Zaidi et al. LGURJCSIT 2024

ISSN: 2521-0122 (Online)
ISSN: 2519-7991 (Print)

LGU Research Journal of Computer Science & IT

doi: 10.54692/lgurjcsit.2024.081551

Vol (8): Issue (1), January – March 2024

Developing MLP based prediction system for anticancer drug response using hybrid features of genomics and cheminformatics

Awais Raza Zaidi[1*], Muhammad Bilal[2], Tuba Majid[3], Abdul Majid[4]
[1,2,4]Biomedical Informatics Research Lab, Department of Computer & Information Sciences, Pakistan Institute of Engineering & Applied Sciences, Nilore, Islamabad, Pakistan.
[3]Experimental Continuum Mechanics Research Group, Department of Mechanical and Process Engineering, ETH Zurich, 8092 Zürich, Switzerland.

Email: awaisraza_19@pieas.edu.pk

ABSTRACT:

Traditional cancer treatment methods have become less effective due to the increasing diversity of cancer types. To address this, precision medicine has gained support within the medical community. This approach tailors treatment to individual patients based on their specific disease characteristics. However, a major challenge lies in accurately predicting how a patient will respond to a specialized drug. Numerous machine learning-based predictive systems have been developed to address this challenge. These systems utilize genomic signatures and the chemical structure of drugs to predict drug activity. In this paper, we introduce a Multi-Layer Perceptron (MLP) based system for predicting the response of anticancer drugs. Our system utilizes hybrid features derived from both genetic expression and the chemical structure of drugs. It is developed using the well-known GDSC dataset (Genomics of Drug Sensitivity in Cancer). Our system achieved a lower Root Mean Square Error (RMSE) value of 0.889, in contrast to the RMSE value of 0.983 obtained by the current state-of-the-art (SOTA) system, SwNet. This indicates superior predictive accuracy. The findings suggest that our proposed research holds promise for the development of targeted drugs for anticancer treatments.

KEYWORDS: MLP, GDSC, Precision Medicine, Gene Expression, SMILES

1. INTRODUCTION

Cancer ranks among the leading causes of mortality globally, with approximately 19.3 million new cases and around 10 million cancer-related deaths reported in 2020 [1,2]. Given its significant impact, cancer has been a focal point of both biological and clinical investigations. Scientists have long explored the underlying mechanisms of cancer development, particularly the interplay between genetic and epigenetic factors. Notable indicators of cancer include dysfunctional gene activity and altered gene expression patterns. Increasing evidence suggests that acquired epigenetic irregularities, alongside genetic mutations, contribute to the onset of cancer [3]. Mainstream cancer treatments encompass surgery, cytotoxic chemotherapy, targeted therapy, radiation therapy, endocrine therapy, and immunotherapy [4]. Despite advancements, patients, especially those with advanced cancer, often face relapse following treatment [5]. Immune resistance to conventional chemotherapy and medications remains a significant challenge in cancer treatment. Traditional chemotherapeutic drugs function by damaging cancer cell DNA, but their non-specific nature and high toxicity pose limitations. Until recently, cancer treatments were uniformly prescribed based solely on the cancer type, adopting a one-size-fits-all approach. For instance, clinical trials across 15 different cancer types led to the approval of Pembrolizumab for treating solid tumors with high microsatellite

instability or mismatch repair deficiency [6]. Another approved medication, Larotrectinib, targets the tropomyosin receptor kinase gene fusion prevalent in numerous cancers [7]. However, the scarcity of comprehensive cancer datasets often necessitates the utilization of conventional treatments [8]. Consequently, there is a decline in prognoses and a subsequent rise in cancer mortality rates [9]. However, the introduction of personalized medicine has revolutionized cancer treatment methodologies. With vast genomic databases now accessible, precision oncology has emerged as a treatment approach. This method takes into account a patient's genetic makeup and biomarkers when formulating therapeutic recommendations. The first FDA approval for such personalized treatment occurred in 2017 when Pembrolizumab was sanctioned for treating solid tumors with high microsatellite instability or mismatch repair deficiency. Clinical trials across 15 different cancer types supported this approval [10]. Another effective medication, Larotrectinib, targets the tropomyosin receptor kinase gene fusion found in numerous tumors [4]. However, many anticancer drug molecules lack well-established biomarkers, posing a challenge. Even targeted therapies, in addition to commonly used

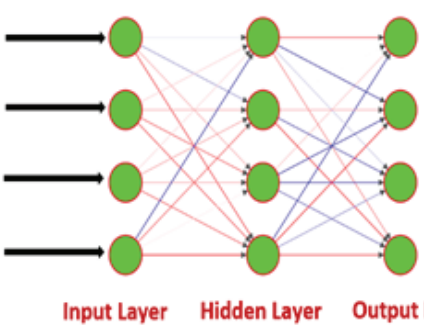

Figure1: Schematic representation of a Multi-Layer Perceptron (MLP) with a single Hidden Layer

cytotoxic drugs, face difficulties in identifying accurate biomarkers. This is because pharmacological targets alone often fall short as therapeutic indicators [11].

