ORIGINAL ARTICLE

Journal Section

Identification of Biological Markers in Cancer Disease using Explainable AI

Muhammad Shahzad[1] | Ruhal Lohana[1] | Khursheed Aurangzeb[2] | Isbah Imtiaz Ali[1] | Muhammad Shahid Anwar[3] | Mahnoor Murtaza[1] | Rauf Ahmed Shams Malick[1] | Piratdin Allayarov[4]

[1]FAST School of Computing, National University of Computer and Emerging Sciences (NUCES-FAST), Karachi 75030, Pakistan (emails: mshahzad@nu.edu.pk)

[2]Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, P. O. Box 51178, Riyadh 11543, Saudi Arabia (email:kaurangzeb@ksu.edu.sa)

[3]Department of AI and Software Gachon University Seongnam-si, 13120, South Korea (e-mail: shahidanwar786@gachon.ac.kr)

[4]Tashkent State University of Economics, 100063, Uzbekistan (email: p.allayarov@tsue.uz)

Correspondence
Muhammad Shahid Anwar
Email: shahidanwar786@gachon.ac.kr

Funding information
This Research is funded by Researchers Supporting Project Number (RSPD2023R947), King Saud University, Riyadh, Saudi Arabia

The research aims to improve the prediction of drug sensitivity on cancer cell lines using gene expression data and molecular fingerprints of drugs. The proposed study uses a deep learning model, **BioMarkerX**, trained on the Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC) datasets utilizing Particle Swarm Optimization (PSO) technique to select specific genes as features. The model achieves high prediction accuracy with an RMSE of 0.40 ± 0.02 and R2 of 0.83 ± 0.03 on the CCLE dataset, and an RMSE of 0.36 ± 0.05 and R2 of 0.83 ± 0.03 on the GDSC dataset. The approach also used an explainable artificial intelligence (XAI) model to discover biological markers linked to cancer development. This can provide insights into targeted therapies for improving cancer treatment outcomes. Overall, the study presents an effective approach for identifying important biological markers relevant to cancer disease, aiding in the development of more efficient anticancer medications.

KEYWORDS
Cell lines, Explainable AI, Drug Sensitivity, Deep Learning, Cancers, Metaheuristic Algorithms.

1 | INTRODUCTION

Cancer cell lines cultivated in a controlled environment possess diverse genomic backgrounds and gene expressions. These cell lines are crucial tools for studying the molecular processes that underlie therapeutic efficacy and for discovering novel anticancer drugs in the field of cancer biology [1]. Initially, computer models known as Quantitative Structure-Activity Relationship (QSAR) approaches were created using a collection of fixed cell lines or tissues [1]. These models depended on pharmacological attributes and worked under the premise that substances with similar chemical and structural properties would have comparable biological effects on established cell lines. The ability of QSAR models to generalise across different cancer cell lines remained constrained, despite the fact that they were heavily used for chemical space exploration and the discovery of novel anticancer drugs. This restriction resulted from treating drug response prediction as a single-task learning issue without taking cell line characteristics into account.

Therefore in the pursuit of personalized cancer treatment, the development of machine learning models has facilitated the prediction of drug response based on both tumor and drug characteristics [1]. Researchers can find molecular genetic variables that influence drug sensitivity by training these algorithms on vast datasets of genomic and molecular data along with accompanying drug response data. The objective is to create exact models that can accurately predict medication sensitivity, allowing doctors to select the best course of treatment for their patients.

However, it could be challenging to create precise drug sensitivity prediction models using computational methods [2, 3, 4, 5]. The calibre of the bioactivity data utilized to train these algorithms is one major obstacle. Although getting high-quality data can be expensive, it is essential to the predictive models' accuracy. In addition, it is challenging to create accurate models due to the complexity of the molecular interactions between medications and diseases.

Therefore drug sensitivity prediction is a critical aspect of precision medicine, as it helps identify the most effective medications for a particular patient's disease.

Deep learning algorithms have been an effective resource for predicting medication sensitivity in recent years. These models employ innovative machine-learning techniques that can find complex relationships between molecular characteristics and medication sensitivity [5]. The Cancer Cell Line Encyclopedia (CCLE) and the Genomics of Drug Sensitivity in Cancer (GDSC) (IC50) are two of the largest datasets accessible for this use. These datasets offer researchers an abundance of genetic, drug, and molecular response information, allowing them to produce more reliable drug sensitivity prediction models. As endorsed by the study [6], such improved screening datasets have assisted the development of cancer therapeutics and biology research.

For a given patient's disease, precision medicine includes drug sensitivity predictions that assist in the discovery of the best medicines. This approach is beneficial and different compared to the common trial-and-error methods of drug discovery and development. In addition, trial-and-error methods may be costly, unsuccessful, and often time-consuming. According to [6], the high-throughput screening methods have allowed for the collection of a large amount of data on the sensitivity of a panel of cancer cell lines and hundreds of substances. The high-throughput screening methods enabled the collection of a large amount of data on the sensitivity of a panel of cancer cell lines and hundreds of substances [6]. In addition the dataset provide a number of molecular and genomic data including DNA sequencing, protein interaction networks, and gene expression.

Furthermore, to achieve a better and deeper understanding of our project's inner workings, we utilized the Explainable AI (XAI) approach as the purpose was to improve the interpretability and transparency of our ML model. This technique help us to better understand the elements that influence its predictions. Moreover, using XAI enabled us to shed light on the model's previously unknown skills and acquire significant insights into its decision-making processes. This technique can predict drug sensitivity on cancer cell lines by assessing data on the cancer cells' genetic and molecular properties, as well as the drug treatments used to target them. The XAI approach may help to recog-

nize the biomarkers or features that are most predictive of drug sensitivity. Moreover, to provide insights into the underlying mechanisms that drive drug response, the XAI approach can be helpful.

Based on the cancer patient's molecular and genetic profiles, this work employed an XAI approach. This technique can predict which cancer patients are most likely to react to a given therapy. In addition, the model can then outline the precise genes or biochemical pathways that are regulating medication response. This can give a clear, understandable explanation of how it made its predictions.

Overall, enhancing the efficacy of treatments and patient outcomes depends on the development of precise drug sensitivity prediction models. Researchers can continue to grow in this critical area of precision medicine by utilizing the most recent datasets and innovative machine-learning methodologies.

Despite significant advancements in the field of drug sensitivity prediction, there are still challenges in creating precise computational models. Obtaining high-quality bioactivity data, which is crucial for training accurate models, can be expensive and time-consuming. Additionally, the complex molecular interactions between medications and diseases make it difficult to develop accurate prediction models.

In our pursuit of predicting drug sensitivity, we employed Explainable AI (XAI) techniques to enhance the transparency and interpretability of our machine-learning model. By leveraging XAI, we gained insights into the factors contributing to our model's predictions and elucidated the previously unexplained capabilities of the model. This approach enabled us to identify crucial traits or biomarkers indicative of drug sensitivity and shed light on the underlying mechanisms governing medication response. Our study aims to further advance drug sensitivity prediction by utilizing cutting-edge datasets and innovative machine-learning methodologies. By developing precise models and leveraging XAI techniques, we strive to improve treatment efficacy and patient outcomes in the field of cancer research.To provide a more accurate description, the uniqueness of our proposed model BioMarkerX can be outlined through the following aspects:

- The primary aim of this study is to incorporate techniques from explainable artificial intelligence (XAI) into the proposed BioMarkerX model.
- By integrating XAI component, the model not only attains precise predictions of drug responses but also offers valuable insights into the biological factors that contribute to the development of cancer. research on interpretable drug response prediction, offering potential opportunities for targeted drug development and personalized medicine.
- The integration of XAI techniques enhances the interpretability and transparency of the BioMarkerX model, allowing for a deeper understanding of the underlying mechanisms involved in drug response prediction.
- The findings of this research provide a foundation for further advancements in the field of explainable drug response prediction and have implications for improving patient outcomes through personalized treatment approaches.

