A computational model for anti-cancer drug sensitivity prediction

Zheming Zhao*, Kezhi Li*†, Chris Toumazou*† and Melpomeni Kalofonou*†
*Department of Electrical and Electronic Engineering, Imperial College London, SW7 2BT, UK
†Centre for Bio-Inspired Technology, Institute of Biomedical Engineering, Imperial College London, SW7 2BT, UK
Email:{zheming.zhao17, kezhi.li, c.toumazou, m.kalofonou}@imperial.ac.uk

***Abstract*—Various methods have been developed to build models for predicting drug response in cancer treatment based on patient data through machine learning algorithms. Drug prediction models can offer better patient data classification, optimising sensitivity identification in cancer therapy for suitable drugs. In this paper, a computational model based on Deep Neural Networks has been designed for prediction of anti-cancer drug response based on genetic expression data using publicly available drug profiling datasets from Cancer Cell Line Encyclopedia (CCLE). The model consists of several parts, including continuous drug response prediction, discretization and a drug sensitivity result output. Regularization and compression of neuron connections is also implemented to make the model compact and efficient, outperforming other widely used algorithms, such as elastic net (EN), random forest (RF), support vector regression (SVR) and simple artificial neural network (ANN) in sensitivity analysis and predictive accuracy.**

***Index Terms*—Cancer treatment, Drug sensitivity prediction, Computational model, Deep Neural Network**

I. Introduction

Cancer, a multifactorial disease with a highly heterogeneous nature has been at the forefront of research for decades, with emphasis to have been put on the development of more targeted cancer diagnostics and therapeutics. Since the 1980s, several cancer-related molecular features have been identified [1], advancing the knowledge based on which cancer therapies and drugs are being developed.

With the development of large-scale pharmacogenomic screening and profiling of cancer cell lines, precision medicine in cancer treatment (PMiCT) has become popular due to its effectiveness in tailoring medical treatment based on each patient's characteristics, offering insight in the role of molecular features that cancer cells exhibit and their correlation with predicting treatment outcome and therefore survival rates [2]. One important aspect of PMiCT is the computational prediction of drug responses based on multiple types of genome-wide molecular data using computational tools. Two of the most significant public resources for data profiling, CCLE and GDSC [3], offer information on cancer gene mutations, genetic expression data (Affymetrix, RNA-seq), copy number variations and drug sensitivity.

Recently, deep learning techniques have been used to aid processes in medical treatment and clinical decision support. However, in a classical statistic or machine learning (ML) regression model, the algorithms tend to make a moderate instead of an accurate prediction, restricted by the complexity of the model and the under study datasets. In 2017, Tiancheng *et al.* developed a software that can accurately diagnose breast cancer using deep neural networks (DNN) [4], demonstrating the effectiveness of DNN in dealing with large datasets and the potential of use in determining cancer treatment strategies.

In this paper, a comprehensive predictive model for anti-cancer drug sensitivity has been developed. The model uses the genetic expression profile as the input, and outputs a score on drug sensitivity for each patient case. It consists of several parts with a pan-cancer single-drug scheme, whereby in each part, 3 main functional sections are constructed: (i) pre-training feature selection, using the support vector regression based recursive feature elimination (RFE), (ii) continuous drug response prediction, with a multiple layer DNN and (iii) sensitivity discretization, with the interquartile range (IQR) to have been used to discretize and transform the drug response to sensitivity prediction. The constructed model has been regularized and compressed using the latest techniques in deep neural network to simplify computation. Compared to other methods such as elastic net (EN), random forest (RF) and support vector regression (SVR), the proposed method is more accurate in prediction and efficient in sensitivity analysis.

A. Previous work and under study dataset

Prior to the CCLE and GDSC projects, models were mainly built based on the NCI-60 database. Francesco *et al.* [5] proposed one response prediction model using genetic programming. In 2012, CCLE was published, allowing mapping of genetic features such as gene expression, copy number and gene mutations to anti-cancer drug response [2], with a designated drug response measurement, 'activity area', to have been introduced to simultaneously capture the efficacy and potency of a drug. Followingly, Jang *et al* [6] compared the performance of different regression and classification models and showed that (i) elastic net has a good prediction performance across all platforms and (ii) genetic expression is the most important input for assessing the role of drug response. An integrated approach was later proposed with the drug target inhibition being taken into consideration for IC50 values prediction, demonstrating a better performance compared to random forest based methods [7]. Other methods such as multitask learning [8] and dual-layer network [9] were also

978-1-5090-0617-5/19/$31.00 ©2019 IEEE

proved to have an improved performance compared to elastic net and other standard regression methods. Only recently, was the first application of machine learning in drug response prediction presented, using CCLE and GDSC datasets [10], offering new insight in the role of multiple molecular factors in prediction of cancer treatment.