The paper is structured as follows: In Section 2, we outline the methodologies employed by previous researchers for drug response prediction. Section 3 details our proposed methodology. Section 4 presents the experimental results, followed by the conclusion.

2. RELATED WORK

In the fields of bioinformatics and chemo-informatics, machine learning has emerged as a crucial research tool. Utilizing statistical, probabilistic, and optimization techniques, machine learning models analyze complex patterns within large, noisy datasets [5]. These models can be utilized for various purposes, including disease detection, diagnosis, prognosis, and drug discovery. Given advancements in genomics and molecular biology, personalized medicine has gained traction as researchers seek correlations between an individual's biological characteristics and treatment responses (biomarkers). The primary focus of cancer therapy is the identification of biomarkers [12]. The computational biology community has made substantial advancements in constructing predictive models that correlate an individual's genomic data with pharmacological responses, spurred by the availability of large datasets. Numerous machine learning techniques have been explored for building these models, including kernel models, sparse linear/non-linear models, elastic net regularized matrix factorization, network-based models, and ensemble models [13]. A simplified representation of a feedforward neural network, depicted in Figure 1, illustrates the data flow from the input layer through hidden layers to the output layer, providing predictive results. This research has yielded promising results, demonstrating the efficacy of algorithms, the predictive potential of various genomic data.

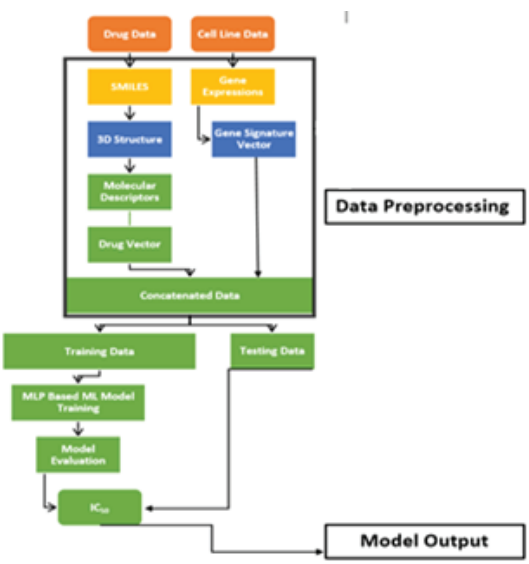

Figure 2: the overall architecture of proposed prediction model

Figure 2 shows, the overall architecture of our proposed prediction model. The chemical structure in SMILE format and gene expression features serve as inputs. Both input vectors are concatenated and fed to the MLP for prediction of the drug's IC50 values for the given cell line.formats, and the advantages of leveraging prior knowledge [14]. Certain studies have integrated drug chemical structures as input for predicting medication bioactivity [15], while others have combined genomic data and chemical signatures in prediction models, resulting in high accuracy in forecasting IC50 values [16,17].

2.1. Proposed prediction system

In this paper, we introduce a Multi-Layer Perceptron (MLP) based system for predicting anticancer drug response. Our system utilizes hybrid features derived from genetic expression and drug chemical structure. The predictive model is developed using the Genomics of Drug Sensitivity in Cancer (GDSC) dataset. Through comparative analysis, we validate the enhanced performance of our model compared to other existing models. The multi-layer perceptron (MLP) is an adaptation of feed-forward neural networks, featuring input, hidden, and output layers [8]. Input data is received by the input layer, while the output layer handles tasks like prediction and classification. Hidden layers, situated between input and output layers, constitute the computational core of the MLP. Data in an MLP flows from the input layer to the output layer, akin to feed-forward neural networks. Training of MLP neurons employs the backpropagation learning algorithm. MLPs excel at solving non-linear problems and can approximate any continuous function. Main applications of MLP include pattern categorization, recognition, prediction, and approximation. See Figure 1 for a schematic representation of a Multi-Layer Perceptron with a single hidden layer. We have constructed a machine learning model utilizing the Multi-Layer Perceptron framework. The model integrates two distinct datasets: one containing drug molecular data, and the other comprising patient cell line information represented by various gene expressions. Our model is capable of forecasting the body's reaction to different cell lines by computing the IC50 values of administered drugs, which signify the minimal concentration inducing 50% cell death.