The subsequent sections of this research report are structured as follows: Section 2 provides a synopsis of the relevant research and reference additivity models. Section 3 discusses the datasets utilized in this context and examines different machine-learning approaches for predicting drug sensitivity. Performance metrics employed in this study are discussed in Section 4. The experimental setup for our model is outlined in Section 5. Findings, including comparisons between the datasets, are presented in Section 6. Finally, Section 7 presents concluding remarks based on the research outcomes.

2 | RELATED WORK

Considerable literature exists on the topic of predicting drug sensitivity using data obtained from cancer cell lines. This area of research attracts significant attention due to the potential benefits of identifying optimal treatment strategies for cancer patients. Recently much research has been done on cancer classification like [7] and cancer diagnostics like [8]. In this section, we will review the existing literature and summarize the current state-of-the-art techniques and methodologies for predicting drug sensitivity in cancer cell lines. In medical contexts, it is essential to determine a patient's medication sensitivity to ensure that effective treatments are prescribed. Thus, simply knowing the response values of a drug may not be adequate. Categorizing cell line-drug couples into sensitive or resistant pairs is a more critical and practical task than estimating their response values.

Cancer precision medicine relies on the presence of somatically acquired tumor alterations, which act as indicators of how well medications and other treatments work. Individual genetic changes, such as driver mutations impacting oncogenes or tumor suppressor genes, or copy-number variations therein, are frequently these indicators [9].

Deep learning models emerge as powerful tools for predicting drug sensitivity in cancer cell lines by analyzing large-scale genomic data. These models extract complex features and learn non-linear relationships between input variables and output variables. Neural networks, in particular, are widely used for this task, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Deep Neural Networks (DNNs). CNNs show promise in analyzing images of cancer cells to predict drug sensitivity, while RNNs analyze temporal data, such as gene expression profiles, to predict drug response. DNNs are used for various tasks, including feature selection, drug combination prediction, and drug resistance prediction. The utilization of deep learning models has the potential to revolutionize precision medicine by enabling personalized treatment options for cancer patients based on their individual genomic and drug sensitivity profiles. [10]

Only a few previous studies use omics and chemical compound features to predict "cell line-drug sensitivity" [11]. These studies implement methods, including autoencoders combined with a neural network [6], SMILES encoding and bidirectional Recurrent Neural Network (bRNN) [11], and different deep learning models.

The datasets of Genomics of Drug Sensitivity in Cancer (GDSC) and Cancer Cell Line Encyclopedia (CCLE) have an enormous amount of gene expression data which we reduced by using oncogenes [12] that are important genes which transform a cell into a tumor cell, so we used those genes to reduce the dimensions of our dataset. These essential genes condensed the gene expression data for each cancer cell line.

A classification approach to predict drug sensitivity in cancer cell lines is discussed using various molecular information. Their methodology involved four preprocessing steps: imputing missing data, transforming IC50 values to binary labels using a maximum concentration threshold, calculating similarities, and standardizing and normalizing the data. The authors employed a thresholding approach to convert a regression problem to a classification problem and evaluated their model using recall, precision, F1 score, and accuracy metrics. Their work addresses the need for accurate drug sensitivity prediction in cancer patients and provides insight into the potential use of classification techniques for this task. However, there may be limitations to their approach, such as the assumptions made in the thresholding step and the choice of similarity measure used.[13]

The response distributions of drugs can vary depending on the chemical and may exhibit low variance for pharmaceuticals that target specific genes and pathways, and high variation for drugs that target generic cellular processes. However, the prediction performance of biologically driven models and genome-wide models is not significantly different despite the latter employing fewer input characteristics. Feature selection is a process of selecting a subset of relevant features from a large set of input features to improve the accuracy and efficiency of the prediction model. In the context of drug sensitivity prediction, feature selection plays a crucial role in identifying the relevant molecular

and genomic features that influence the sensitivity of cancer cells to different drugs. [2]

One of the main challenges in drug sensitivity prediction is dealing with high-dimensional data, which can lead to overfitting and poor generalization performance. Feature selection helps to address this challenge by reducing the dimensionality of the data and selecting the most informative features that contribute to prediction accuracy.

Feature selection algorithms can be categorized into three types that can be used depending on the nature of the data and the specific task: filters, wrappers, and embedded techniques [14]. Filter methods, for example, evaluate the relevance of each feature independently of the prediction model, and select the most informative features based on some statistical measure. Wrapper methods, on the other hand, incorporate the prediction model as part of the feature selection process and select the features that lead to the best model performance. Embedded methods, which are integrated into the model training process, select the most relevant features during model training. In previous research, both data-driven and biologically motivated feature selection methods were utilized [2].

The identification of effective cancer therapies relies on predicting drug sensitivity in individual patients. However, high-dimensional datasets with thousands of biological features can lead to overfitting and unreliable predictions. A recent study investigated different feature selection methods for drug sensitivity prediction using a dataset of hundreds of cancer cell lines and over 18,000 biological features. The study compared biologically driven feature selection based on drug targets and signaling pathways with data-driven methods based on genome-wide gene expression. The results showed that, in general, the baseline genome-wide set of features or data-driven feature selection performed better than biologically driven features. However, some individual drugs had better predictive performance when using biologically driven features, and these models often included only a small number of features. The study highlights the importance of feature selection strategies in drug sensitivity prediction and provides insights into the trade-off between biological relevance and predictive accuracy. [2]

In some of the previous works, the gene expression data that is gathered from the feature selection algorithms will be further used to merge with the compound Morgan fingerprints to forecast drug sensitivity data using a deep feed-forward neural network [11, 6, 5]. This provides a compliant structure which generalized it to be used for new drugs and cancer cell lines as well [5].

Avoiding overfitting the data is also one of the goals of feature subset selection. Another reason for selecting the features is not just to reduce the dimensions of data but also to remove biases. Biases are frequently present in data sets from various investigations. There are two basic forms of biases associated with experimentally measured data, aside from variations in research design (such as selection of cell lines and medications, and sample size), namely bias in molecular characterization and bias in response assay. These are equivalent to biases in input feature and output label selection when viewed in the context of a machine learning task. They represent the biggest obstacle to collaborative learning and learning transfer attempts across various data sources. Feature preprocessing can help to reduce the first kind of bias.[1]

One of the studies had implemented a module named HiDRA (Hierarchical Network for Drug Response Prediction with Attention) which used the Morgan Fingerprints and gene expression data which were aggregated at the pathway level using an attention mechanism. The main benefit was that the model was interpretable and capable of finding the essential features that were useful to predict drug sensitivity by using a hierarchical attention module.[15]

It is to note that in one of the works, 1856 essential genes are used, after which Pearson's correlation is calculated for each cell-line pair using "expression fold-changes of these essential genes"[16]. Similarly, another work computed both the Pearson and Spearman correlation coefficients between the observed and predicted IC50 values for the samples in the testing dataset. To find out the top drugs they counted the number of times a specific gene was selected in the predictive genes [12]. These works help with finding out about the essential genes which contribute the most in predicting drug sensitivity.

The prediction of drug responses is a critical challenge in the field of pharmacology. Previous studies have attempted to address this issue by utilizing various machine learning techniques, including regularized linear regression, k-nearest neighbors (KNN), Random Forests, and support vector machines (SVM). However, although these methods are effective at generating accurate predictions, their interpretability is often limited [2]. To overcome this limitation, some researchers have proposed multitask learning as a means of improving drug response prediction by leveraging knowledge gained from learning about multiple drugs. Despite the efficacy of these techniques, there is a pressing need for more interpretable models.

