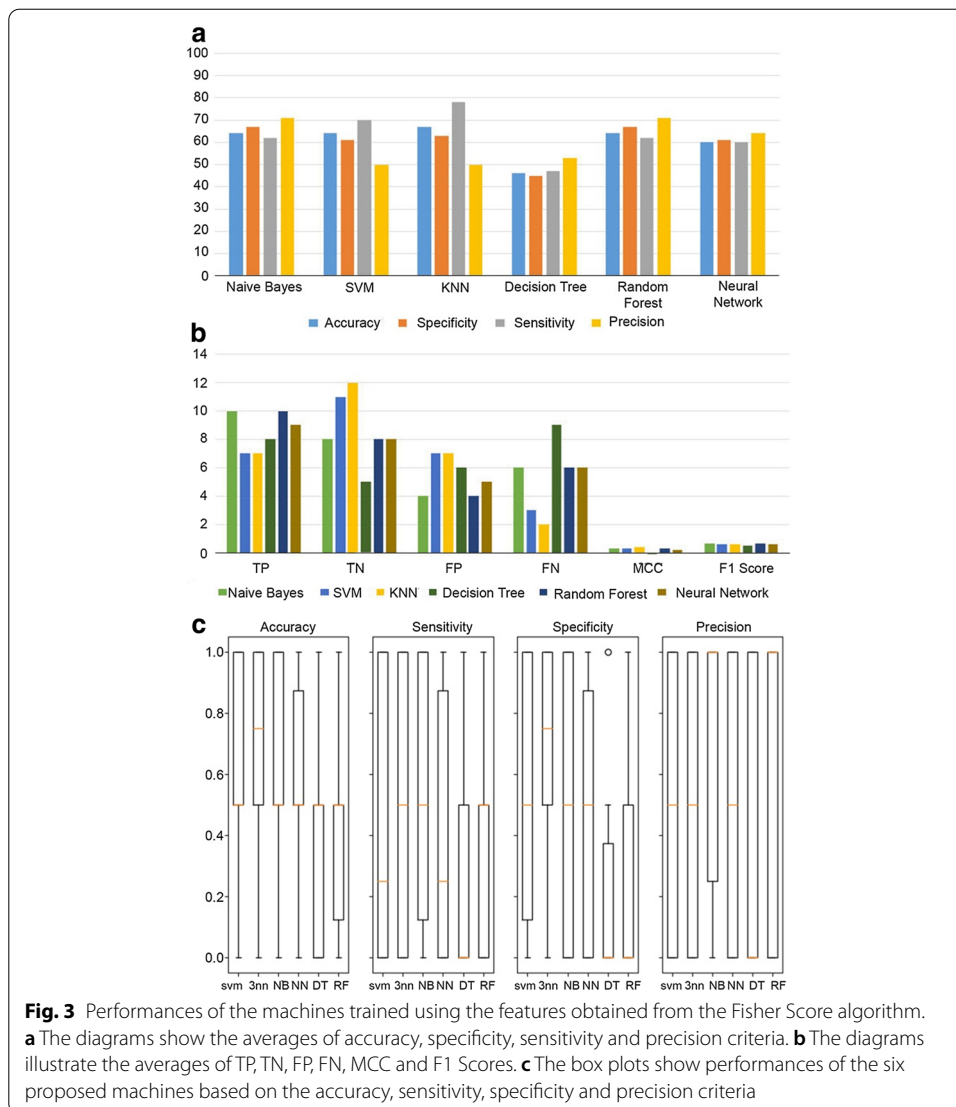
Naïve Bayes and Random Forest algorithms, with an average specificity criterion of 67%, are the best machines to correctly detect negative samples. On the other hand, the Decision Tree algorithm with the average specificity of 45% has the weakest performance in this regard (Fig. 3). Furthermore, the KNN algorithm has the best performance in terms of specificity based on the obtained box plots (Fig. 3c). In addition, the KNN algorithm with 78% average sensitivity is the best machine to correctly detect positive samples. The SVM algorithm comes afterward with an average sensitivity of 70% (Fig. 3). According to the obtained box plots, the Naïve Bayes and KNN algorithms are the best choices in terms of Sensitivity criteria, respectively. In addition, among the



above algorithms, Naïve Bayes and Random Forest algorithms with an average precision of 71% are the most precise machines. Similarly based on the calculated box plots, the Naïve Bayes algorithm performed better than other algorithms in terms of precision criteria. Finally, according to the MCC criteria, the KNN algorithm and according to the F1 Score criterion, the Random Forest and Naïve Bayes algorithms have the best performances (Fig. 3).

Determining specific mirs for extracted DE genes

The specific mirs for the extracted 73 DEGs were harvested from the miRCancer and miRDB databases for related cancer types (Table 1). These results determine that the expression profile of mirs are down for upregulated genes and are up for down regulated ones (Table 3; Fig. 4).



mir-target network topology

The mir target network has been topologically analyzed using the degree centrality and the results revealed the hub genes which should be evaluated for their performance in the chemoresistance process (Table 4).