2.2. Datasets Preprocessing stage

The initial phase of the proposed methodologyinvolves preprocessing the input datasets. Specifically, we retrieve data from the Genomics of Drug Sensitivity in Cancer (GDSC) dataset [9]. This widely recognized dataset provides comprehensive information for studying cancer biology and predicting medication responses in individual patients. The GDSC dataset encompasses over 75,000 experiments examining the effects of 138 anticancer drugs on more than a thousand cell lines from diverse cancer types. Additionally, baseline data includes information on gene copy number, expression, and somatic mutations in 75 cancer-related genes [18]. The chemical structures of compounds listed in GDSC are sourced from PubChem [19]. To preprocess the input data for our model, we initially acquired the Simplified Molecular Input Line Entry System (SMILES) representation of compounds from PubChem. SMILES is a format used to represent three-dimensional structural information in a machine-readable manner. Subsequently, we converted the SMILES representation of each compound into Morgan Fingerprints using RDKit [20]. This conversion aimed to align the compounds with a mathematical notation. We then calculated the similarity between chemical compounds using the Tanimoto coefficient. Drug sensitivity with respect to its corresponding cell lines was assessed by predicting the IC50 values of the compounds.

2.3. Model development stage

In the second stage of our proposed approach, we trained Multi-layer Perceptron (MLP) models using two sets of input data: molecular structures of chemical drugs represented by Morgan Fingerprints, and gene expression data from GDSC. Our model architecture comprises a 5-layer MLP neural network with specific neuron counts in each layer: 800 neurons in the first layer, 400 neurons in the second layer, 200 neurons in the third layer, 50 neurons in the fourth layer, and 1 neuron in the fifth (final) layer. We opted for the Parametric Rectified Linear Unit (PReLU) activation function for improved generalization, which outperformed other activation functions. A formal mathematical definition for PReLU is presented follow.

$$f(y_i) = \begin{cases} y_i, & \text{if } y_i > 0 \\ a_i y_i, & \text{if } y_i \leq 0 \end{cases} \qquad (1)$$

Here yi is any input at the ith channel, ai is negative slope:
if a_i=0, f behaves as ReLU
if a_i>0, f behaves leaky ReLU
if a_i is a learnable parameter, f behaves like PReLU.

$$f(yi) = \{^{yi}_{aiyi} \quad if \ yi \tag{2}$$

Given the large dataset volume, we applied Batch Normalization and included a dropout layer to prevent overfitting. To evaluate the model's performance, we employed 10-fold cross-validation during training and testing.

2.4. Results and discussion

To assess the performance of our proposed system against existing models, we selected four previously published state-of-the-art models that were trained and tested on the GDSC dataset. Our model demonstrated superior accuracy in predicting IC50 values compared to these models. The relative performance of our model compared to others, measured in terms of mean square error (MSE), is summarized in Table 1.
Table 1 performance comparison with previously models, bold results are denoting best among all participating in comparison.

Table 1: performance comparison with previously models, bold results are denoting best among all participating in comparison

Dataset	Model	Avg. MSE
GDSC	XGBoost	1.256
GDSC	**MLP**	**0.889**
GDSC	SwNet	0.983
GDSC	SRMF	0.987
GDSC	KMBTL	1.2642

• Kernelized Bayesian Multi-Task Learning (KBMTL): This method combines binary classification or regression with kernel-based non-linear dimensionality reduction to create a novel Bayesian approach [21].
• Similarity Regularization Matrix Factorization (SRMF): Utilizing chemical structures of drugs and baseline levels of gene expression in various cell lines, this model predicts the anticancer drug responses of respective cell lines [22].

• Self-Attention Gene Weight Layer Network (SWnet): SWnet compiles the chemical drugs dataset using a graph neural network (GNN) training model, while the cell line dataset is trained with a convolutional neural network (CNN). Both datasets are then merged into a single dataset for predicting IC50 values [23].
In our study, we conducted several experiments to evaluate the performance of our machine learning model. Initially, we trained our system using XGBoost and optimized key parameters to enhance prediction accuracy. The best result achieved with this model was a mean square error (MSE) of 1.256. In the second experiment, we trained our model using an MLP-based approach, which resulted in an improved MSE of 0.956. Finally, we applied 10-fold cross-validation to further refine our MLP model, achieving the lowest MSE of 0.889. Our proposed machine learning model demonstrated superior performance, with a lower MSE of 0.889 compared to other prediction models, including SRMF, KBMTL, XGboost, and SWnet, as depicted in Table 1.