In other studies, researchers have investigated essential genes that play a crucial role in drug sensitivity prediction. Suphavilai et al. utilized 1856 essential genes and calculated Pearson's correlation coefficients for each cell-line pair using the expression fold-changes of these genes [16]. Similarly, Li et al. computed both Pearson and Spearman correlation coefficients between the observed and predicted IC50 values for samples in the testing dataset. They also identified the top drugs by counting the number of times a specific gene was selected as a predictive gene [12]. These works provide valuable insights into the genes that contribute the most to drug sensitivity prediction.

Gillani et al. proposed a model using XAI mentioning that when making predictions using Machine Learning (ML) models with high local accuracy, Explainable Artificial Intelligence (XAI) has been recognised as a useful working method for determining the applicability of key features. Therefore, XAI results may result in precise clinical predictions.[17]

Taken together, these studies highlight the importance of developing interpretable models for drug response prediction and identifying essential genes that contribute to drug sensitivity. Such insights can potentially inform the design of more effective therapies, ultimately benefiting patients and the healthcare industry.

Although existing approaches have shown promising results in predicting drug sensitivity, they often lack the ability to provide interpretable explanations for the obtained outcomes. /

In this research work, we have handled interpretability issues by incorporating explainable artificial intelligence (XAI) techniques in this proposed work. This will not only give remarkable results but also provide thorough explanations of biological understandings and decision-making.

Many previous studies have already made significant contributions to drug sensitivity prediction models. But still, there is a need for improvement in model interpretation and explainibility, so as to discover biological markers associated with cancer development. Our proposed **BioMarkerX** model aims to overcome this limitation by integrating XAI techniques.

By integrating XAI, we are able to accurately predict drug responses as well as gain important insight into the underlying molecular mechanisms that underlie the onset of cancer. This study makes a substantial contribution to the growing field of interpretable drug response prediction and offers exciting prospects for the development of personalized drug development medication.

3 | MATERIALS AND METHODS

3.1 | Data

Data on drug sensitivity and cell line expression patterns are retrieved from the two open databases CCLE and GDSC. We downloaded the compound chemical structure files from PubChem.

The CCLE and GDSC datasets were carefully examined and cleaned to remove redundant and unneeded data. Next, a dataset containing oncogenes was discovered, which are also referred to as Eukaryotic genes and have the power to predominate among cancer's oncogenic components. It is made up of proto-oncogenes, mutant forms of

normal cellular genes that have been preserved throughout development and are essential for healthy cells' physiological functions. To further improve the accuracy and specificity of the datasets, the oncogenes dataset was then compared with the CCLE and GDSC datasets, respectively, to identify the cell lines that coincided with the oncogenes dataset.

Cancer Cell Line Encyclopedia (CCLE). The dataset we downloaded had 1406 cancer cell lines and 19222 genes. These were downloaded from the CCLE website. We further selected cell lines with respect to the data we had against 238 drugs for this paper. IC50 values are used to indicate drug responses and are measured in μM units. A low IC50 value suggests that the specific drug is effective on the given cell line, while a high IC50 value indicates that the drug is not effective. To compare with the previous study, we utilized the negative logarithm of IC50 values (measured in μM) which is represented as negative log10 IC50. We downloaded the 1D and 2D compound structures of 238 drugs from the PubChem library named *PubChemPy* using Python. The Morgan fingerprints of the compounds, which were also acquired from a Python library named *rdkit*, were merged with the data using a 256-bit size, resulting in feature vectors of length 256 for each compound. The final dataset has 143802 rows and 1739 columns.

Genomics of Drug Sensitivity in Cancer (GDSC). The gene expression data was of around 812 cell lines and 19440 genes which were downloaded from the DepMap Portal. The same way was used to reduce the dataset. We first extracted the oncogenes from it to reduce the gene expression data as mentioned and then kept only the cell lines for which we had the data against the drugs. The structure files for the compounds, in both 1D and 2D format, are also obtained from PubChem and processed using the same method as those used for CCLE. The final dataset has 125946 rows and 1698 columns.

3.2 | Decision Tree Regressor

Regression difficulties are resolved using a machine-learning approach called a decision tree regressor. It divides the input space recursively into sets according to the values of the input characteristics. The feature that offers the greatest information gain at each level of the tree is chosen as the splitting criterion.

Creating a model that can forecast the target variable (i.e., the output) based on the input features is the aim of the decision tree regressor. To do this, a tree structure is constructed, with the internal nodes standing in for the feature tests and the leaf nodes for the anticipated values. The algorithm moves through the decision tree from root to leaf node, choosing the feature tests that most closely match the input features along the way. The leaf node is then used to get the output value. Both categorical and numerical input features, as well as missing values, can be handled by the decision tree regressor. It is prone to overfitting, though this can be avoided by using strategies like pruning and establishing a tree's maximum depth.

All things considered, the decision tree regressor is a potent and understandable machine learning algorithm that can be used to a number of regression tasks, such as the forecasting of housing values, stock prices, and customer churn rates.

3.3 | Explainable Artificial Intelligence

Explainable AI (XAI) is one of the state-of-the-art techniques that interprets the AI model decision-making which is more transparent and understandable to humans.

XAI can forecast drug sensitivity on cancer cell lines by analyzing data on the genetic and molecular characteristics of the cancer cells as well as the pharmacological therapies utilized to target them. XAI can help identify the traits or biomarkers that are most suggestive of drug sensitivity and offer insights into the underlying mechanisms governing

medication response.

Researchers can create more efficient medicines and treatments by using XAI to comprehend the underlying processes of drug sensitivity in cancer cells. These therapies and treatments will be specifically adapted to the unique genetic and molecular characteristics of each patient's cancer. By avoiding ineffective treatments, this strategy has the potential to enhance patient outcomes and lower healthcare expenditures.

3.4 | Particle Swarm Optimization

In PSO, a population of potential solutions (called particles) is assessed using a fitness function, and the particles then modify their locations in the search space in order to converge on the best option. The "personal best" answer is the one that each particle holds onto, while the "global best" solution is the best one that any particle in the population has come up with. Every time a particle discovers a better solution, the global best solution, which directs all particles towards the optimum, is updated. The neural network's weights can be updated using the global best solution in deep learning to enhance performance.

PSO, one of the bio-inspired algorithms, is uncomplicated in its search for the optimal solution in the problem space. It differs from other optimization techniques in that it only needs the objective function and does not rely on the gradient or any differential form of the objective. Moreover, there aren't many hyperparameters. Another benefit is how simple it is to parallelize PSO.

To update particles' velocities, we use equations 1 and 2:

$$V_i{}^{t+1} = W.V_i{}^t + c_1 U_1{}^t (P_{b_1}{}^t - P_i{}^t) + c_2 U_2{}^t (g_b{}^t - P_i{}^t) \tag{1}$$

To move the particles to their new positions, the following equation was used:

$$P_i{}^{t+1} = P_i{}^t + v_i{}^{t+1} \tag{2}$$

3.5 | Deep Learning and Neural Network

Artificial neural networks (NNs), which, as the name implies, are inspired by their biological counterparts, are the foundation of the deep learning class of machine learning (ML) techniques. Artificial NNs are made up of a number of connected layers, each of which contains a number of units, also known as neurons [18]. In contrast to deep neural networks (DNN), which often include numerous processing layers, shallow neural networks (NN) are made up of an input layer, one hidden layer, and one output layer [19]. These models can learn complicated nonlinear functions thanks to this feature. Also, unlike most classic ML techniques, DL approaches can learn higher-order representations straight from the raw input data, therefore they often do not need substantial feature selection before training [20].

Each unit takes its inputs from the layer above it in the hidden or output layers. A weight is assigned to each of the edges, or connections between nodes in adjacent layers, which reflect the relative relevance of a specific input. To determine its output value, each unit weights the total of its inputs and applies an activation function [19]. Up until the last layer's projected output values, this information is continuously sent forward.

