Enhanced Deep Learning Model for Personalized Cancer Treatment

HANAN AHMED1, SAFWAT HAMAD1, HOWIDA A. SHEDEED1, AND ASHRAF SAAD HUSSEIN2

1Faculty of Computer and Information Sciences, Ain Shams University, Cairo 11566, Egypt
2Faculty of Computer Science and Engineering, King Salman International University, El Tur 8701301, Egypt

Corresponding author: Hanan Ahmed (hanan.ahmed20100@cis.asu.edu.eg)

ABSTRACT

Personalized medicine provides more safe and effective treatment by individualizing the choice of drug and dose based on an individual’s genetic profile. Cancer patients’ response to anti-cancer treatments (drugs) is one of the foremost challenges in personalized medicine that releases the target treatment. Both size and availability of drug sensitivity data have motivated researchers to develop Artificial Intelligence (AI), based models, for predicting drug response to advance cancer treatment. The concerned AI models include Machine Learning (ML) and the recently advanced Deep Learning (DL) based models. This paper introduces both; a data federation method and a DL-based model for predicting drug response. The fundamental goal is to generalize the predictor so it will be able to predict the response to different drugs accurately. As the data has a considerable effect on any AI model, the data federation is utilized to consolidate the data. The proposed consolidation process is carried out to make each cell line contains gene expression data, its mutation profile, and drug response data. ML models such as Support Vector Machine (SVM) and Linear Regression (LR) are used along with Principal Component Analysis (PCA) for feature reduction, and the AI models have been tested with and without data federation. The results show that data federation enhanced the accuracy and decreased the Mean Square Error (MSE) by almost 25%. The proposed DL model uses dimension reduction encoders. The encoder is a DL model that uses unsupervised learning. It is trained by integrating an encoder with a decoder to achieve equality between the input and output. The proposed model has achieved the best accuracy compared to some other recent models in terms of the Pearson correlation coefficient (PCC) as a performance measure. In addition, the results show that the Enhanced Deep Drug Response prediction (Enhanced Deep-DR) model has achieved the best PCC value even with the largest number of genes and drugs, which proves the high capacity and efficiency of the proposed model. Convolutional Neural Network (CNN) based-model is also implemented; it achieves higher accuracy in predicting the drug response than in some other DL-based models but less than the Enhanced Deep learning. The Enhanced Deep-DR achieves better accuracy within the range of 5% to 12% than other DL-models.

INDEX TERMS

Artificial intelligence, artificial neural networks, biomedical, feedforward neural networks, personalized medicine, drug response prediction.

I. INTRODUCTION

Pharmacogenomics is one of the fundamental components of personalized medicine. It promises a safer and more effective drug treatment by individualizing the choice of drug and dose based on an individual’s genetic profile [1], [2]. To personalize medicine, genetic biomarkers and mutation profiles are used to study drug sensitivity [3], [4]. Drug sensitivity assays serve as a standard, high-throughput experimental platform for measuring the response of cancer cells to drug treatments. The standardized protocols of sensitivity assays, along with rapid improvement of technologies for genomic profiling, have led researchers to generate large pharmacogenomic drug response datasets for anti-cancer drug discovery [5], [6], [7].

Recently, soft computational models such as AI and artificial neural networks (ANN) are used for Genome-wide
investigations and analysis [8]. ML models have become a natural fit for analytically predicting the response of cell lines to drug treatments. Researchers strive to develop highly predictive ML drug response models [9], [10], [11]. Different ML models are used, such as recommendation systems [12], ranking methods [13], [14], generative models [15], [16], feature analysis [17], network modeling [18], and ensemble models [19] [20]. Given recent advances in ANNs, DL has become a favorite approach across a variety of scientific disciplines for discovering hidden patterns in large volumes of complex data, including the prediction of drug response in cancer cell lines [15], [21], [22], [23], [24].

Many DL models use different biomarker features for learning, such as gene expression profile, mutation profile, pathways, methylation, copy number, etc. Some other features are utilized, such as the chemical substructure of drugs [23], [25], to enhance the predictions.