II. Problem Formulation

The aim of the proposed model is to predict the anti-cancer drug sensitivity based on genetic expression and drug response data derived from the CCLE dataset, while achieving minimization of the errors of the predicted drug response, proposed as per below:

$$\text{Minimize} \quad J(f,\vartheta) = \frac{1}{n}\sum_{i=1}^{n}\left[y - f(\vartheta, x)\right]^2$$

where f denotes the model (mapping) we want to find, ϑ is the hyper-parameter in it; y and x present the anti-cancer drug response and genetic expression respectively. The genetic expression (RNAseq) and drug response data from CCLE were selected for model training. The genetic expression data contain 56318 genes from 1047 human cell lines, while the drug response data contain IC50, EC50, Amax and ActArea measurements of 24 anti-cancer drugs across more than 400 human cell lines for each of them respectively.

III. Method

The proposed model is based on a pan-cancer single-drug response prediction scheme that offers sensitivity prediction for a single therapeutic compound based on datasets derived from different cancer types. Each part consists of three blocks: feature selection, response (continuous) prediction and discretization, with the proposed model's overall structure to be illustrated in Fig. 1. Feature selection reduces the input dimension and filters the noisy features. Response on continuous prediction is constructed to predict the target ActArea values from the genetic expression input. The discretization part mainly focuses on transforming the continuous prediction results to the sensitivity prediction scores. Since our model scheme is based on pan-cancer single-drug datasets (different entries from different cancer types are used for the prediction of a single drug response), for each drug an independent model is trained using the proposed model.

A. Feature Selection

Recursive feature elimination (RFE) was used, utilising an external estimator (usually a model) that assigns weights to features (e.g. the coefficients of a linear model). This algorithm selects smaller feature sets recursively, which means that in each iteration a smaller number of features remain until the desired number of features is reached. The advantage of RFE is that it usually selects the most informative features with respect to the target, while maintaining a strong potential relationship between target and features, reducing the dimensions by deleting the irrelevant features. For the estimator, the epsilon-support vector regression (SVR) was

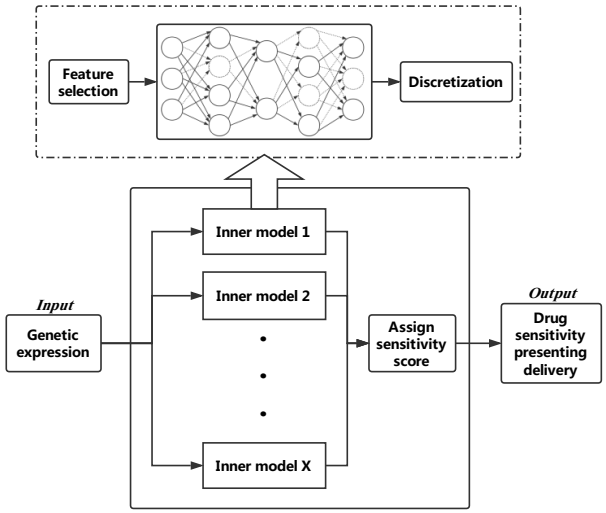

Fig. 1. The proposed overall computational model

chosen to be used for the scope of this work, based on the principle of solving the optimisation problem as below: $minimize \quad \frac{1}{2}\|w\|^2 \quad \text{subject to} \quad \begin{cases} y - \hat{y} \le \epsilon \\ \hat{y} - y \le \epsilon \end{cases}$ where $\hat{y}$ denotes the predicted target, $\hat{y} = \langle w, x\rangle + b$, $\langle\ ,\ \rangle$ denotes the dot production, and ϵ is the designated margin (tolerance). For feature selection, the ϵ margin in the SVR estimator was set at 0.1 and the desired feature number was set at 1000.