The objective of a training NN is to reduce the error defined by a suitable loss function, which is accomplished by minimising the difference between the predicted output values and the true values. The projected and actual output values are compared once an input sample has been forwarded and propagated all the way to the output layer, and the error is calculated using the defined loss function. The backpropagation algorithm can be used to train the network

by calculating the gradient of the loss function and propagating the error from the output layer to the input layer in the reverse direction.[19]. The gradients with respect to the weights can be computed in this way. Gradient methods, such as stochastic gradient descent (SGD) or its variations, such as the Adam algorithm, can then be used to modify the weights [21]. As a result, learning is accomplished by repeatedly changing the weights.

4 | PERFORMANCE METRICS

Given that the drug sensitivity model is constantly urged to try to prevent major errors, the RMSE is more sensitive to unusually large errors and is therefore more appropriate for measuring drug sensitivity. Therefore determining root mean squared as our deep learning model's error metric.We also use the R^2 coefficient of determination to evaluate how well our deep learning models predict the future

$$RMSE = \sqrt{\sum (y_i - \tilde{y}_i)^2 / N} \tag{3}$$

$$R^2 = 1 - \sum \left(y_i - \tilde{y_i}^{r0}\right)^2 / \sum (y_i - \overline{y_i})^2 \tag{4}$$

where N is the test data size and where y_i is the target drug sensitivity data, $\widetilde{y_i}$ is the predicted counterpart of the i-th input data, and $\bar{y}$ is the average value of the target drug data.

Equations 3 and 4 shows the equations of RMSE and R^2.

5 | EXPERIMENTAL SETUP OF BIOMARKERX

In this study, we have proposed a deep learning model, BioMarkerX as given in Figure 1, integrated with Explainable AI, that predicts drug sensitivity in cancer cell lines. The hyper-parameters of the deep learning model were selected through a rigorous process that involved grid search and empirical testing, guided by the principles of achieving optimal predictive performance while maintaining computational feasibility.

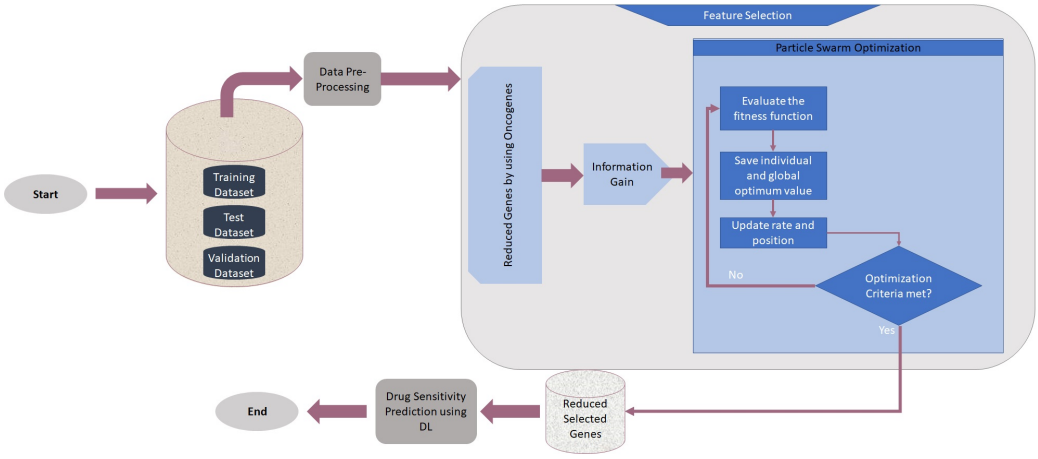
The elu activation function was specifically chosen for its advantages over other activation functions, such as tanh. The elu function helps in reducing the vanishing gradient problem, allowing the model to learn more efficiently, and thus, was considered more suitable for our deep learning architecture.

The data for the cell lines' expression and Morgan compound fingerprints were merged to serve as the input for BioMarkerX. Initially, we utilized an autoencoder for dimensionality reduction, but due to the opaque nature of the reduced feature space, which is not conducive to explainable AI, we shifted towards feature selection techniques. This shift was crucial in ensuring that the output of the model remained interpretable, a key factor in explainable AI which aims to make the decision-making process of machine learning models transparent.

The Particle Swarm Optimization (PSO) method was employed for feature selection, enhancing the precision of the input data by identifying the most relevant features. This selection process is crucial as it directly impacts the model's ability to generalize beyond the training data.

In the architecture of BioMarkerX, the deep feed-forward neural network plays a pivotal role, with an input layer designed to handle a combination of 256 compound signatures and 500 cell line features. The subsequent layers, each

FIGURE 1 **The BioMarkerX model diagram is presented as follows: (a) Data is acquired from the CCLE and GDSC datasets, followed by the selection of oncogenes' expression data and data normalization.The PSO technique is then applied for feature selection. (b) 2D drug structures are extracted from PubChem datasets and converted into Morgan fingerprints, serving as drug features. These features are merged with gene expression data to generate the final response data. (c) Both genomic and drug features are concatenated as input for training the neural network model. Model performance is evaluated using RMSE and R2 metrics.**



utilizing the elu activation function, are structured with decreasing numbers of neurons: 1,000, 800, 500, and 100, respectively. The output layer is distinct in that it does not employ an activation function, allowing for a continuous range of output values, and is composed of a single neuron reflecting the predicted drug sensitivity. The network employs the RMSE loss function and is initialized with the He normal initializer to optimize the training process.

Algorithm 1 provides a detailed pseudocode of BioMarkerX, illustrating the sequential steps of the model from input preparation to final prediction, embodying our methodological framework. To validate the model, a hidden test set is employed alongside a 10-fold cross-validation strategy to ensure robustness and unbiased performance evaluation.

Lastly, SHAP (SHapley Additive exPlanations) values are computed to dissect the contribution of each input feature to the output, enabling us to unravel the complexities of the model's predictions. The SHAP values give us a granular view of feature influence, providing a clearer explanation of the model's behavior and ensuring that our predictions are not only accurate but also understandable.

6 | RESULTS AND DISCUSSION

Based on the results presented in Table 1 and Table 2, it can be concluded that our BioMarkerX model outperformed the other models in terms of predictive performance for the GDSC dataset, achieving the lowest RMSE and highest R^2 scores using both 10-fold cross-validation and leave-tissue-out validation. The Genetic Algorithm + Neural Network (GA + NN) model also showed competitive performance, with RMSE values of 0.35 ± 0.05 and 0.45 ± 0.03, and R^2 scores of 0.81 ± 0.03 and 0.73 ± 0.03 for 10-fold cross-validation and leave-tissue-out validation, respectively.

The DeepDSC model had a higher RMSE and lower R^2 compared to the other models, with RMSE values of 0.23 ± 0.02 and 0.28 ± 0.08, and R^2 scores of 0.78 ± 0.04 and 0.73 ± 0.12 for 10-fold cross-validation and leave-tissue-out

Algorithm 1
Pseudocode of BioMarkerX

Input:
Divide the set of features into a Training set, a Validation set, and a Test set.
Output:
Step#1: Top-ranked features from Information Gain (IG)
Step#2: Optimal feature subset from Particle Swarm Optimization (PSO)
Model: Prediction model for Drug Sensitivity

Begin: // General steps for Step One Feature Selection of the proposed model
$S \leftarrow 0$;
$valueInfoGain \leftarrow$ Calculate the information gain for all the n features;
Sort the results of features from step 2;
if $valueInfoGain > S$ **then**
    Select the features;
end if
Step#1 $\leftarrow$ Subset of the selected attributes from step 6;
$x_i \leftarrow$ Initialized randomly;
$v_i \leftarrow$ Initialized randomly;
$iterations \leftarrow$ Initialized randomly;
$p$best $\leftarrow$ Initialized randomly;
$g$best $\leftarrow$ Initialized randomly;
$P \leftarrow$ Generate random particles;
for each particle i **do**
    Calculate fitness function;
    Update $p$best, $g$best;
end for
while $iterations$ **do**
    for each particle i **do**
        Update v_i, x_i;
        if $x_i >$ limit **then**
            $x_i \leftarrow$ limit;
        end if
        Calculate fitness function;
        Update $p$best, $g$best;
    end for
end while
End Feature Selection