mir enrichment

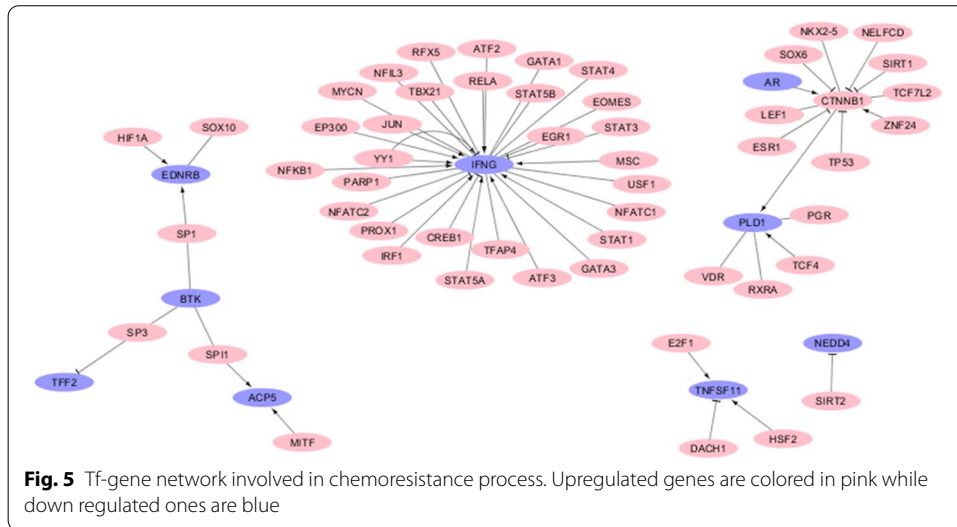
The upregulated and down regulated mirs in the pathways related to chemoresistance were enriched using DIANA mirPath v.3 database. The related pathways were extracted from KEGG database using the standard P-value of 0.05 (The extracted related pathway: Additional file 1).

Table 4 The extracted hub genes and mirs nominated for performance evaluation in the chemoresistance process

Name	Degree
PTGER3	7
YWHAH	6
CTNNB1	6
ANKRD50	5
EDNRB	5
ACSL6	4
PDCD6IP	3
hsa-mir-206	3
GNAI2	3
hsa-mir-486-5p	3
PLD1	3
hsa-mir-760	3
TMED5	3
hsa-mir-661	3

Table 5 The detected hub genes based on the in-degree centrality in the TF-gene interaction network

Hub genes	In-degree factor
*IFNG	30
CTNNB1	10



TF network topology

Three factors including the in-degree, the out-degree and the betweenness centralities have been noted in TF-gene interaction network for topological analysis. The detected hub genes are related to the in-degree centrality and are listed in Table 5.

Table 6 The TFs related to both upregulated and down regulated genes. In the upregulated genes the TFs are down and in the down regulated genes the TFs are up

TFs-ups	TARGET-downs	Relation	TFs-ups	TARGET-downs	Relation	TFS-downs	Targets-Ups	Repression
ATF3	IFNG	Activation	PROX1	IFNG	Repression	AR	CTNNB1	Activation
CREB1	IFNG	Activation	RELA	IFNG	Activation	CTNNB1	PLD1	Activation
CREB1	IFNG	Repression	RELA	IFNG	Unknown	ESR1	CTNNB1	Repression
CREB1	IFNG	Unknown	RFX5	IFNG	Unknown	LEF1	CTNNB1	Unknown
DACH1	TNFSF11	Repression	SIRT2	NEDD4	Repression	NELFCD	CTNNB1	Repression
E2F1	TNFSF11	Activation	SOX10	EDNRB	Unknown	NKX2-5	CTNNB1	Unknown
EGR1	IFNG	Unknown	SP1	BTK	Unknown	PGR	PLD1	Unknown
EOMES	IFNG	Unknown	SP1	EDNRB	Activation	RXRA	PLD1	Unknown
EP300	IFNG	Activation	SP1	EDNRB	Unknown	SIRT1	CTNNB1	Repression
GATA1	IFNG	Unknown	SP3	BTK	Unknown	SOX6	CTNNB1	Repression
GATA3	IFNG	Unknown	SPI1	ACPS	Activation	SP3	TFF2	Repression
HIF1A	EDNRB	Activation	SPI1	BTK	Unknown	TCF4	PLD1	Activation
HSF2	TNFSF11	Activation	STAT1	IFNG	Activation	TCF7L2	CTNNB1	Unknown
IRF1	IFNG	Activation	STAT1	IFNG	Repression	TP53	CTNNB1	Repression
JUN	IFNG	Activation	STAT1	IFNG	Unknown	VDR	PLD1	Unknown
JUN	IFNG	Unknown	STAT3	IFNG	Repression	ZNF24	CTNNB1	Activation
MITF	ACPS	Activation	STAT4	IFNG	Unknown			
MSC	IFNG	Activation	STAT5A	IFNG	Activation			
MYCN	IFNG	Unknown	STAT5B	IFNG	Unknown			
NFATC1	IFNG	Unknown	TBX21	IFNG	Activation			
NFATC2	IFNG	Unknown	TBX21	IFNG	Unknown			
NFIL3	IFNG	Unknown	TFAP4	IFNG	Unknown			
NFKB1	IFNG	Activation	USF1	IFNG	Unknown			
NFKB1	IFNG	Unknown	YY1	IFNG	Activation			
PARP1	IFNG	Unknown	YY1	IFNG	Repression			
			YY1	IFNG	Unknown			

The network illustrated in Fig. 5 has been constructed by merging upregulated and down regulated genes and their corresponding regulatory TFs. Topology analysis has been performed based on this network. According to the results obtained from the trrust database, the TFs were down for upregulated genes and were up for down regulated genes (Table 6).

TF enrichment

The transcription factors regulating the genes harvested from the feature selection step were enriched using Enrichr database in the oncogenes and chemoresistance pathways. In the Enrichr database the pathways were harvested from KEGG and the adjusted P-value was significant. The results are shown in Table 7.

Biological description

A literature review was performed on the 73 DE genes which were harvested by the CountIF filter in the feature selection step. A group of these genes were reported in the literature as chemoresistance genes (Additional file 2). The other ones were identified to have a vital role in the oncogenesis pathway and other important cell functions.