B. Deep Neural Network

A DNN has been developed to process the features after dimension reduction. The proposed DNN has 12 dense layers which are adopted in the model. Other structures have been tested, including convolutional neural network (CNN) and recurrent neural network (RNN). Empirically, the conventional dense layers achieve a better result. In this model, tanh and ReLU are set as the activation functions alternately to combine the advantages of both nonlinear functions.

1) Hyper-parameters Optimization: In the neural network construction, a heuristic optimization approach has been used to decide on the parameters, with input data to be split at 70% for training and 30% for validation. Firstly, one hidden layer with 10 nodes and 'ReLU' activation function in each node was initialized. The maximum number of nodes for each layer was set at 200. The number of nodes was then increased in the current layer by 10. If the mean square error (MSE) of an unseen (validating) set improved, the function was repeated; if the MSE could not be improved or the maximum number of nodes was reached, the current node number was kept the same. The number of layers could also increase with the addition of one more hidden layer with 10 nodes through the use of the 'ReLU' activation function. Finally, in the training phase, the *Adam* [11] optimization algorithm was used, due to the combining advantages of AdaGrad and RMSProp and its computational efficiency and ease of implementation.

2) Regularization and Compression: In order to improve the predictive accuracy, avoid overfitting and make the model

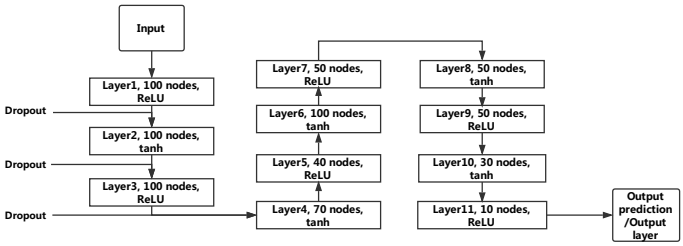

Fig. 2. The proposed optimized neural network.

more efficient, several regularization and compression techniques have been applied. In detail, l_1 and l_2 regularizations were adopted in the cost function, as they make values of weight matrices decrease and sparse. Also, *Dropout* [12], a regularization technique that randomly drops units (along with their edges) with a pre-defined probability *p*, was applied to the first 3 hidden layers, with a dropping rate *p* set at 0.4, reducing the overfitting problem significantly in several individual datasets. Finally, a compression technique introduced in [13] was applied to the neural network. After pruning, trained quantization and huffman coding, the model was further simplified, while maintaining similar performance. Following the aforementioned processes used for computation and process, the resulted DNN is summarized in Fig. 2.

C. Discretization

In order to assess drug sensitivity and therefore define drug response effectiveness, discretization of the ActArea values was performed, with the resulting scores defined as 'sensitive', 'resistant' or 'moderate'. To avoid errors caused by outliers in the measurements, the upper and lower quartiles were used in the discretization scheme as follows: first the distribution and the 25, 75 percentile values were defined for each of the 24 drugs; if an ActArea value was equal or larger than the 75 percentile value, it was marked as sensitive; if the ActArea value was equal or less than the 25 percentile value, it was marked as resistant, otherwise, the ActArea value was marked as moderate.

IV. Results

A. Preprocessing

With the help of the Python packages *cmapPy* and *pandas*, the selected data files from CCLE were imported and visualized. In the dataset, the index of the input matrix was identified by the cell line name, with each name to be defined by (i) the sample name and (ii) the tissue name (cancer type). All the index annotation details can be found in the ATCC database [14]. The columns identify human genes, with their primary annotations and names defined in Zerbino *et al.*'s study [15]. For the drug response set, the ActArea is selected from four recorded response measurements: EC50, IC50, Amax and ActArea, with all extracted from drug effectiveness results of *in vitro* experiments. Cell lines that did not have a measured ActArea value and measured responses without a corresponding genetic expression value were removed from the analysis. For the expression set alone, genes that had zero expression values across all cell lines were also deleted. In this work, min-max normalization was used to scale the range of features to $(0, 1)$ interval.

B. Continuous prediction

The performance of our proposed NN based model was compared to the elastic net (EN) regression, random forest (RF) regression, simple ANN with random feature selections (with same number of dense layers and linear activation function). The performance of each algorithm was measured by the mean squared error (MSE) and validated using Monte Carlo cross-validation. In the Monte Carlo cross-validation, the full training dataset was partitioned by random sampling into two subsets: a training set and a validating (or testing) set. The proportion of the data going into each subset was defined as 70% of the data for training and the rest 30% for testing, which led to an average of 130 samples in the testing set. This process was repeated 5 times, as each time a new partition was created for each independent model training for each drug. The MSEs were calculated for each test set, with the average overall performance to be shown in Table I.