Begin: // Prediction of Drug Sensitivity
Use the selected features as input to our neural network merged with the Morgan Fingerprints and drug response;
Divide our data into training, testing, and validation sets;
Train the neural network over the training data;
Test the neural network over the validation and testing data;
Evaluate the performance of the prediction results;
End Prediction

TABLE 1 Results on GDSC Dataset

	Validation Scheme	RMSE	R^2
GA + NN	10-fold CV	0.36 ± 0.05	0.85 ± 0.03
GA + NN	leave-tissue-out	0.45 ± 0.03	0.72 ± 0.03
DeepDSC [6]	10-fold CV	0.52 ± 0.01	0.78 ± 0.01
DeepDSC [6]	leave-tissue-out	0.64 ± 0.05	0.66 ± 0.07
NeuPD [4]	10-fold CV	0.490 ± 0.02	0.929 ± 0.03
BioMarkerX	10-fold CV	0.35 ± 0.05	0.89 ± 0.03
BioMarkerX	leave-tissue-out	0.40 ± 0.02	0.88 ± 0.03

TABLE 2 Results on CCLE Dataset

	Validation Scheme	RMSE	R^2
GA + NN	10-fold CV	0.35 ± 0.05	0.81 ± 0.03
GA + NN	leave-tissue-out	0.45 ± 0.03	0.73 ± 0.03
DeepDSC [6]	10-fold CV	0.23 ± 0.02	0.78 ± 0.04
DeepDSC [6]	leave-tissue-out	0.28 ± 0.08	0.73 ± 0.12
NeuPD [4]	10-fold CV	1.784 ± 0.02	0.543 ± 0.03
BioMarkerX	10-fold CV	0.32 ± 0.05	0.86 ± 0.03
BioMarkerX	leave-tissue-out	0.38 ± 0.03	0.87 ± 0.03

validation, respectively.

For the NeuPD model, the performance was notably lower, with RMSE values of 1.784 ± 0.02 for 10-fold cross-validation, indicating a less accurate prediction capability. The R^2 score for NeuPD was 0.543 ± 0.03, also lower compared to other models.

For the CCLE dataset, the DeepDSC model achieved the lowest RMSE using 10-fold cross-validation, while our BioMarkerX model had the highest R^2 using both 10-fold cross-validation and leave-tissue-out validation. The GA + NN model also showed competitive performance on this dataset.

Overall, our BioMarkerX model appears to be a promising approach for predicting cancer drug sensitivity based on gene expression data. Figures 2 and 3 show the graphical comparison of RMSE and R^2 values on these models.

6.1 | Comparison on GDSC Data

Table 1 shows the comparative performance of three models, GA + NN, BioMarkerX, and DeepDSC, in predicting drug response on the GDSC dataset using two different validation schemes: 10-fold cross-validation and leave-tissue-out.

6.2 | Comparison with CCLE Data

Analyzing the performance metrics, BioMarkerX emerged as the top-performing model in our study. In the 10-fold cross-validation (CV) approach, BioMarkerX achieved an impressive Root Mean Square Error (RMSE) of 0.35 ± 0.05

FIGURE 2 Visual Comparison on GDSC Dataset

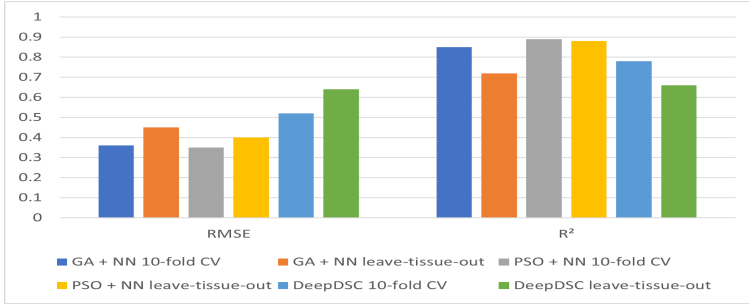
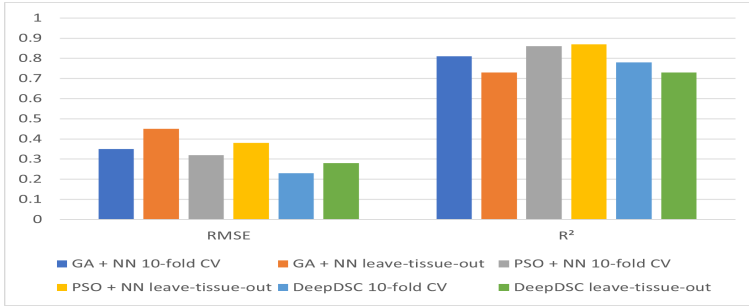


FIGURE 3 Visual Comparison on CCLE Dataset



and an R^2 score of 0.89 ± 0.03, and it maintained its strong performance in the leave-tissue-out validation scheme with an RMSE of 0.40 ± 0.02 and an R^2 score of 0.88 ± 0.03.

The Genetic Algorithm + Neural Network (GA + NN) model also showed commendable results, particularly in the 10-fold CV with an RMSE of 0.36 ± 0.05 and an R^2 score of 0.85 ± 0.03. However, its performance slightly declined in the leave-tissue-out scheme, recording an RMSE of 0.45 ± 0.03 and an R^2 score of 0.72 ± 0.03.

In contrast, the DeepDSC model displayed moderate performance in the 10-fold CV, with an RMSE of 0.52 ± 0.01 and an R^2 score of 0.78 ± 0.01. Its performance was notably weaker in the leave-tissue-out scheme, yielding the highest RMSE of 0.64 ± 0.05 and the lowest R^2 score of 0.66 ± 0.07 among the evaluated models.

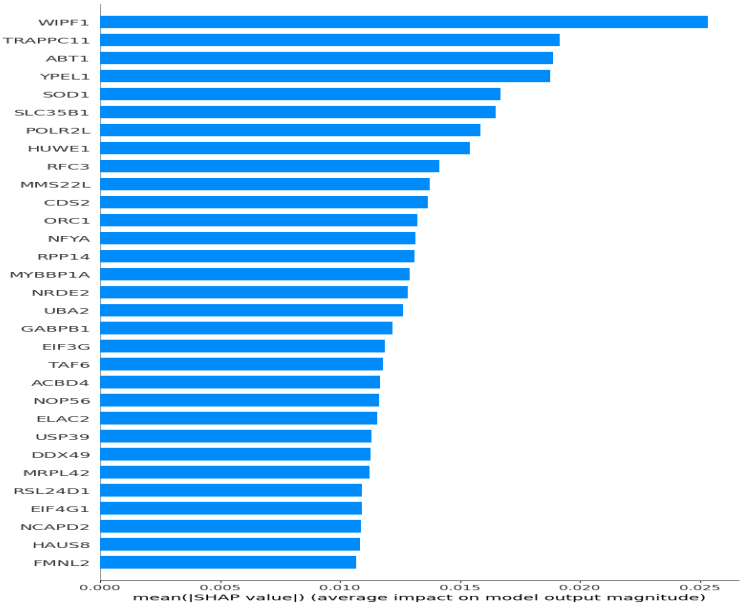
The NeuPD model, although not outperforming BioMarkerX, showed a strong R^2 score of 0.929 ± 0.03 in the 10-fold CV, despite having an RMSE of 0.490 ± 0.02, indicating its potential in certain aspects of genomic profile-based drug response prediction.