Table 7 The annotations of transcription factors regulating the harvested genes from the feature selection step

Term	P-value	Adjusted P-value	Genes
Pathways in cancer	1.2063332677701377E-19	3.715506464732024E-17	STAT5A;STAT5B;TCF7L2;JUN;SPI1;STAT1;LEF1;STAT3;MITF;HIF1A;ESR1;RELA;NFKB1;AR;RXRA;SP1;E2F1;STAT4;EP300;CTNNB1;TP53
Wnt signaling pathway	9.705297353623653E-9	1.5732797815347815E-7	TCF7L2;JUN;LEF1;EP300;NFATC2;CTNNB1;NFATC1;TP53
Non-small cell lung cancer	2.334429995619187E-8	3.126106255003085E-7	STAT5A;STAT5B;RXRA;STAT3;E2F1;TP53
TNF signaling pathway	5.026455978790371E-7	4.553377769021866E-6	ATF2;JUN;CREB1;IRF1;RELA;NFKB1
JAK-STAT signaling pathway	4.820060593301115E-6	3.620923567650594E-5	STAT5A;STAT5B;STAT1;STAT3;TAT4;EP300
Small cell lung cancer	5.15158695877854E-6	3.6899739146599775E-5	RXRA;E2F1;TP53;RELA;NFKB1
HIF-1 signaling pathway	7.356942182113052E-6	5.035418204646267E-5	STAT3;EP300;HIF1A;RELA;NFKB1
cAMP signaling pathway	2.2332816777813463E-5	1.4037770546054178E-4	JUN;CREB1;EP300;NFATC1;RELA;NFKB1
Apoptosis	4.156130787036355E-5	2.3274332407403588E-4	JUN;PARP1;TP53;RELA;NFKB1
Adherens junction	4.260972052484695E-5	2.3435346288665825E-4	TCF7L2;LEF1;EP300;CTNNB1
MAPK signaling pathway	1.39454465445684E-4	7.158662559545113E-4	ATF2;JUN;NFATC1;TP53;RELA;NFKB1
Toll-like receptor signaling pathway	1.786661621004991E-4	8.59830905108652E-4	JUN;STAT1;RELA;NFKB1
PI3K-Akt signaling pathway	3.7197722896986754E-4	0.0016604200945321624	ATF2;CREB1;RXRA;TP53;RELA;NFKB1
cGMP-PKG signaling pathway	0.0010478054962433662	0.0042463696426704835	ATF2;CREB1;NFATC2;NFATC1
ErbB signaling pathway	0.0015715511019872403	0.006205612043744487	STAT5A;STAT5B;JUN
NF-kappa B signaling pathway	0.002160916108972604	0.008116611726384903	PARP1;RELA;NFKB1
Sphingolipid signaling pathway	0.004084635694037895	0.014800797573690256	TP53;RELA;NFKB1
Cell cycle	0.004582612813265843	0.01641214821495209	E2F1;EP300;TP53
FoxO signaling pathway	0.005453058494504465	0.01930508064721121	STAT3;EP300;SIRT1
Hippo signaling pathway	0.009248140092633821	0.03130139723660678	TCF7L2;LEF1;CTNNB1

Therefore, although they were not reported as chemoresistance genes we propose that these genes might be potentially chemoresistance (Table 8). Future studies can be performed to validate these results.

Discussion

The main aim of this study was to train a machine for the detection of sensitive and resistant samples to Cisplatin in different cancer cell lines including ovarian, pancreatic and lung cancers. Six machines were developed based on different algorithms including Naïve bayes, SVM, KNN, Decision tree, Random Forest and Neural network and the results were evaluated by the accuracy, specificity, sensitivity, precision, MCC and F1 Scores. Furthermore, a series of systems biology analyses were performed using the DE genes harvested from the feature selection step to further improve our study.