Our constructed model achieved an average 0.1538 in MSE measurement for all 24 drugs, ranging from a minimum of 0.0714 to a maximum of 0.2689. Compared to the other methods implemented as part of this investigation, the proposed model strongly outperformed each one of them in all 24 drugs. The average reduction in MSE for our neural network based model over elastic net across 24 drugs exceeded 50% and achieved maximum 78.1% for drug PD-0325901. Also, for the drugs 17-AAG, Topotecan, PD-0325901, Paclitaxel and AZD6244, whereby the rest 4 methods showed noticable predicted errors, our model could still capture the pattern and demonstrate a great performance of an average of 0.2187 in MSE.

C. Sensitivity prediction

With respect to the continuous prediction, the sensitivity prediction of each drug was derived as follows: first the original ActArea values of all 24 drugs were discretized into 3 categories (sensitive, resistant and moderate) based on their IQR; the predicted continuous results were then classified based on the same condition of the original values classification under the assumption that the response distribution remains unchange. The prediction accuracy of all 24 drugs can be seen in Table II, with the results to verify the improved performance of the proposed algorithm.

D. Timing aspect

Fig. 3 shows the accuracy and the training time specifications between our proposed method and the rest of the 4 tested algorithms. The training time is measured in seconds/each drug and then processed in logarithmic transformation. Compared to elastic net, our proposed method not only performed better in terms of accuracy but also reduced the training time by a large proportion.

TABLE I
THE PREDICTIVE MSE ACROSS 24 DRUGS OVER RANDOM FOREST, SVR, ELASTIC NET, SIMPLE NEURAL NETWORK AND THE PROPOSED METHOD (PM)

Drug name	RF	SVR	EN	NN	PM	Drug name	RF	SVR	EN	NN	PM
AEW541	0.358	0.324	0.293	0.426	**0.126**	Irinotecan	0.683	0.694	0.626	1.071	**0.107**
Nilotnib	0.522	0.521	0.456	0.525	**0.215**	Topotecan	1.060	1.013	0.915	1.468	**0.152**
17-AAG	1.079	0.933	0.899	1.108	**0.206**	LBW242	0.375	0.462	0.500	0.358	**0.140**
PHA-665752	0.223	0.225	0.233	0.241	**0.071**	PD-0325901	1.716	1.039	1.230	2.388	**0.268**
Lapatinib	0.347	0.298	0.313	0.448	**0.121**	PD-0332991	0.287	0.346	0.288	0.373	**0.118**
Nutlin-3	0.268	0.262	0.245	0.243	**0.123**	Paclitaxel	1.651	1.387	1.091	2.103	**0.256**
AZD0530	0.693	0.583	0.611	0.679	**0.178**	AZD6244	1.048	0.882	0.974	1.470	**0.210**
PF2341066	0.314	0.307	0.291	0.389	**0.156**	PLX4720	0.510	0.514	0.334	0.579	**0.226**
L-685458	0.313	0.290	0.303	0.376	**0.146**	RAF265	0.569	0.539	0.521	0.560	**0.100**
ZD-6474	0.556	0.569	0.472	0.576	**0.138**	TAE684	0.724	0.594	0.593	0.732	**0.162**
Panobinostat	0.465	0.457	0.466	0.736	**0.105**	TKI258	0.420	0.339	0.299	0.512	**0.157**
Sorafinib	0.244	0.218	0.273	0.238	**0.080**	Erlotinib	0.355	0.320	0.341	0.437	**0.121**

TABLE II
SENSITIVITY PREDICTION IN PERCENTILE SCORES

From all 24 drugs	RF	SVR	EN	ANN	PM
Minimum	0.422	0.446	0.451	0.439	**0.642**
25 perc.	0.467	0.479	0.489	0.475	**0.720**
Mean	0.513	0.520	0.528	0.521	**0.763**
50 perc. (Median)	0.510	0.520	0.531	0.521	**0.771**
75 perc.	0.541	0.553	0.562	0.551	**0.807**
Maximum	0.589	0.591	0.603	0.597	**0.872**