For example, the graph convolutional network for drug response prediction (GraphDRP) [26] mainly uses features extracted from drugs and gene mutations. The GraphDRP models the drug as a molecular graph that reflects interactions between the atoms inside the drug. To construct this graph, a convolutional network model is used and learned from other datasets related to chemical compounds, such as PubChem. The only data used from cell lines is a binary vector of 735 dimensions which indicates genomic mutations. The cancer drug response doesn’t depend only on gene mutations, and the number of considered genes is relatively small. Also [27] proposes a twin Convolutional Neural Network (tCNN) that uses two CNNs: one for extracting features from the drug, and the other for extracting features from the cell line. The tCNN uses the simplified molecular input line entry specification (SMILES) format. SMILES represents drugs as a string which makes them lose their molecular representation. An ensemble transfer learning (ETL) framework [23] targets to predict drug response to both new cancer cases and drugs. The ETL pipeline is implemented using a gradient boosting model and two deep neural networks. The ETL performance is tested on Three drugs only. Multi-Omics Late Integration (MOLI) is a DL-based model which uses the mutation profile, copy number, and gene expression as input to predict drug response. The MOLI is validated on five chemotherapy agents and two targeted therapeutics.

Regardless of the learning approach, the target of models is to improve the generalization performance. The generalization has different targets; for some models, generalization refers to the aggregated accuracy of model predictions on a set of unseen data samples, whereas some other models generalize the prediction response to unseen drugs [23] [25], and others refer to building general models for predicting the drug response of patients with different types of cancers as in [28]. Drug response prediction still faces significant challenges, such as inconsistencies across studies in genes and response profiling, predictive models trained on one data source may not perform as well on another, and it is unclear how to improve model generalization with increasing the amounts of cell lines or drug data.

In this paper, we propose a data federation method to improve the data quality. The data federation improves the performance accuracy of the model. The performance accuracy of traditional machine learning models such as: SVM and LR is enhanced by MSE reduction of 83.59% and 81%, respectively. Also, we propose an Enhanced Deep-DR model based on the Deep-DR model [29]. It uses a deep neural network (DNN) model to predict the drug response to 265 different drugs. Both gene expression and mutation profile are used as input in the predictive process. The proposed model reduces the value of MSE by 25% compared to the MSE of Deep-DR. The proposed model uses Cancer Genome Atlas (TCGA) and Cancer Cell Line Encyclopedia (CCLE) for gene expressions and the data of mutation profile, and Genomics of Drug Sensitivity in Cancer (GDSC) for drug response. In addition, the proposed model utilizes data federation to enhance data quality.

The rest of the paper is adjusted as follows: Section 2 provides the Data pre-processing and section 3 displays a brief description of the Deep-DR. Next, section 4 introduces the Enhanced Deep-DR. Then, section 5 describes the ML models, and section 6 depicts the Convolutional Network. After that, section 7 presents the experimental results, and section 8 introduces a comparative study between the proposed model and other recent models. Finally, section 9 includes the conclusion and future work.

II. DATA PRE-PROCESSING
In this section of the paper, both, the used datasets and the data preprocessing, will be introduced.

A. DATASETS
In the proposed work, three different datasets are used; the Cancer Genome Atlas (TCGA) [30], Cancer Cell Line Encyclopedia (CCLE) [31] [32], and the Genomics of Drug Sensitivity in Cancer (GDSC) [5] [33]. CCLE project aims to characterize the genetic characteristics of cancer cell lines in an accurate manner. CCLE includes mutation status for 25 oncogenes across 486 cancer cell lines, DNA copy number variations for 23,316 genes across 1,043 cancer cell lines, and mRNA expressions for 54,675 mRNAs across 127 cancer cell lines [31]. In 2019, the CCLE database underwent a crucial update, including newly released DNA methylation data, whole genome sequencing data, and RNA-seq data [34]. The TCGA dataset consists of 2.5 petabytes of data, including the genomic profiles of tumors and matching normal tissues from more than 11,000 patients representing 33 types of human cancers [30]. The GDSC [5] project provides drug response gene expression from 1,001 cancer-cell lines and drug response data in the form of half maximal inhibitory concentration (IC50) values for cancer genomic data for 265 drugs.
B. DATA PREPARATION

The human DNA contains about 3 billion bases and about 20,000 genes on 23 pairs of chromosomes, and each DNA produces lots of data that requires storage. TCGA dataset consists of 2.5 petabytes of data. Due to the vast amount of gene expression data, such data needs to be summarized. So, the number of Transcripts Per Million (TPM) of the gene is used [35]. TPM is computed for both CCLE and TCGA given the total numbers of cell lines, tumors, and genes, like C, T, and G, respectively. In this way, we can compute the gene expression using (1) and (2).