Table 8 potential chemoresistance genes proposed for further studies

Potential chemoresistance genes	Main role in cell
klhdc10	Except klhdc10, another client protein for kelch is phosphatase 5 (PP5) which in response to ROS inactivates ASK1. After this interaction, PP5 phosphatase activity will be suppressed. Furthermore, kelch mediates H ₂ O ₂ -induced sustained activation of ASK1 and cell death in Neuro2A cells This data proposes that Slim/KLHDC10 is an activator for ASK1 and its activation through suppression of pp5 leads to oxidative stress-induced cell death [62]
MCRS1	Nucleolar MCRS1 which is called MSP58. It is proposed that TOJ3, an avian homologue of MSP58, is associated with Jun-induced cell transformation as well as tumorigenesis. Other studies have identified MSP58 as an oncogene hence its transformation activity was blocked by interaction with the PTEN tumor suppressor. It has also been reported that there exist different expression levels of MSP58 in human glioma and colorectal cancer [63]
MSH4	MSH4 mutation has been associated with lung cancer [64]. Furthermore, although hMSH4 and hMSH5 are not involved in DNA-mismatch correction but they participate in the Mutual recombination and appropriate separation of homologous chromosomes at meiosis [65]
nucb2	It has been proposed that NUCB2 is associated with metastasis in melanoma. Several studies have demonstrated that KLF4 levels are increased in melanoma cells leading to apoptosis inhibition and metastasis [66]
sh3gl2	It is proposed that SH3GL2 can have a tumor suppressor role in brain since deletion mutations in the locus of this gene can cause pilocytic astrocytomas. It has also been shown that through regulation of SH3GL2 gene, miR-330 affects proliferation, migration, invasion, cell cycle and apoptosis of human glioblastoma [67]
TMED5	malignant phenotypes including increased cell proliferation, EMT progression, apoptosis inhibition, cell migration and invasion, and drug resistance can emerge as a result of higher expression of TMED5 [68]. TMED5 is also proposed to play a role in NCI/ADR-RES drug-resistance (MDR) [69]
TMEM119	TMEM119 plays a role in migration and invasion of gastric cancer cells through activation of STAT3 signaling pathway which is found to be strongly correlated with the invasion, metastasis, and prognosis of gastric cancer [70]. Moreover, Zheng et al. showed that down-regulation of TMEM119 reduces Bcl-2 levels and increases Bax and caspase-3 levels in SGC-7901 cells [71]
TMEM219	The over-expression of TMEM219 gene which is localized in the membrane of breast, prostate, and pancreatic tumor cells, can suppress tumor growth. Other studies have revealed that expression levels of IGFBP 3 as well as its death receptor are in close relation to inefficient prognosis and low survival rate in pancreatic ductal adenocarcinoma [72]
WIPI1	WIPI1 up regulation has been detected in a variety of tumors. It has also been proposed that WIPI1 plays a role as an autophagy activator through TORC1 suppression. Furthermore, it has been stated that WIPI1 up regulation results in both lower relapses and higher survival rates in breast cancer [73]
YWHAH	YWHAH gene can be considered as a potential target for therapeutic agents as it is down-regulated in liposarcoma cells after Doxorubicin treatment [74]. Furthermore in another study, Kibel AS et al. proposed that YWHAH expression levels is positively correlated with malignant Prostate Cancer [75]
znf507	It has been proposed that over-expression of ZNF507 is associated with pancreatic, periampullary adenocarcinoma [76] as well as ovarian high-grade serous carcinomas (as an amplified 'driver' gene) [77]
ACP5	(TRAP-ACP5) can be used as a marker for predicting cancer progression and aggressiveness as it plays critical roles in many biologic processes such as bone resorption, osteoclast differentiation and, cell motility promotion through the modulation of focal adhesion kinase phosphorylation. It also acts as a metalloenzyme in activated osteoclasts and macrophages and also serves as a metastasis driver in cancer. It has been recently demonstrated that TRAP, through TGFβ ₂ /TβR and CD44 signaling pathway, results in metastatic MDA-MB-231 breast cancer cells [78]
ANXA6	Several cancer types including Melanoma, CC, Epithelial Carcinoma, BC, GC, PCa, ALL, CML, large-cell lymphoma and myeloma have been proposed to be related to dysregulation of AnxA6 Therefore, AnxA6 is a potential biomarker for identification, cure and prognosis of certain cancers [79]. Moreover, recent studies have proposed AnxA6 as a suggested target for inhibition of pancreatic cancer via antibodies [80]

Table 8 (continued)

Potential chemoresistance genes	Main role in cell
Atp6v1g3	BSND and ATP6V1G3 has been proposed as novel immunohistochemical markers for the differential diagnosis of chromophobe RCC from other RCC subtypes and also diagnosis of chromophobe RCC metastasis to distant organs [81]
CWF19I2	CWF19 like cell cycle control factor 2 (PMID: 143,884). Breast cancer development has been related to ERBB2, MYC, GSTT1, PIK3CA and CWF19L2 [82]
DSC1	Dsc2 level is observed to decrease in colorectal cancer. The alterations in Dsc expression pattern can cause significant changes in desmosome function [83]. DSC1 can also be considered as a biomarker for tumor differentiation, and it can be a prognostic marker for lung cancer [84]
DUT	DUT expression provides a discernible phenotype in a variety of cancers which can be used for prediction of patients response to chemotherapy as well as overall survival. It is significant to note that resistance to thymidylate synthase inhibitors is related to 3–fivefold increased expression levels of dUTPase in HT29 and A549 cells [85]
EDNRB	EDNRB methylation is helpful in screening of oral pre-malignancy and malignancy conditions [86]. hyper-methylation of the EDNRB gene has commonly occurred in NSCLC. Since the rate of EDNRB methylation is significantly higher in squamous cell carcinoma than adenocarcinomas, it can be used to distinguish SCCs from adenocarcinoma of the lung. since the downregulation of EDNRB after hypermethylation of the EDNRB gene is necessary for lung cancer tumorigenesis and is associated to tumor-related death [87]
FADS1	FADS1 rs174549 polymorphism is a useful factor for oral cancer PFS prediction, specifically in chemoradiotherapy patients. It can also be considered as a potential target for future of personalized treatment [88]
FAM65b	FAM65B can be considered as a suitable target for therapeutic approaches based on cancer stem cell elimination. The reason is that overexpression of FAM65B is observed in Prostate tumors such as PC3. These tumors have stem like characteristics. For example, they are pro-angiogenic and strongly self-renewal [89]
FAM89b	Fam89b is proposed to be a suitable target for chemotherapeutic strategies since it is a TGF- β pathway suppressor and signaling pathways induced by TGF- β have tumor-suppressing or tumor promoting effects based on type and stage of the cancer [90]
Gnai2	Previous studies have demonstrated GNAI2 as a main regulator of oncogenesis and an upstream driver of cancer development in the Ovarian cancer [91]. In addition to ovarian cancer, up-regulation of GNAI2 has also been observed in Hepatocellular Carcinoma. This protein acts through activation of the Ras-ERK/MAPK Mitogenic pathway by membrane recruitment of Rap1 GTPase-activating protein and moderation of GTP-bound Rap1 and also through the enhancement of cell survival by activation of AKT and inhibition of apoptosis by regulating Bcl-2 levels [92]

It was concluded that the machines which were trained using the features extracted from the Fisher Score algorithm performed better than the ones trained by the same set of features reduced using the PCA algorithm. The reason is due to the richer distinguishing information in the features selected by the Feature Score algorithm than the reduced features of the PCA algorithm. Exceptionally, the KNN algorithm performed similarly in both cases. The similarity of the KNN algorithm performance in both cases is due to the preservation of the data arrangement in the reduced space after the implementation of the PCA dimension reduction algorithm. Since the KNN algorithm decides the fate of a data based on its K nearest neighbors, maintenance of the data arrangement after the dimension reduction has led in the same results in both cases. Furthermore, the KNN and the Naïve bayes algorithms are proposed as the most appropriate machines for prediction. However, it should be noted that the appropriate machine must be selected based on the considered specific application.