Overall, these findings underscore the effectiveness of the BioMarkerX model in predicting drug responses from genomic profiles, suggesting its significant potential in facilitating personalized medicine and drug discovery, particularly in the context of cancer treatment.

6.3 | XAI SHAP Results

Table 2 summarizes the results of comparing different models on the CCLE dataset using two validation schemes: 10-fold cross-validation and leave-tissue-out. Our proposed BioMarkerX model demonstrated strong performance

FIGURE 4 GDSC Dataset- Mean Shap values on the x-axis and genes on the y-axis



with an RMSE of 0.35 ± 0.03 and R^2 of 0.86 ± 0.03 under 10-fold cross-validation, and an RMSE of 0.38 ± 0.03 and R^2 of 0.87 ± 0.03 under leave-tissue-out. The GA + NN model also performed well, achieving an RMSE of 0.35 ± 0.03 and R^2 of 0.81 ± 0.03 under 10-fold cross-validation and an RMSE of 0.45 ± 0.03 and R^2 of 0.73 ± 0.03 under leave-tissue-out. The DeepDSC model obtained an RMSE of 0.23 ± 0.02 and R^2 of 0.78 ± 0.04 under 10-fold cross-validation, but its performance dropped significantly under leave-tissue-out, with an RMSE of 0.28 ± 0.08 and R^2 of 0.73 ± 0.12.

Under both validation approaches, BioMarkerX performed comparably to the GA + NN model and outperformed the DeepDSC model. Furthermore, BioMarkerX used more pharmacological data than DeepDSC, implying that it is more generalizable. As a result, we conclude that the BioMarkerX model is a good predictor of drug response on the CCLE dataset.

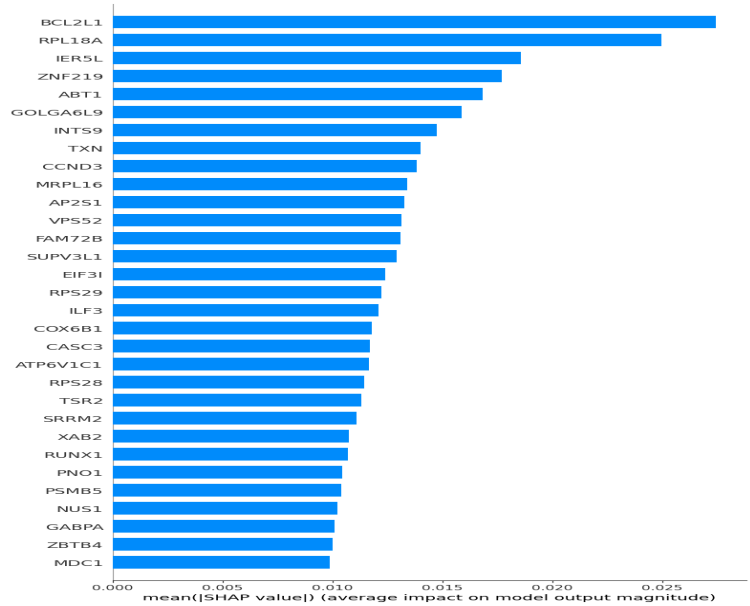
The two figures we've provided are SHAP (Shapley Additive Explanations) summary plots, which are used to interpret the output of machine learning models, particularly the BioMarkerX model in your case. These plots show the average impact of each feature on the model's output.

In the first image (Figure 4), the top feature by importance is BCL2L1, which has the highest mean absolute SHAP value, indicating that it has the greatest average impact on the model's predictions. The length of the bar represents the average magnitude of the feature's effect on the model's output, without considering the direction of the effect (whether it increases or decreases the predicted value).

Similarly, in the second image (Figure 5), the top feature by importance seems to be WIPI1, followed by TRAPPC11 and ABT1, each with their respective mean absolute SHAP values indicated by the length of their bars.

From these plots, we can infer that BCL2L1 is a significant gene in the context of the GDSC dataset, influencing the model's output to a greater extent than other genes. This aligns with biomedical research findings that overexpression of BCL-2 can allow cancer cells to evade apoptosis, contributing to the disease's progression. [22]

FIGURE 5 CCLE Dataset- Mean Shap values on the x-axis and genes on the y-axis



The SHAP summary plots are crucial for understanding the model's predictions, as they provide a visual representation of the contribution of each feature to the decision-making process. Features with longer bars are more influential in the model's predictions.

These plots are valuable tools for researchers and clinicians as they offer insights into the potential biological significance of different genes in cancer and help in identifying targets for further investigation or therapeutic intervention.

Figures 6 and 7 showcase the SHAP violin summary charts, an integral part of our study's analytical arsenal. These charts employ the visual language of violin plots to articulate the distribution and density of SHAP values across all observations for individual features within our datasets. Each violin plot encapsulates the spread and frequency of the SHAP values, with the width of the plot at different SHAP value levels indicating the density of points—broader sections mean more data points with that SHAP value.

The density curves within these violins allow us to intuitively grasp the distribution of the impact each feature has on the model's prediction for any given observation. For example, a feature with a wide violin at a high positive SHAP value indicates a strong positive impact on the model's output for many observations. Conversely, a feature with a narrow shape suggests less variability in its impact on the model's output.

What sets the SHAP violin summary chart apart from conventional violin plots is its capacity to convey a deeper narrative about the feature's behavior in the model—it not only reveals the distribution of the SHAP values but also their skewness or asymmetry. This skewness can signify the directional tendency of a feature's impact on the model's predictions.

Incorporating SHAP violin summary charts into our research provides a multifaceted view of how individual features drive the model's predictions. This is crucial for validating the interpretability of our BioMarkerX model and for ensuring that its decision-making process aligns with biological expectations and domain expertise. By scrutinizing

FIGURE 6 GDSC Dataset - SHAP violin summary plot

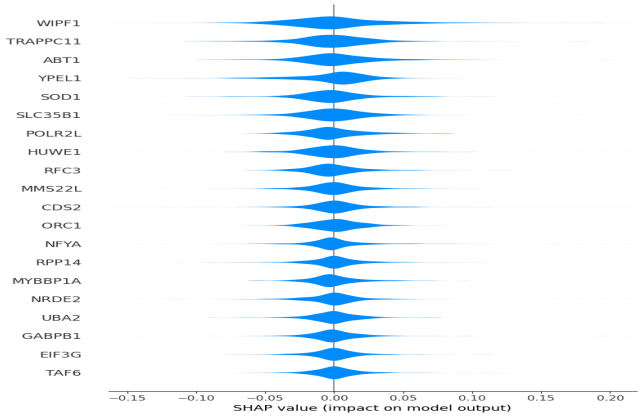
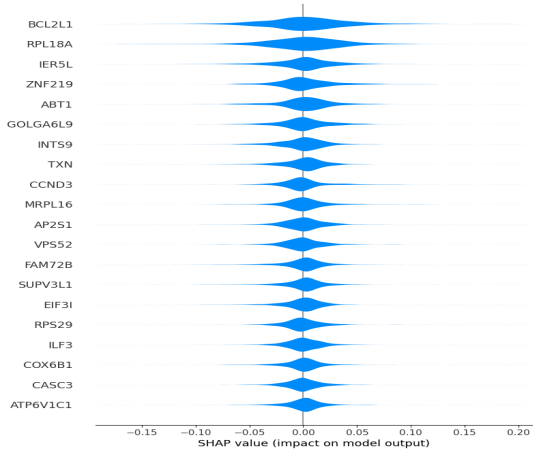


FIGURE 7 CCLE Dataset - SHAP violin summary plot



these plots, we can discern the nuances of feature contributions, validate model behavior against known biological pathways, and possibly uncover novel insights into the molecular underpinnings of cancer.