\[ E^{\text{CCLE}} = \log_2(\text{tpm}^{\text{CCLE}}_g + 1), \]
\[ E^{\text{TCGA}} = \log_2(\text{tpm}^{\text{TCGA}}_g + 1), \]

where tpm\(^{\text{CCLE}}_g\) is the number of transcripts per million of gene \(g \in \{1, G\}\) in cell line \(c \in \{1, C\}\)).

\[ E^{\text{TCGA}} = \log_2(\text{tpm}^{\text{TCGA}}_c + 1), \]

where tpm\(^{\text{TCGA}}_c\) is the number of transcripts per million of gene \(g \in \{1, G\}\) in tumor \(t \in \{1, T\}\)).

For CCLE, gene expression TPM files for genes of 935 cell lines are downloaded from CTD\(^2\) Portal [35]. For TCGA, the Pan-Cancer dataset is used and TPM values are downloaded from UCSC Xena [36], [37] for 10536 TCGA samples. Genes over all samples that have a mean less than one are removed; as they are considered as low information burden. As the model is interested only in tumors, normal samples are removed so the samples become 9807 tumors. Normal samples are identified and excluded using the TCGA barcode [38].

Files of mutation data (Mutation Annotation Format (MAF) [39]) are downloaded for both TCGA PanCancer [40] and CCLE [41] directly from their websites. Only four mutation types are non-synonymous mutations (missense and nonsense mutations), frameshift insertions, and deletions in the mutation profile are used to construct the mutation matrix. The mutation matrices are binary matrices, including samples or cell lines as rows and all genes as columns, so:

\[ M^{\text{CCLE}} = [m^{\text{CCLE}}_{g,c}] \text{and} M^{\text{TCGA}} = [m^{\text{TCGA}}_{g,t}], \]

where \(m^{\text{CCLE}}_{g,c}\) and \(m^{\text{TCGA}}_{g,t}\) are the mutation states will be: (1 for mutation and 0 for wildtype) of gene \(g\) in \(c\) and \(t\), respectively.

Genes with no mutations in CCLE and TCGA samples are eliminated.

Drug response data for the CCLE dataset is downloaded from GDSC [5], GDSC contains responses of 990 cell lines only for 265 anti-cancer drugs measured by IC\(_{50}\). IC\(_{50}\) is measured in \(\mu\)M and is represented on the log scale. Only drug response data is used from GDSC, and missing data are imputed by a weighted mean, as in [29].

III. DEEP-DR

Deep-DR [29] is implemented to predict IC\(_{50}\) values. IC\(_{50}\) is a quantitative measure that indicates how much of a particular inhibitory substance (e.g. drug) is needed to inhibit, in vitro, a given biological process or biological component by 50%. The drug response data in GDSC is given in form of IC\(_{50}\), so the predicted results will be comparable to the real ones. Deep-DR predicts IC\(_{50}\) based on genomic profiles of a cell or a tumor. The implementation of the Deep-DR model is divided into two successive parts shown in Fig. 1:

1) The first part, which includes two auto-encoders, one for mutations and other one for gene expression, is implemented via two neural networks for dimensional reduction.

2) The second part is a prediction network which predicts the IC\(_{50}\) values of 265 drugs and uses the output of the auto-encoders as an input.

A. AUTO-ENCODERS

The Deep-DR is composed of two auto-encoders; one for gene expression and the other for mutations. Auto-encoder is an unsupervised DL (feedforward) that includes symmetric pair of the encoder-decoder as shown in Fig. 2. It is a recent dimensional reduction method depending on minimizing the loss between the input and reconstructed data (decoded) output. Auto-encoders are learned on TCGA mutations and gene expressions dataset to optimize the capability and capture high-order features. The input for the gene encoder is 18281 genes and 15363 genes for the mutation encoder. The inputs of two encoders are not equal; that means the gene expression involves some genes that do not have a mutation, or their mutation profiles are not included. The normal genes
should not be included. The used activation function for encoders is the Rectified Linear Unit (ReLU).

After the training process of auto-encoders, encoders are used in the Deep-DR model, and the decoders are neglected.