Using the classifying features extracted, mir-target network and TF-gene network were constructed and enrichment and topology analyses were performed to

detect hub genes and hub TFs. Based on the degree centrality, PTGER3, YWHAH, CTNNB1, ANKRD50, EDNRB and ACSL6 target genes were detected as the chemoresistance hub genes according to mir-topology. PTGER3 is encoding PTGER3 protein, a member of the G-protein coupled receptor family. In this study, the degree of the PTGER3 was obtained to be seven which is higher than that of other obtained hub genes. Furthermore, this gene is also reported to be a cisplatin-resistant gene through Ras-MAPK/Erk-ETS1-ELK1/CFTR1 pathways [93]. YWHAH and CTNNB1 with the degree centrality of six were identified as the second robust hub genes in the row. It has been reported that after chemotherapy, YWHAH is upregulated in prostate cancer cells and is down regulated in Liposarcoma, representing the potency of this gene in chemoresistance [74, 75]. Moreover, it has been shown that CTNNB1 has a vital role in cancer regulatory pathways such as Gastric cancer signaling [94]. With a degree centrality of 5, ANKRD50 and EDNRB were the next obtained hub genes. According to previous studies, it has been reported that EDNRB-methylation is a very common phenomenon in NSCLC. Due to the higher rate of EDNRB methylation in Squamous Cell Carcinoma (SCCs), it can be used to distinguish between SCCs and lung Adenocarcinoma [87].

The TF-gene network topology analysis was also performed and the results specified two hub genes including IFNG and CTNNB1. IFNG is a protein coding gene and it is involved in Folate Metabolism. Furthermore, Yaghoobi et al. have evaluated the IFNG and its antisense (IFNG-AS1) roles in breast cancer and have proposed the involvement of IFNG and IFNG-AS1 in the pathogenesis of breast cancer [95]. In another study, Gao et al. investigated the role of IFNG pathway in the anti-CTLA resistance mechanism. Anti-CTLA-4 produces IFNG to enhance T cell responses. Their data revealed that defects in the IFNG signaling pathway leads to resistance to anti-CTLA-4 therapy [96].

With a systems biology approach including machine learning methods, feature selection, topological analysis, enrichment analysis and finally literature review, we managed to obtain a set of genes which play critical roles in chemoresistance processes. We also have nominated a set of potentially chemoresistance genes which could be used in further studies.

Conclusion

In this study, machine learning approach as well as systems biology analysis was used to extract the genes which commonly separated cisplatin resistant samples from the sensitive ones in lung, pancreatic and ovarian cancers. Furthermore, six classifiers were trained to distinguish between chemoresistance samples from the sensitive ones. As a result, KNN and Naïve Bayes algorithms were selected as the most practical machines according to a set of calculated measures. Moreover, the results of our systems biology analysis indicated several potential chemoresistance genes among which PTGER3, YWHAH, CTNNB1, ANKRD50, EDNRB, ACSL6, IFNG and, CTNNB1 are topologically more important than others. Our results have been validated against different databases such as UniProt, Enrichr and DIANA mirPath v.3 and the papers extracted from the literature. Therefore, this *in silico* study as well as its predictions can pave the way for further experimental researches.

Abbreviations

DEG: Differentially Expressed Genes; ODE: Ordinary differential equation; PDE: Partial differential equation; SVM: Support vector machines; PCA: Principal components analysis; ML: Machine learning; NSCLC: Non-small cell lung cancer; FS: Fisher Score; LOO: Leave one out.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40537-021-00477-z>.

Additional file 1. The extracted related pathway.

Additional file 2. The reported chemoresistance genes.

Acknowledgements

None

Authors' contributions

The specific contributions made by each author is as follows: AA: Conceptualization, Methodology, Implementation, Writing—Original Draft, Writing—Review & Editing. NSM: Methodology, Implementation, Validation, Writing—Review & Editing. JZ: Methodology, Implementation, Validation, Writing—Review & Editing. MR: Methodology, Implementation, Validation, Writing—Review & Editing. SSA: Conceptualization, Writing—Review & Editing. AAR: Conceptualization, Writing—Review & Editing. All authors read and approved the final manuscript.

Funding

None.

Availability of data and materials

The dataset used in this study can be obtained from the corresponding author on reasonable request.

Declaration

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Institute of Fundamental Medicine and Biology, Kazan Federal University, Kazan, Russia. ²Department of Biophysics, Faculty of Biological Sciences, Tarbiat Modares University, Tehran, Iran. ³Bioinformatics and Computational Omics Lab (BioCOOL), Department of Biophysics, Faculty of Biological Sciences, Tarbiat Modares University, Tehran, Iran. ⁴Department of Computer Engineering, University of Kurdistan, Sanandaj, Iran.