3. CONCLUSION

In this study, we introduced an MLP-based machine learning model that utilizes two distinct types of data from the GDSC dataset: gene expressions and chemical structures of drugs. The chemical structural information is transformed into mathematical descriptors derived from SMILES notation and 3D chemical structures. Our model demonstrates enhanced accuracy in predicting IC50 values compared to previously published state-of-the-art (SOTA) models, achieved a Root Mean Square Error (RMSE) value of 0.889, in contrast to the RMSE value of 0.983 obtained by the current state-of-the-art (SOTA) system, SwNet. This indicates superior predictive accuracy. The findings suggest that our proposed research holds promise for the development of targeted drugs for anticancer treatments, with the continual growth of available data, our aim is to deploy our model on larger datasets in the future to further improve its performance. Our model represents a significant advancement for researchers in the field of precision medicine and cancer therapy.

REFERENCES

[1] H. Sung et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians,*

71(3), pp. 209-249, 2021.

[2] B. W. K. P. Stewart and C. P. Wild, *International agency for research on cancer. World cancer report, 2014.*

[3] S. B. Baylin et al., "Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer," *Human molecular genetics*, 10(7), pp. 687-692, 2001.

[4] Corrado, G., Salutari, V., Palluzzi, E., Distefano, M. G., Scambia, G., & Ferrandina, G. (2017). Optimizing treatment in recurrent epithelial ovarian cancer. Expert review of anticancer therapy, 17(12), 1147-1158.

[5] P. G. Casali, "Adjuvant chemotherapy for soft tissue sarcoma," *American Society of Clinical Oncology Educational Book,* 35(1), pp. 629-633, 2015.

[6] X. Wang et al., "Drug resistance and combating drug resistance in cancer," *Cancer Drug Resistance,* 2(2), pp. 141, 2019.

[7] J. Baselga et al., "AACR cancer progress report 2015," *Clinical Cancer Research,* 21(19_Supplement), S1-S128, 2015.

[8] N. Federman and R. McDermott, "Larotrectinib, a highly selective tropomyosin receptor kinase (TRK) inhibitor for the treatment of TRK fusion cancer," *Expert Review of Clinical Pharmacology,* 12(10), pp. 931-939, 2019.

9) L. Marcus et al., "FDA Approval Summary: Pembrolizumab for the Treatment of Microsatellite Instability-High Solid Tumors FDA Approval Summary: Pembrolizumab for MSI-H Solid Tumors" *Clinical Cancer Research,* 25(13), pp. 3753-3758, 2019.

10) D. Chakravarty et al., "OncoKB: a precision oncology knowledge base," *JCO precision oncology,* 1, pp. 1-16, 2017.

11) P. Geeleher et al., "Clinical drug response can be predicted using baseline gene expression levels and in vitrodrug sensitivity in cell lines," *Genome biology,* 15(3), pp. 1-12, 2024.

12) Z. Dong et al., "Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection," *BMC cancer,* 15(1), pp.1-12, 2015.

13) L. J. Scott, "Larotrectinib: first global approval," *Drugs,* 79(2), pp. 201-206, 2019.

14) M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nature Reviews Genetics,* 16(6), pp. 321-332, 2015.

15) P. Larranaga et al., "Machine learning in bioinformatics," *Briefings in bioinformatics,* 7(1), pp. 86-112, 2006.

16) H. Bhaskar et al., "Machine learning in bioinformatics: A brief survey and recommendations for practitioners," *Computers in biology and medicine,* 36(10), 1104-1125, 2006.

17) P. Raj et al., "The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases," *Academic Press,* 2020.

18) M. J. Garnett et al., "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature,* 483(7391), pp. 570-575, 2012.

19) J. Barretina et al., "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature,* 483(7391), pp. 603-607, 2012.

20) S. Kim et al., "PubChem substance and compound databases," *Nucleic acids research, 44(D1)*, pp. D1202-D1213, 2016.

21) M. Gönen and A. A. Margolin, "Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning," *Bioinformatics,* 30(17), pp. i556-i563, 2014.

22) L. Wang et al., "Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization," BMC cancer, 17(1),pp. 1-12, 2017.

23) Z. Zuo et al., "SWnet: a deep learning model for drug response prediction from cancer genomic signatures and compound chemical structures," *BMC bioinformatics,* 22(1), pp. 1-16, (2021).