**B. PREDICTION NETWORK DEEP-DR**

The prediction network is a feedforward network to predict IC\textsubscript{50} anti-cancer drugs. Prediction network input is connected with the output of encoders trained in the previous sub-section as shown in Fig 1. The used activation function is ReLU except for the output layer which is a linear function. The Adam optimizer is utilized with MSE as a loss function. He’s uniform distribution [42] is used to initialize weights. The entire model is trained using CCLE data which is divided into training, validation, and testing sets. The validation dataset is used to avoid model overfitting. The data should be in the form of:

\[
(M_{\text{CCLE}}(:, c), E_{\text{CCLE}}(:, c), IC_{\text{CCLE}}(e))
\]

where \( M_{\text{CCLE}}(:, c), E_{\text{CCLE}}(:, c) \) is the input pair as the mutation and gene expression profiles for cell line \( c \). and \( IC_{\text{CCLE}}(e) \) is the output (drug response for the same cell line \( c \) from the GDSC dataset).

**IV. ENHANCED DEEP-DR**

Enhanced Deep-DR is a model based on Deep-DR explained in the previous section with some modifications. The modifications can be summarized as follow:

1) Data integration: integrates the mutations with their gene expression to have a sample with complete information.
2) Data federation: allows the multiple datasets (CCLE and GDSC) to act as one dataset with a new format.
3) Prediction Network: This predicts the IC\textsubscript{50} values of 265 drugs and uses the output of the auto-encoders as an input.

**A. DATA INTEGRATION**

To integrate mutation data and its gene expression, the gene annotation must be unified. The integration process is proposed to integrate the gene data to create a complete gene profile with both the gene expression and its mutations.

Gene expression files use *Ensembl ID* [43] to refer to the gene and mutation files use Hogo symbols as reference for genes. Hence, the same gene is identified with its synonyms.

*Ensembl IDs* are different, according to their assembly map version. In the assembly map, *Ensembl IDs* are mapped to their genes Hogo symbols. According to the assembly map version of each file, *Ensembl IDs* are converted to Hogo symbols. In order to remove non-mutated genes, Hogo symbols of genes are not included in the selected mutation file of the dataset removed from its gene expression file. Both mutated genes in CCLE and TCGA are combined to increase the model’s generalization capability. The total mutated genes without repetition are 19, 702 genes.

**B. DATA FEDERATION**

Data Federation [44] is a software method that allows multiple databases to act as one. The new database has its own format. The datasets that need to be federated are CCLE gene expression files from CTD\textsuperscript{2} Portal, MAF files from CCLE, and CCLE drug response from GDSC. Hence, after the federation, each cell line will contain gene expression data, its own mutation profile, and its drug response data. As CCLE gene expression TPM files, downloaded from CTD\textsuperscript{2} Portal, do not contain any ids. The file name, which is a unique sequence of letters and numbers, is used. The file name represents a sort of *ID* that is used in CCLE called *analysis id*. The mutation file contains the *DepMap ID* (new id system) which is the only identity. The drug response data from GDSC are identified using *COSMIC_ID*. Therefore, three different ids are used for each file. To federate these data, we should find a way to map those different id systems to each other.

The keys for mapping are:

1) “CCLE\_id\_mapping.txt” from CTD\textsuperscript{2} [35].
2) “sample\_info.csv” from GDSC [5].

The “CCLE\_id\_mapping.txt” is used to map the analysis id to *ccle\_name*. The “sample\_info.csv” is the key to map *DepMap ID* to the *ccle\_name*. As a result, gene expression files are converted from the *ccle\_name* to the corresponding *DepMap ID*; that means each gene profile has its mutation profile \([M_{\text{CCLE}}(:, c), E_{\text{CCLE}}(:, c), IC_{\text{CCLE}}(e)]\). But not all *DepMap IDs* have *COSMIC_ID* so the *DepMap IDs* that do not have *COSMIC_ID* are neglected. As a result, the CCLE dataset becomes 619 cell lines only.

**C. AUTO-ENCODERS**

The proposed model is composed of two auto-encoders one for gene expression and one for mutation. The auto-encoders are trained using the TCGA dataset (gene expression and mutation). Both auto-encoders have the same architecture. The encoder architecture consists of three layers each with 1024, 128, and 64 nodes respectively. The decoder is also 64, 128, and 1024 nodes respectively. The input layer for each encoder is 19, 702 genes and the output layer of each decoder is 19, 702. After the training process, the decoder is neglected (see Fig. 3). The genes considered in the gene expression are the same as those considered in the mutated one. Any gene that does not appear as a mutated gene in one of TCGA or CCLE is out of consideration. As cancer is a gene...
disease, any mutated gene is under consideration due to the existence of some mutated genes that cause the resistance of drugs or make them not effective. The increase in the number of mutated genes enhances the generalization and helps to discover more about the relation and hidden patterns between the drug response and the important genes. Both CCLE and TCGA contain tumor tissues so each mutated gene must be in consideration, while the existence of some genes mutated may cause the ineffectiveness of some drugs.