Received: 4 March 2021 Accepted: 24 May 2021

Published online: 05 July 2021

References

1. Mazumdar M, Glassman J. Categorizing a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Stat Med*. 2000;19(1):113–32.
2. Urruticoechea A, Alemany R, Balart J, Villanueva A, Viñals F, Capella G. Recent advances in cancer therapy: an overview. *Curr Pharm Des*. 2010;16(1):3–10.
3. Damin DC, Lazzaron AR. Evolving treatment strategies for colorectal cancer: a critical review of current therapeutic options. *World J Gastroenterol*: WJG. 2014;20(4):877.
4. Khalil DN, Smith EL, Brentjens RJ, Wolchok JD. The future of cancer treatment: immunomodulation, CARs and combination immunotherapy. *Nat Rev Clin Oncol*. 2016;13(5):273.
5. Raguz S, Yagüe E. Resistance to chemotherapy: new treatments and novel insights into an old problem. *Br J Cancer*. 2008;99(3):387–91.
6. Rebutti M, Michiels C. Molecular aspects of cancer cell resistance to chemotherapy. *Biochem Pharmacol*. 2013;85(9):1219–26.
7. Housman G, et al. Drug resistance in cancer: an overview. *J Cancers*. 2014;6(3):1769–92.
8. Uramoto H, Tanaka F. Recurrence after surgery in patients with NSCLC. *Transl Lung Cancer Res*. 2014;3(4):242.
9. Lippert TH, Ruoff H-J, Volm M. Intrinsic and acquired drug resistance in malignant tumors. *Arzneimittelforschung*. 2008;58(06):261–4.
10. Kelderman S, Schumacher TN, Haanen JB. Acquired and intrinsic resistance in cancer immunotherapy. *Mol Oncol*. 2014;8(6):1132–9.

11. Lloyd KL, Cree IA, Savage RS. Prediction of resistance to chemotherapy in ovarian cancer: a systematic review. *BMC Cancer*. 2015;15(1):117.
12. Sekine I, Shimizu C, Nishio K, Saijo N, Tamura T. A literature review of molecular markers predictive of clinical response to cytotoxic chemotherapy in patients with breast cancer. *Int J Clin Oncol*. 2009;14(2):112–9.
13. Cortazar P, Johnson BE. Review of the efficacy of individualized chemotherapy selected by in vitro drug sensitivity testing for patients with cancer. *J Clin Oncol*. 1999;17(5):1625–1625.
14. Fruehauf JP, Alberts DS. Assay-assisted treatment selection for women with breast or ovarian cancer. In: *Chemosensitivity testing in oncology*. Springer; 2003. p. 126–145.
15. Sekine I, Minna JD, Nishio K, Saijo N, Tamura T. Genes regulating the sensitivity of solid tumor cell lines to cytotoxic agents: a literature review. *Jpn J Clin Oncol*. 2007;37(5):329–36.
16. Sekine I, Minna JD, Nishio K, Tamura T, Saijo N. A literature review of molecular markers predictive of clinical response to cytotoxic chemotherapy in patients with lung cancer. *J Thorac Oncol*. 2006;1(1):31–7.
17. Słodkowska EA, Ross JS. MammaPrint™ 70-gene signature: another milestone in personalized medical care for breast cancer patients. *Expert Rev Mol Diagn*. 2009;9(5):417–22.
18. Camidge DR, Pao W, Sequist LV. Acquired resistance to TKIs in solid tumours: learning from lung cancer. *Nat Rev Clin Oncol*. 2014;11(8):473.
19. Sawyers C. Targeted cancer therapy. *Nature*. 2004;432(7015):294.
20. Sun X, Hu B. Mathematical modeling and computational prediction of cancer drug resistance. *Brief Bioinform*. 2017;19(6):1382–99.
21. Huang C, et al. Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy. *Sci Rep*. 2018;8(1):16444.
22. Ali M, Aittokallio T. Machine learning and feature selection for drug response prediction in precision oncology applications. *Biophys Rev*. 2019;11(1):31–9.
23. Chen R, Liu X, Jin S, Lin J, Liu J. Machine learning for drug-target interaction prediction. *Mol Cells*. 2018;23(9):2208.
24. Liu R, Zhang G, Yang Z. Towards rapid prediction of drug-resistant cancer cell phenotypes: single cell mass spectrometry combined with machine learning. *Chem Commun*. 2019;55(5):616–9.
25. Li H, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9.
26. A. Team RC. R: A language and environment for statistical computing. Vienna; 2013.
27. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118–27.
28. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9(5):2579–605.
29. Rostami M, Moradi P. A clustering based genetic algorithm for feature selection. In: *Information and Knowledge Technology (IKT)*. 2014. p. 112–116.
30. Moradi P, Rostami M. A graph theoretic approach for unsupervised feature selection. *Eng Appl Artif Intell*. 2015;44:33–45.
31. Moradi P, Rostami M. Integration of graph clustering with ant colony optimization for feature selection. *Knowledge Based Syst*. 2015;84:144–61.
32. Rostami M, Forouzandeh S, Berahmand K, Soltani M. Integration of multi-objective PSO based feature selection and node centrality for medical datasets. *Genomics*. 2020;112(6):4370–84.
33. Berahmand K, Haghani S, Rostami M, Li Y. A new attributed graph clustering by using label propagation in complex networks. *J King Saud Univ Comput Inf Sci*. 2020. <https://doi.org/10.1016/j.jksuci.2020.08.013>.
34. Rostami M, Berahmand K, Forouzandeh S. A novel community detection based genetic algorithm for feature selection. *J Big Data*. 2021;8(1):2.
35. Rostami M, Berahmand K, Forouzandeh S. A novel method of constrained feature selection by the measurement of pairwise constraints uncertainty. *J Big Data*. 2020;7(1):83.
36. Liu Y, Nie F, Gao Q, Gao X, Han J, Shao L. Flexible unsupervised feature extraction for image classification. *Neural Netw*. 2019;115:65–71.
37. Wang H, Zhang Y, Zhang J, Li T, Peng L. A factor graph model for unsupervised feature selection. *Inf Sci*. 2019;480:144–59.
38. Tang X, Dai Y, Xiang Y. Feature selection based on feature interactions with application to text categorization. *Expert Syst Appl*. 2019;120:207–16.
39. Wahid A, et al. Feature selection and classification for gene expression data using novel correlation based overlapping score method via Chou's 5-steps rule. *Chemom Intell Lab Syst*. 2020;199:103958.
40. Saeys Y, Inza I, Larrañaga P. review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;23(19):2507–17.
41. Yazdi KM, et al. Prediction optimization of diffusion paths in social networks using integration of ant colony and densest subgraph algorithms. *J High Speed Netw*. 2020;26:141–53.
42. Yazdi KM et al. Improving recommender systems accuracy in social networks using popularity. In: *2019 20th international conference on parallel and distributed computing, applications and technologies (PDCAT)*. 2019. p. 301–307.
43. Gao W, Hu L, Zhang P, He J. Feature selection considering the composition of feature relevancy. *Pattern Recognit Lett*. 2018;112:70–4.
44. Abdulla M, Khasawneh MT. G-Forest: An ensemble method for cost-sensitive feature selection in gene expression microarrays. *Artif Intell Med*. 2020;108:101941.
45. Rostami M, Berahmand K, Nasiri E, Forouzandeh S. Review of swarm intelligence-based feature selection methods. *Eng Appl Artif Intell*. 2021;100:104210.
46. Lever J, Krzywinski M, Altman N. Points of significance: principal component analysis. ed: Nature Publishing Group; 2017.
47. Lewis DD. Naive (Bayes) at forty: the independence assumption in information retrieval. Springer; 1998. p. 4–15.
48. Zhang Z. Introduction to machine learning: k-nearest neighbors. *Ann Transl Med*. 2016;4(11):218.
49. Barros RC, Basgalupp MP, Freitas AA, De Carvalho AC. Evolutionary design of decision-tree algorithms tailored to microarray gene expression data sets. *IEEE Trans Evol Comput*. 2013;18(6):873–92.