Ultimately, our research aims to demonstrate the efficacy of SHAP violin summary charts in dissecting and understanding complex machine learning models, particularly in the high-stakes context of biomedical research, where interpretability is as vital as predictive accuracy.

In our study, we have harnessed the power of SHAP Waterfall plots, as demonstrated in Figures 8 and 9, to dissect the decision-making process of the BioMarkerX model. These plots are instrumental in visualizing the step-by-step contribution of each feature to the final prediction outcome. Starting from the base value, which is the average model output over the dataset, each bar represents the contribution of an individual feature to shifting the prediction away from the base value towards the final output.

The features are ordered to show their contribution to the prediction, with the most significant influencers at the top of the plot. Positive contributions that increase the prediction value are represented by red bars, while negative

contributions that decrease the prediction value are shown in blue. This ordering provides a hierarchical view of feature importance, clarifying which factors are most influential in the model's predictions.

For example, in the plot for the GDSC dataset, we can observe that the feature MYC has a negative SHAP value, meaning it contributed to a decrease in the model's prediction value for this particular instance. Conversely, ABT1 has a positive SHAP value, indicating it contributed to an increase in the prediction value.

The Waterfall plots serve as an excellent tool for pinpointing key drivers of the model's predictions, enabling us to understand not just the 'what' but also the 'why' behind the model's output. By analyzing these plots, we can ascertain the individual and collective impact of the features, which is paramount in validating the model's accuracy and reliability. Furthermore, these visualizations are crucial for communicating complex model behaviors to stakeholders, ensuring transparency and fostering trust in the model's predictive capabilities.

In summary, the SHAP Waterfall plots are an invaluable asset in our research, simplifying the complexity of model interpretability and providing clear insights into the intricate dynamics of the BioMarkerX model's predictive functionality.

FIGURE 8 **Waterfall Model of sample data point of GDSC**

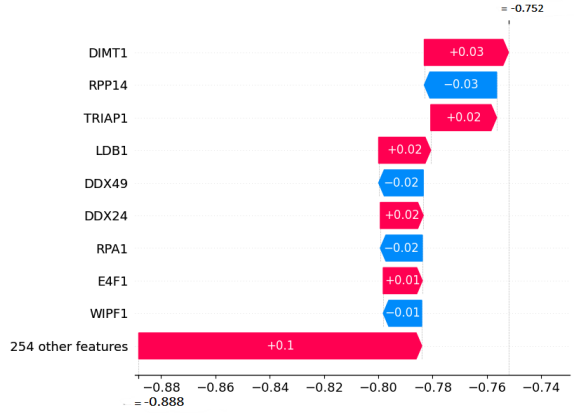
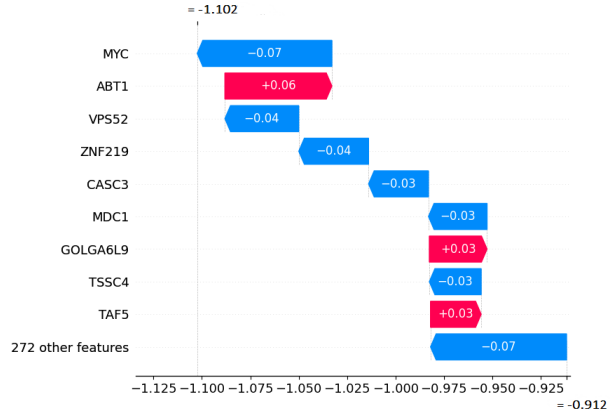


FIGURE 9 **Waterfall Model of sample data point of CCLE**



7 | CONCLUSION

Concluding our work, this study developed a BioMarkerX deep learning model to predict the cancer cell line drug sensitivity by using GDSC and CCLE drug datasets. Gene expression data and Morgan fingerprints were selected as the genomic aspects of cell lines and chemical characteristics of substances, respectively. The Particle Swarm Optimization (PSO) technique was utilized for feature selection. The 10-fold cross-validation results showed that BioMarkerX had high interpolation power to fill in missing drug sensitivity data, with an RMSE of 0.32 ± 0.05 and R^2 of 0.86 ± 0.03. Results from leave-tissue-out indicated that BioMarkerX likewise had very low extrapolation errors, with an RMSE of 0.38 ± 0.03 and R^2 of 0.87 ± 0.03. Overall, the BioMarkerX model can be considered an effective method for predicting drug response on the CCLE and GDSC datasets. Additionally, an explainable artificial intelligence (XAI) model was employed to identify biological indicators associated with cancer development, which can provide valuable insights for targeted drug development. Conclusively, we have tried to achieve two goals, i.e., to improve the performance of the existing drug sensitivity prediction problem and to identify the key genes that contribute to the drug sensitivity prediction model using the XAI component. This study provides a useful approach for identifying critical biological indicators relevant to cancer disease, which can help in the creation of more effective anticancer drugs.

By including XAI, the model, BioMarkerX, generates precise predictions of treatment response and provides insightful knowledge about the molecular factors underlying the development of cancer.

This study opens potential opportunities for targeted drug development and personalized treatment by adding to the body of knowledge on interpretable drug response prediction. The incorporation of XAI approaches improves the BioMarkerX model's interpretability and transparency and enables a deeper comprehension of the underlying mechanisms underpinning drug response prediction. The outcomes of this study have implications for improving patient outcomes through individualized treatment strategies and lay the groundwork for future developments in the field of explainable medication response prediction.

This study, however, has some limitations. It only works on omics data. It does not include consideration for drug pathways and similar data. This can be extended by integrating other omics data like mutation.

Author Contribution

Conceptualization, M. Shahzad and Rauf Ahmed S. M.; methodology, M. Shahzad; software, Ruhal; validation, Ismbah and Mahnoor; formal analysis, M. Shahzad.; investigation, Rauf Ahmed S. M. and M. Shahzad; resources, M. Shahid A.; data curation, M. Shahzad; writing—original draft preparation, M. Shahzad and Isbah I.; writing—review and editing, M. Shahzad, Isbah, Mahnoor M. and Ruhal; visualization, Piratdin A.; supervision, M. Shahzad and Rauf Ahmed S. M.; project administration, Khursheed A.; funding acquisition, M. Shahid A.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The datasets used in this research are available at https://www.cancerrxgene.org, https://cancer.sanger.ac.uk/cosmic, https://pubchem.ncbi.nlm.nih.gov, and https://drive.google.com/drive/folders/1RqPDQ5eKAEAG1i5hQpHFv2qBQ$_q$*nSKJw?usp sharing*.

Conflict of Interest

We have no conflicts of interest to disclose about the subject matter of this paper.

references

[1] Xia F. A cross-study analysis of drug response prediction in cancer cell lines. Briefings in Bioinformatics;23(1). https://doi.org/10.1093/bib/bbab356.

[2] Koras K, Juraeva D, Kreis J, Mazur J, Staub E, Szczurek E. Feature selection strategies for Drug Sensitivity Prediction. Scientific Reports;10(1).

[3] Ullah N, Khan MS, Khan JA, Choi A, Anwar MS. A robust end-to-end deep learning-based approach for effective and reliable BTD using MR images. Sensors 2022;22(19):7575.

[4] Shahzad M, Tahir MA, Alhussein M, Mobin A, Shams Malick RA, Anwar MS. NeuPD—A Neural Network-Based Approach to Predict Antineoplastic Drug Response. Diagnostics 2023;13(12):2043.

[5] Tang Y, Gottlieb A. Explainable drug sensitivity prediction through cancer pathway enrichment. Scientific Reports;11(1). https://doi.org/10.1038/s41598-021-82612-7.

[6] Li M. DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines. IEEE/ACM Transactions on Computational Biology and Bioinformatics;18(2):575–582. https://doi.org/10.1109/tcbb.2019.2919581.