D. PREDICTION NETWORK OF ENHANCED DEEP-DR

The proposed enhancement is to increase the accuracy of the Deep-DR and make it more generalized by increasing the number of input genes. Hence, both the mutated gene lists of TCGA and CCLE are combined and used as input data for the gene expression encoder and mutation encoder. The input layer for both encoders is 19,702 genes. Therefore, the model can handle more mutations, find hidden patterns related to those genes and predict the drug response with better performance. Also, for prediction network training, we perform data federation explained in the previous sub-section which enhances the data quality and, as a result, the prediction accuracy.

The prediction network is a feedforward network to predict the IC50 values of different drugs. The network is four network layers with 128, 128, 128, and 265 nodes, respectively (see Fig 4). Moreover, the input layer is a merge between the 64 output from the gene expression encoder and the other 64 output of the mutation encoder. The used activation function is ReLU except for the output layer which is a linear function. The Adam optimizer is utilized with MSE as a loss function. He’s uniform distribution [42] is used to initialize weights.

In both prediction training and auto-encoders, a batch size of 64 samples is used. In the auto-encoder training phase, the Auto-encoders are trained for 200 epochs to enhance the auto-encoder accuracy. The training process, for the prediction network, is stopped when the MSE value has stopped decreasing for 2 consecutive epochs on the validation set, so the overfitting is avoided.

![FIGURE 4. The sequence diagram for enhanced Deep-DR.](image)

V. TRADITIONAL MACHINE LEARNING ALGORITHMS

To prove the effectiveness of the data integration, and data federation in improving the data quality and its effect on enhancing the model efficiency, two traditional machine learning models are used support vector machine (SVM) and Linear regression (LR).

SVM is one of the most robust machine learning models used for both classification and regression. SVM is a supervised learning model that depends on finding the N-dimensional hyperplane that classifies the points based on the famous theory called Kernel Trick. Using the SVM in regression introduced in 1996 was named SVR. SVR uses only a subset of the training data and neglects points that lie beyond the margin. The SVM is known by its efficiency in high dimensional space, and in cases, where the number of dimensions is higher than the number of samples. SVR is an attempt to solve the objective function (3) with respect to (4) as a constraint:

\[
\min \frac{1}{2} \|w\|^2, \quad \text{Subject to : } |y_i - \langle w, x_i \rangle - b| \leq \varepsilon
\]

where \(x_i\) is a training sample with target value \(y_i\). The inner product plus intercept \(\langle w, x_i \rangle - b\) is the prediction for that sample, and \(\varepsilon\) is a free parameter that serves as a threshold. All predictions have to be within a \(\varepsilon\) range of the true predictions.

LR is a statistical model used for understanding the relationship between input and output (numerical values). Later, LR is used in ML as a supervised learning model to predict continuous values (numerical) as well. The LR depends on assuming a linear relationship between input variables and the output as in (5).

\[
y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon
\]

where \(i \in [1, p]\), \(n\) the observations, \(y_i\) the dependent variables, \(x_i\) the explanatory variables, \(\beta_0\) the y-intercept, and \(\beta_i\) is the slope coefficient for each \(x_i\) and \(\varepsilon\) model’s error.

In [29], both models are used with Principal Component Analysis (PCA) as a dimension reduction method. PCA is the most common method used for dimension reduction of large datasets. PCA depends on constructing the covariance matrix among all possible pairs of variables as in (6) and (7), and then Eigenvectors and eigenvalues will be computed. The Eigenvectors represent the principal vectors and the eigenvalues are the amount of variance. Principal components represent the directions of the data that explain a maximal amount of variance. The maximal amount of variance means the more information it has.

\[
x_i^* = (x_i - \mu_i) / \sigma_i
\]

\[
C = \text{Cov}(X)
\]

where \(x\) is data item in n space \(x = (x_1, x_2, \ldots, \ldots, x_n)\), \(i \in [1, n]\), \(\mu_i\) is the mean of dimension \(i\), \(\sigma_i\) is the standard deviation for the same dimension and Cov(X) is \((n \times n)\) covariance matrix.