50. Díaz-Uriarte R, De Andres SA. Gene selection and classification of microarray data using random forest. *BMC Bioinform.* 2006;7(1):3.
51. Lancashire LJ, Lemetre C, Ball GR. An introduction to artificial neural networks in bioinformatics—application to complex microarray and mass spectrometry datasets in cancer studies. *Brief Bioinform.* 2009;10(3):315–29.
52. Elisseeff A, Pontil M. Leave-one-out error and stability of learning algorithms with applications. *NATO Sci Ser Sub Ser iii Comput Syst Sci.* 2003;190:111–30.
53. Heidaryan E. A note on model selection based on the percentage of accuracy-precision. *J Energy Resour Technol.* 2019. <https://doi.org/10.1115/1.4041844>.
54. Altman DG, Bland JM. Diagnostic tests. 1: sensitivity and specificity. *BMJ Br Med J.* 1994;308(6943):1552.
55. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics.* 2020;21(1):6.
56. Chen EY, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform.* 2013;14(1):128.
57. Xie B, Ding Q, Han H, Wu D. miRCancer: a microRNA–cancer association database constructed by text mining on literature. *Bioinformatics.* 2013;29(5):638–44.
58. Liu W, Wang X. Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. *Genome Biol.* 2019;20(1):1–10.
59. Vlachos IS, et al. DIANA-miRPath v3. 0: deciphering microRNA function with experimental support. *Nucleic Acids Res.* 2015;43(W1):W460–6.
60. Han H, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* 2018;46(D1):D380–6.
61. Shannon P, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13(11):2498–504.
62. Sekine Y, et al. The Kelch repeat protein KLHDC10 regulates oxidative stress-induced ASK1 activation by suppressing PPS. *Mol Cell.* 2012;48(5):692–704.
63. Zhong M, et al. Expression of MSP58 in hepatocellular carcinoma. *Med Oncol.* 2013;30(2):539.
64. Chae YK, et al. Genomic landscape of DNA repair genes in cancer. *Oncotarget.* 2016;7(17):23312.
65. Fischer F. The function of mismatch repair proteins in response to DNA damage caused by chemotherapeutic agents, University of Zurich; 2007.
66. Zhang D, et al. Regulation of the adaptation to ER stress by KLF4 facilitates melanoma cell metastasis via upregulating NUCB2 expression. *J Exp Clin Cancer Res.* 2018;37(1):176.
67. Qu S, et al. MicroRNA-330 is an oncogenic factor in glioblastoma cells by regulating SH3GL2 gene. *PLoS ONE.* 2012;7(9):e46010.
68. Yang Z, et al. GRSF1-mediated MIR-G-1 promotes malignant behavior and nuclear autophagy by directly upregulating TMED5 and LMNB1 in cervical cancer cells. *Autophagy.* 2019;15(4):668–85.
69. Vert A, Castro J, Ribo M, Vilanova M, Benito A. Transcriptional profiling of NCI/ADR-RES cells unveils a complex network of signaling pathways and molecular mechanisms of drug resistance. *Onco Targets Ther.* 2018;11:221.
70. Zheng P, Wang W, Muxi Ji QZ, Feng Y, Zhou F, He Q. TMEM119 promotes gastric cancer cell migration and invasion through STAT3 signaling pathway. *OncoTargets Ther.* 2018;11:5835.
71. Zheng P, et al. TMEM119 silencing inhibits cell viability and causes the apoptosis of gastric cancer SGC-7901 cells. *Oncol Lett.* 2018;15(6):8281–6.
72. Gheysarzadeh A, Bakhtiari H, Ansari A, Emami Razavi A, Emami MH, Mofid MR. The insulin-like growth factor binding protein-3 and its death receptor in pancreatic ductal adenocarcinoma poor prognosis. *J Cell Physiol.* 2019;234(12):23537–46.
73. Lee M, Cheung G, Nair R, Done S. Defining the roles of COIL and WIPI1 in breast cancer metastasis. ed: AACR; 2012.
74. Daigeler A, et al. Heterogeneous in vitro effects of doxorubicin on gene expression in primary human liposarcoma cultures. *BMC Cancer.* 2008;8(1):313.
75. Kibel AS, et al. Genetic variants in cell cycle control pathway confer susceptibility to aggressive prostate carcinoma. *Prostate.* 2016;76(5):479–90.
76. Sandhu V. A systems biology approach to integrated molecular analysis in pancreatic and periampullary adenocarcinoma. 2016.
77. Shih I-M, Nakayama K, Wu G, Nakayama N, Zhang J, Wang T-L. Amplification of the ch19p13.2 NACC1 locus in ovarian high-grade serous carcinoma. *Mod Pathol.* 2011;24(5):638.
78. Xia L, et al. ACP5, a direct transcriptional target of FoxM1, promotes tumor metastasis and indicates poor prognosis in hepatocellular carcinoma. *Oncogene.* 2014;33(11):1395.
79. Qi H, Liu S, Guo C, Wang J, Greenaway FT, Sun M. Role of annexin A6 in cancer. *Oncol Lett.* 2015;10(4):1947–52.
80. O'Sullivan D, et al. A novel inhibitory anti-invasive MAb isolated using phenotypic screening highlights AnxA6 as a functionally relevant target protein in pancreatic cancer. *Br J Cancer.* 2017;117(9):1326.
81. Shinmura K, et al. BSND and ATP6V1G3: novel immunohistochemical markers for chromophobe renal cell carcinoma. *Medicine.* 2015;94(24):e989.
82. Eo H-S, Heo JY, Choi Y, Hwang Y, Choi H-S. A pathway-based classification of breast cancer integrating data on differentially expressed genes, copy number variations and MicroRNA target genes. *Mol Cells.* 2012;34(4):393–8.
83. Khan K, Hardy R, Haq A, Ogunbiyi O, Morton D, Chidgey M. Desmocollin switching in colorectal cancer. *Br J Cancer.* 2006;95(10):1367.
84. Cui T, et al. Diagnostic and prognostic impact of desmocollins in human lung cancer. *J Clin Pathol.* 2012;65(12):1100–6.
85. Ladner RD. The role of dUTPase and uracil-DNA repair in cancer chemotherapy. *Curr Protein Pept Sci.* 2001;2(4):361–70.
86. Schussel J, et al. EDNRB and DCC salivary rinse hypermethylation has a similar performance as expert clinical examination in discrimination of oral cancer/dysplasia versus benign lesions. *Clin Cancer Res.* 2013;19(12):3268–75.
87. Chen S-C, et al. Aberrant promoter methylation of EDNRB in lung cancer in Taiwan. *Oncol Rep.* 2006;15(1):167–72.