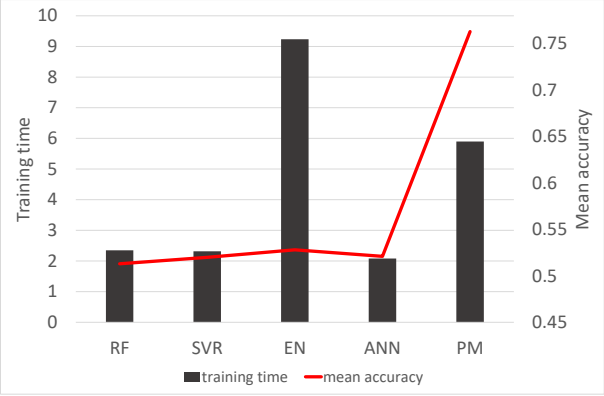

Fig. 3. Training and inference time comparison

V. CONCLUSION

In this work, a computational model for prediction of cancer drug sensitivity was developed and tested on a range of genetic expression derived from a drug profiling dataset (CCLE). The proposed model, based on the principle of Deep Neural Networks, performs feature selection, continuous prediction and discretization, using the genetic profiling information as the model's input, calculating a drug sensitivity score as the output (sensitive, resistant or moderate). Comparison metrics were also generated to evaluate the performance of the model for prediction of 24 drugs, demonstrating better prediction efficiency compared to all other methods, with a maximum 78% reduction in MSE compared to the elastic net method. The model achieved the best mean sensitivity in prediction with an accuracy of 0.7631, while reducing considerably the processing training time.

ACKNOWLEDGEMENT

The authors would like to acknowledge the Cancer Research UK Multidisciplinary Award (C54044/A25292) for supporting this research.

REFERENCES

[1] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer Genome Landscapes," *Science*, vol. 339, no. 6127, pp. 1546–1558, 2013.

[2] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, et al., "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, 2012.

[3] N. Stransky, M. Ghandi, G.V. Kryukov, L.A. Garraway, J. Lehár, et al., "Pharmacogenomic agreement between two cancer cell line data sets," *Nature*, vol. 528, no. 7580, pp. 84, 2015.

[4] T. He, M. Puppala, R. Ogunti, J.J. Mancuso, X. Yu, et al., "Deep learning analytics for diagnostic support of breast cancer disease management," *IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 365–368, 2017.

[5] F. Archetti, I. Giordani, and L. Vanneschi, "Genetic programming for anticancer therapeutic response prediction using the NCI-60 dataset," *Computers & Operations Research*, vol. 37, no. 8, pp. 1395–1405, 2010.

[6] I.S. Jang, E.C. Neto, J. Guinney, S.H. Friend, and A.A. Margolin, "Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data.," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 63–74, 2014.

[7] N. Berlow, S. Haider, Q. Wan, M. Geltzeiler, L.E. Davis, C. Keller, and R. Pal, "An Integrated Approach to Anti-Cancer Drug Sensitivity Prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 6, pp. 995–1008, 2014.

[8] H. Yuan, I. Paskov, H. Paskov, A.J. González, and C.S. Leslie, "Multitask learning improves prediction of cancer drug sensitivity," *Scientific Reports*, vol. 6, no. 1, pp. 31619, 2016.

[9] N. Zhang, H. Wang, Y. Fang, J. Wang, X. Zheng, and X.S. Liu, "Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug Network Model," *PLOS Computational Biology*, vol. 11, no. 9, pp. e1004498, 2015.

[10] M.Q. Ding, L. Chen, G.F. Cooper, J.D. Young, and X. Lu, "Precision oncology beyond targeted therapy: Combining omics data with machine learning matches the majority of cancer cells to effective therapeutics," *Molecular Cancer Research*, vol. 16, no. 2, pp. 269–278, 2018.

[11] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[12] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[13] S. Han, H. Mao, and W.J. Dally, "Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding," *CoRR*, vol. abs/1510.00149, 2015.

[14] "ATCC: The Global Bioresource Center," [Online]. https://www.atcc.org/.

[15] D.R Zerbino, P. Achuthan, W. Akanni, M.R. Amode, et al., "Ensembl 2018," *Nucleic Acids Research*, vol. 46, no. D1, pp. D754–D761, 2017.