We apply both SVM and LR models with PCA for dimension reduction on the data after integration and federation as input.

VI. CONVOLUTIONAL NEURAL NETWORK (CNN)

As CNN is a fully connected neural network, the prediction network is replaced with CNN. The CNN input is the data after reduction using the previously trained encoders.
CNN deals with images which means 2D matrix input. Each sample data is reduced and converted to a $128 \times 128$ matrix with a single channel. The sample is 128 dimensions where 64 are from the mutation encoder and the other 64 are from the gene expression encoder as in [29]. Encoders prove the efficiency of their use in reduction.

CNN is a multilayer perceptron with regularization to avoid overfitting. CNN uses 3 different types of layers: convolutional layer, pooling layer, and fully connected layer.

The used network consists of 3 convolutional layers with (50, 100, and 100) filters for each layer respectively and kernel size (3, 3) for each layer. For max-pooling layers, the max-pooling size is (2, 2). The Activation function is ReLU as in (8) and the Adam optimizer is used with MSE. He’s uniform distribution [31] is used to initialize weight.

$$f(x) = \max(0, x) \quad (8)$$

As CNN is mainly used for image classification, to make CNN predict continuous values for IC$_{50}$, a Flatten layer is added, and then 2 dense layers with 64 neurons as well. The ReLU function is used as an activation function. The output layer with 265 neurons and linear activation function are shown in Fig. 5.

**VII. EXPERIMENTAL RESULTS**

The proposed model target is to predict drug response measured by IC$_{50}$ using both gene expression and mutation profiles for both TCGA and CCLE. TCGA contains 9807 tumors of 33 cancer types and CCLE contains 619 cell lines of 25 tissues with their drug response from GDSC. After data preparation, 19,702 genes are presented for both gene expression and mutation profiles. After the imputation of missing values as in [29], the range of log IC$_{50}$ is from 9.8 to 12.8. Two Auto-encoders are used; one for gene expression profile and the other one for mutation profile. Auto-encoders are used for capturing features and reducing dimensions.

The encoders are 4-layers feedforward networks trained on TCGA data. The prediction model is the integration of a 5-layers feedforward network with the encoders extracted from pre-trained auto-encoders. The output neurons of encoders are connected to the input of the prediction network as illustrated in Fig. 1. The last layer of the predictive network represents the predicted values of IC$_{50}$. The entire model is trained on the CCLE dataset which is divided into 80% for training, 10% for validation, and 10% for testing with 100-fold for cross-validation.

The mutation auto-encoder achieves an MSE value of 0.00594 on the TCGA dataset used for training while the gene expression auto-encoder achieves an MSE value equal to 0.98087. The validation set is used to avoid overfitting.

To test the stability of the Enhanced Deep-DR, the model is run 100 times on different training, validation, and testing sets. The model achieved an average overall MSE in IC$_{50}$ equal to 1.80, 1.80, and 1.78 in training, validation and testing data, respectively. The model converges with an average number of epochs equal to 11.69.

The Enhanced Deep-DR is compared with the original Deep-DR in [29] and CNN as in Table 1. The overall MSE over the testing set is improved. Therefore, that proves the model generalization and the effect of the data federation on the overall performance. The MSE of training and validation increases as a normal consequence of boosting the number of input genes. Despite the rise in MSE value in both training and validation, the MSE value of testing is decreased by 25% which proves the enhancement of model generalization and accuracy. The decrement in the average number of epochs to 11.69 indicates the increase in convergence speed and reduction of execution time. The CNN model has better MSE on both training and validation than Enhanced Deep-DR but less for the testing set. Enhanced Deep-DR has better generalization and prediction performance than CNN according to MSE values on the test set.

**VIII. COMPARATIVE STUDY**

This comparative study is carried out to prove the generalization of the proposed model and the effect of the data federation in improving the data quality. Some recent models used to predict the drug response are selected such as tDDN [15], MC-RR [5], consDeepSignaling [33], and ADRML [17].
Those models utilize different types of features, number of drugs, and learning methods. ADRML [25] predicts drug response with Manifold Learning. The ADRML uses Manifold Learning to predict the drug response by constructing a bipartite graph between drugs and cell lines. ADRML features are the gene expression profile, copy number variation, and mutation profile for cell lines and IC50 for drug response. Deep transfer learning (tDDN) [23] employs Deep learning models for predicting drug responses. The used features are gene expression and drug target pathways. The MC-RR [45] is an ensemble learning algorithm that uses low-rank matrix completion (MC) and ridge regression (RR). The consDeepSignaling [46] is a deep learning model that utilizes 46 signaling pathways, gene expressions, and copy number variation. The DNN [28] is a deep learning framework to predict drug response (IC50) using gene expression.