88. Chen F, He B, Yan L, Qiu Y, Lin L, Cai L. FADS1 rs174549 polymorphism may predict a favorable response to chemoradiotherapy in oral cancer patients. *J Oral Maxillofac Surg.* 2017;75(1):214–20.
89. Zhang K, Waxman DJ. PC3 prostate tumor-initiating cells with molecular profile FAM65B high/MFI2 low/LEF1 low increase tumor angiogenesis. *Mol Cancer.* 2010;9(1):319.
90. Mironova N, Patutina O, Brenner E, Kurilshikov A, Vlassov V, Zenkova M. The systemic tumor response to RNase A treatment affects the expression of genes involved in maintaining cell malignancy. *Oncotarget.* 2017;8(45):78796.
91. Raymond JR, Appleton KM, Pierce JY, Peterson YK. Suppression of GNAI2 message in ovarian cancer. *J Ovarian Res.* 2014;7(1):6.
92. Jung-Yi Jiang R-JL, Lee S-J. A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification. *IEEE Trans Knowl Data Eng.* 2011;23(3):335–49.
93. Rodriguez-Aguayo C, et al. "PTGER3 induces ovary tumorigenesis and confers resistance to cisplatin therapy through up-regulation Ras-MAPK/Erk-ETS1-ELK1/CFTR1 axis," (in eng). *EBioMedicine.* 2019;40:290–304.
94. Tanabe S, Kawabata T, Aoyagi K, Yokozaki H, Sasaki H. "Gene expression and pathway analysis of CTNNB1 in cancer and stem cells," (in eng). *World J Stem Cells.* 2016;8(11):384–95.
95. Yaghoobi H, Azizi H, Oskooei VK, Taheri M, Ghafouri-Fard S. Assessment of expression of interferon γ (IFN- γ) gene and its antisense (IFNG-AS1) in breast cancer (in eng). *World Journal Surg Oncol.* 2018;16(1):211–211.
96. Gao J, et al. Loss of IFN- γ pathway genes in tumor cells as a mechanism of resistance to anti-CTLA-4 therapy. *Cell.* 2016;167(2):397–404.e9.

Publisher's Note

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

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)