[7] Yaqub M, Jinchao F, Ahmed S, Mehmood A, Chuhan IS, Manan MA, et al. DeepLabV3, IBCO-based ALCResNet: A fully automated classification, and grading system for brain tumor. Alexandria Engineering Journal 2023;76:609–627. https://www.sciencedirect.com/science/article/pii/S1110016823005379.

[8] Mehmood M, Abbasi SH, Aurangzeb K, Majeed MF, Anwar MS, Alhussein M. A classifier model for prostate cancer diagnosis using CNNs and transfer learning with multi-parametric MRI. Frontiers in Oncology 2023;13:1225490.

[9] Levatić J, Mutational signatures are markers of drug sensitivity of cancer cells;. https://doi.org/10.1101/2021.05.19.444811.

[10] Partin A, Brettin TS, Zhu Y, Narykov O, Clyde A, Overbeek J, et al. Deep learning methods for drug response prediction in cancer: Predominant and emerging trends. Frontiers in Medicine 2023;10. https://www.frontiersin.org/articles/10.3389/fmed.2023.1086097.

[11] Manica M, Oskooei A, Born J, Subramanian V, Saez-Rodriguez J, Rodriguez Martinez M. Toward Explainable Anticancer Compound Sensitivity Prediction via Multimodal Attention-Based Convolutional Encoders. Mol Pharmaceutics;16:4797– 4806. https://doi.org/10.1021/acs.molpharmaceut.9b00520.

[12] Li Y, Umbach DM, Krahn JM, Shats I, Li X, Li L. Predicting tumor response to drugs based on gene-expression biomarkers of sensitivity learned from cancer cell lines. BMC Genomics;22(1).

[13] Ahmadi Moughari F, Eslahchi C. A computational method for drug sensitivity prediction of cancer cell lines based on various molecular information. PLOS ONE;16(4). https://doi.org/10.1371/journal.pone.0250620.

[14] Hira ZM, Gillies DF. A review of feature selection and feature extraction methods applied on microarray data. Advances in Bioinformatics;2015:1–13.

[15] Jin I, Nam H. Hidra: Hierarchical Network for drug response prediction with attention. Journal of Chemical Information and Modeling;61(8):3858–3867.

[16] Suphavilai C, Bertrand D, Nagarajan N. Predicting cancer drug response using a recommender system. Bioinformatics;34(22):3907–3914.

[17] Gillani IS, Shahzad M, Mobin A, Munawar MR, Awan MU, Asif M. Explainable AI in Drug Sensitivity Prediction on Cancer Cell Lines. In: 2022 International Conference on Emerging Trends in Smart Technologies (ICETST), 2022;. p. 1–5. https://doi.org/10.1109/ICETST55735.2022.9922931.

[18] Baptista D, Ferreira PG, Rocha M. Deep learning for drug response prediction in cancer. Briefings in Bioinformatics 2020 01;22(1):360–379. https://doi.org/10.1093/bib/bbz171.

[19] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May;521(7553):436–444. https://doi.org/10.1038/nature14539.

[20] Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence 2013;35(8):1798–1828.

[21] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. CoRR 2014;abs/1412.6980.

[22] Letai AG. Diagnosing and exploiting cancer's addiction to blocks in apoptosis. Nature Reviews Cancer 2008;8(2):121–132.

Muhammad Shahzad is Assistant Professor in the Department of Computer Science, FAST - National University of Computer and Emerging Sciences. He received his Ph.D. degree in Computer Science from FAST-NUCES in 2023. MS degree in Computer Science from Hamdard University in 2007. He achieved his B.S. degree in computer science from Shah Abdul Latif University. Additionally, he actively participates in various biomedicine projects and has authored multiple papers in this domain. His research focuses on algorithms, machine learning, health informatics, bioinformatics, and computational biology. Notably, he was awarded a scholarship from HEC Pakistan, which enabled him to work as a visiting scholar at Lancaster University, UK.

Ruhal Lohana received a B.S. degree in Computer Science from FAST-NUCES. He has worked on several projects, mainly on computer vision, machine learning and deep learning. His research interests include computer vision, machine learning and bioinformatics.

Khursheed Aurangzeb (Senior Member IEEE) is Associate professor in the Department of Computer Engineering, College of Computer and Information Sciences at King Saud University (KSU), Riyadh, Saudi Arabia. He received his Ph.D. degree in Electronics Design from Mid Sweden University Sweden in June 2013, MS degree in Electrical Engineering (System on Chip Design) from Linkoping University, Sweden in 2009. He received his B.S. degree in Computer Engineering from COMSATS Institute of Information Technology Abbottabad, Pakistan in 2006. Dr. Khursheed has authored and co-authored more than 90 publications including IEEE / ACM / Springer / Hindawi / MDPI journals, and flagship conference papers. He has obtained more than 15 years of excellent experience as instructor and re-

searcher in data analytics, machine/deep learning, signal processing, electronics circuits/systems and embedded systems. He has been involved in many research projects as a principal investigator and a co-principal investigator. His research interest is in the diverse fields of embedded systems, computer architecture, signal processing, wireless sensor networks, communication, and camera-based sensor networks with an emphasis on big data and machine/deep learning with applications in smart grids, precision agriculture, and healthcare.

Isbah Imtiaz Ali completed her Bachelors in Computer Science from FAST-NUCES. She has worked on a handful of machine learning and deep learning projects including object classification and detection. Her research interests include machine learning, data science and bioinformatics.

Muhammad Shahid Anwar is currently working as an Assistant Professor in the Department of AI and Software at Gachon University, Seongnam, South Korea. He received his Ph.D. degree in Information and Communication Engineering from the School of Information and Electronics, Beijing Institute of Technology, Beijing, China in 2021 and his M.Sc. in Telecommunications Technology from Aston University, Birmingham, U.K., in 2012. Dr. Shahid has authored and co-authored more than 60 publications including IEEE, Springer, IET, Hindawi, MDPI, Frontiers journals, and flagship conference papers. He has been honored with the “Outstanding Scholar of the Year 2020 Award" from the CSC Scholarship Council under the Ministry of Education China. Dr. Shahid also received the “Excellent Student of the Year 2020 Award" from the Beijing Institute of Technology, China. He has been serving as an editorial board member in CSSE Tech Science and a reviewer of several Journals including ACM and IEEE Transactions. His research interests include 360-degree videos, immersive Media (Virtual Reality, AR), Metaverse, and Quality of Experience (QoE) evaluations of VR telemedicine and healthcare systems. He is focusing on deep learning-based VR video evaluations and developed several Machine Learning based QoE Prediction models.

Mahnoor Murtaza earned her Bachelor in Computer Science degree from FAST-NUCES. She has worked on various projects related to machine learning and deep learning. Her research interest includes bioinformatics and health informatics.

Rauf Ahmed Shams Malick received a PhD degree from the University of Karachi. He has been a Visiting Scholar with NİG, Japan, and UCLA, USA. He has founded several companies with state-of-the-art products related to social media, location-based analytics, and organizational networks. He is involved in complex system research and pursuing problems in the area of biological networks, networked economics, and personality traits. He has a distinguished background in designing novel solutions for complex systems. He is currently affiliated with the Department of Computer Science, National University of Computer and Emerging Sciences, as an Assistant Professor, continuing research in the specialized scientific area of computer science, complex networks, social computing, bioinformatics, and integrated systems. He has authored several articles along with chapters in different books.

Allayarov Piratdin is an associate professor of Mathematical methods in Economics Department at Tashkent State University of Economics in Uzbekistan. He graduated a Bachelor's degree in Finance from Karakalpak State University and a Master's degree in Accounting from the Karakalpak State University. He obtained Ph.D. in Economics at Hunan University in China. He participated as a local consultant in projects of international organizations such as the World Bank, UN Development Program, GIZ, UNICEF, Asian Development Bank. His primary research is in international trade and economics. He published more than 20 papers in local and international journals.