As shown in Table 2 and 3, Enhanced Deep-DR uses the largest number of genes and drugs to improve the system generalization. It utilizes 19702 genes that have mutations from TCGA and CCLE datasets while DNN uses 16445 genes. All the remaining models implement less than 2000 genes. Enhanced Deep-DR also utilizes 265 drugs, while the nearest one is DNN which uses 251 drugs. The rest models use less than 100 drugs.

Table 4 compares the average of Pearson correlation coefficient (PCC) which indicates the consistency between the real and predicted data. The highest PCC is achieved by Enhanced Deep-DR with a value of 0.89. The PCC of Enhanced Deep-DR ranges between [0.71, 0.96] while the Deep-DR [29] PCC values range [0.7, 0.96] which are almost the same. The tDDN is the second best one with PCC equal = 0.8841, then consDeepSignaling with PCC equal = 0.85. As a result of the previous, the efficiency of deep learning based models gains the best results [47] over other algorithms. See Fig. 6 for more illustrations.

In [29], the author applies SVM and LR with PCA as dimension reduction method to prove the efficiency of the proposed Deep-DR. A separate SVM is trained to predict each drug response; in other words, 265 SVM models are trained and the MSE values are calculated. The same models with the same condition are implemented with the proposed data federation and the results are compared in Table 5. As shown in Table 5, the MSE values are reduced by 85% and 83.5% using LR and SVM models respectively. The decrease in MSE values indicates the effect of the data federation to improve the data quality.

**IX. CONCLUSION AND FUTURE WORK**

This paper introduces both a data federation method and a model that predicts the response of 265 anti-cancer drugs. The data federation is used to integrate the data and improve its quality. Traditional machine learning models such as SVM and LR are used with PCA with and without data federation and the result shows that data federation enhances the accuracy and decreased the mean square error MSE to 1.46 and 1.94, respectively. The DL model, an enhancement to the Deep-DR model (Enhanced Deep-DR), uses encoders for dimension reduction and achieves the best accuracy result with MSE on the test dataset equal to 1.872593. The proposed one shows its reliability and superior performance over the four different models by achieving the best PCC. The introduced Enhanced Deep-DR enhances the model generalization by increasing the number of mutated
genes that may be responsible for the response to different drugs. Also, the introduced data federation method achieves better data. Both mutated genes in TCGA and CCLE are considered as input for the model. The proposed data enhances the model convergence, generalization, and prediction accuracy. The Enhanced Deep-DR achieves better PPC, which means better consistency between real and predicted data.

In future work, Enhanced Deep-DR will be applied to other datasets to prove the methodology’s generalization and efficiency. More dimension reduction methods will be used to replace encoders and test their performance with the pre-treatment accuracy. The Enhanced Deep-DR achieves better PCC, are considered as input for the model. The proposed data federation method different drugs. Also, the introduced data federation method H. Ahmed

REFERENCES

[1] D. M. Roden, H. L. McLeod, M. V. Relling, M. S. Williams, G. A. Mensah, J. F. Peterson, and S. L. Van Driest, “Pharmacogenomics,” Lancet, vol. 394, no. 10197, pp. 521–532, 2019.

[2] M. V. Relling and W. E. Evans, “Pharmacogenomics in the clinic,” Nature, vol. 526, no. 7577, pp. 343–350, Oct. 2015.

[3] M. Mina, F. Raynard, D. Tavernari, E. Battistello, S. Sungalee, S. Saghafinia, T. Laessle, F. Sanchez-Vega, N. Schultz, E. Oriochio, and G. Ciriello, “Conditional selection of genomic alterations dictates cancer evolution and oncogenic dependencies,” Cancer Cell, vol. 32, no. 2, pp. 155–168, Aug. 2017.

[4] S. Park and B. Lehner, “Cancer type-dependent genetic interactions between cancer driver alterations indicate plasticity of epistasis across cell types,” Mol. Syst. Biol., vol. 11, no. 7, p. 824, Jul. 2015.

[5] W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, S. Ramaswamy, P. A. Futreal, A. D. Hamer, M. R. Stratton, C. Benes, U. McDermott, and M. J. Garnett, “Genomics of drug sensitivity in cancer (GDSC): A resource for therapeutic biomarker,” Nucleic Acids Res., vol. 41, pp. 955–961, Jan. 2013.

[6] B. Seaehorde-Ludlow, M. G. Rees, J. H. Cheah, M. Cokol, E. V. Price, M. E. Coletti, V. Jones, N. E. Bodycombe, C. K. Soule, J. Gould, and B. Alexander, “Harnessing connectivity in a large-scale small-molecule sensitivity dataset,” Cancer Discovery, vol. 5, no. 11, pp. 1210–1223, Nov. 2015.

[7] M. R. Grever, S. A. Schepartz, and B. A. Chabner, “The national cancer institute: Cancer drug discovery and development program,” Semin Oncol, vol. 19, no. 6, pp. 623–638, 1992.

[8] M. Ringnér, C. Peterson, and J. Khan. “Analyzing array data using supervised methods,” Pharmacogenomics, vol. 3, no. 3, pp. 403–415, May 2002.

[9] J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, J. F. Peterson, and S. L. Van Driest, “Pharmacogenomics,” Nature Biotechnol., vol. 29, no. 1, pp. 143–189, Jan. 2011.

[10] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, K. Khrabrov, and A. Zhavoronkov, “The cornucopia of meaningful leads: A community effort to assess and improve drug prediction models with applications in dose-response predictions,” Sci. Rep., vol. 10, no. 1, pp. 1–11, Dec. 2020.

[11] J. Chong, H. Yue, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, and A. Reddy, “The cancer cell line encyclopedia enables predictive modelling of anti-cancer drug sensitivity,” Nature, vol. 548, no. 7667, pp. 201–206, Jun. 2017.

[12] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, K. Khrabrov, and A. Zhavoronkov, “The cornucopia of meaningful leads: A community effort to assess and improve drug prediction models with applications in dose-response predictions,” Sci. Rep., vol. 10, no. 1, pp. 1–11, Dec. 2020.
HANAN AHMED received the bachelor’s and M.Sc. degrees in machine learning in computer vision from the Scientific Computing Department, Faculty of Computer and Information Sciences, Cairo, Egypt, in 2010 and 2016, respectively. She is currently pursuing the Ph.D. degree in deep learning applications in biomedical.

She was a teaching assistant for several undergraduate courses and worked as an Associate Editor of International Journal of Software Science and Computational Intelligence (IJSSCI) (IGI Global). Her research interests include machine learning and deep learning applications.

SAFWAT HAMAD received the bachelor’s degree, in 2000, the M.Sc. degree in modeling, simulation, and visualization from the Faculty of Computer and Information Sciences (FCIS), Ain Shams University (ASU), Cairo, Egypt, and the joint Ph.D. degree in high-performance computing from the Computer Science and Engineering Department, University of Connecticut, USA, and FCIS, ASU, in 2008.

He was a teaching assistant for several undergraduate courses. He has been the Chair of the Scientific Computing Department, FCIS, ASU, since 2017. He is currently a Professor. His research interests include image and video processing, computational biology, machine learning, encryption, and security.

ASHRAF SAAD HUSSEIN received the Ph.D. degree in aerospace engineering, in 1999. He is currently the President of King Salman International University, Egypt. He is also a Professor in scientific computing at the Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt. His research interests include modeling and simulation, data science, natural language processing, and high-performance computing applications. From 1996 to 2010, he served as a Senior Research and Development Engineer and then as a Visiting Scientist at the IBM Technology Development Center and Cairo Center for Advanced Studies. He joined Ain Shams University, in 2001, where he delivered several courses, activities, and projects related to academia and research. He was appointed at the Supreme Council of Egyptian Universities, in 2005, to lead several country-wise projects funded by the World Bank and EC, until 2010. In 2010, he started establishing the Electronic and Knowledge Services Center, Supreme Council of Egyptian Universities, to offer several e-services to the stakeholders of government universities. From 2013 to 2019, he had a secondment to the headquarters of the Arab Open University to serve as the Dean of Engineering, and Computing, then as the Vice President for Information and Education Technology. He has authored more than 100 published scientific papers and issued patents. He is a member of several professional associations and a Senior Member of the Association of Computing Machinery (ACM).