An Artificial Intelligence Approach Based on Hybrid CNN-XGB Model to Achieve High Prediction Accuracy through Feature Extraction, Classification and Regression for Enhancing Drug Discovery in Biomedicine

Mukesh Madanan[1],Biju T.Sayed[2],Nurul Akhmal Mohd Zulkefli[3],Nitha C.Velayudhan[4]

[1]Department of Computer Science, Dhofar University,Salalah,Oman
[2]Department of Computer Science,Dhofar University,Salalah,Oman
[3]Department of Computer Science,Dhofar University,Salalah,Oman
[4]Department of Computer Science & Engineering
Noorul Islam Centre for Higher Education
Tamil Nadu,India

Received: November 16, 2020. Revised: May 3, 2021. Accepted: May 25, 2021. Published: June 2, 2021.

Abstract— In the field of biomedicine, drug discovery is the cycle by which new and upcoming medicines are tested and invented to cure ailments. Drug discovery and improvement is an extensive, complex, and exorbitant cycle, settled in with a serious extent of vulnerability that a drug will really be successful or not. Developing new drugs have several challenges to enrich the current field of biomedicine. Among these ultimatums, predicting the reaction of the cell line to the injected or consumed drug is a significant point and this can minimize the cost of drug discovery in sophisticated fashion with a stress on the minimum computational time. Herein, the paper proposes a deep neural network structure as the Convolutional Neural Network (CNN) to detain the gene expression features of the cell line and then use the resulting abstract features as the input data of the XGBoost for drug response prediction. Dataset constituting previously identified molecular features of cancers associated to anti-cancer drugs are used for comparison with existing methods and proposed Hybrid CNNXGB model. The results evidently depicted that the predicted model can attain considerable enhanced performance in the prediction accuracy of drug efficiency.

Keywords— Convolutional Neural Network, Drug Discovery, Multi Fusion Neural Network, XGBoost

I. Introduction

Drug discovery, which is the process of finding new candidate drugs, is significant for biomedicine [1]. Innovative remedy and mechanism endorsement in the US usually acquires a standard of seven to twelve years of accurate testing and numerous safety phases, and many individuals are involved from the pre-clinical testing phase to the sanction of the drug by FDA [2]. In addition to this, the expenses for the improvement of health checkup devices run into huge dollars, and an ongoing report recommends that the whole expense for novel drug testing and discovery could be in overabundance of $1 billion [3]. Also, majority of drugs being created neglect to arrive at the market because of reasons relating to poisonous or low adequacy, erroneous prediction of drug etc. [4]. Thus, in spite of our understanding about the new diseases continuously, efficiency in producing new treatments constantly has not been in the satisfactory level for the past few years.

Generally, the process of dynamic drug discovery comprises of five stages. In first stage, the predicting model is initialized utilizing the reaction information in the midst of the drugs and helping cell lines. The drugs reactions on the objective cell line are predicted in second stage. In the third stage, top capable drugs from the second stage are filtered and are selected for the next phase. In fourth stage, the experiments are carried out with the involvement of test subjects/individuals. Finally, in the fifth stage, information regarding the revealed reactions is utilized to refresh the predicting method to make it more exact. During the stages from the second to fifth, the experiments are iteratively repeated number of times until a satisfactory result is achieved. The stage four is the most costly and advanced stage in drug innovation, and the expense can only be diminished by discontinuing the circle at right time as conceivable drug is effective. The activity of drug discovery is depicted in Fig.1

.

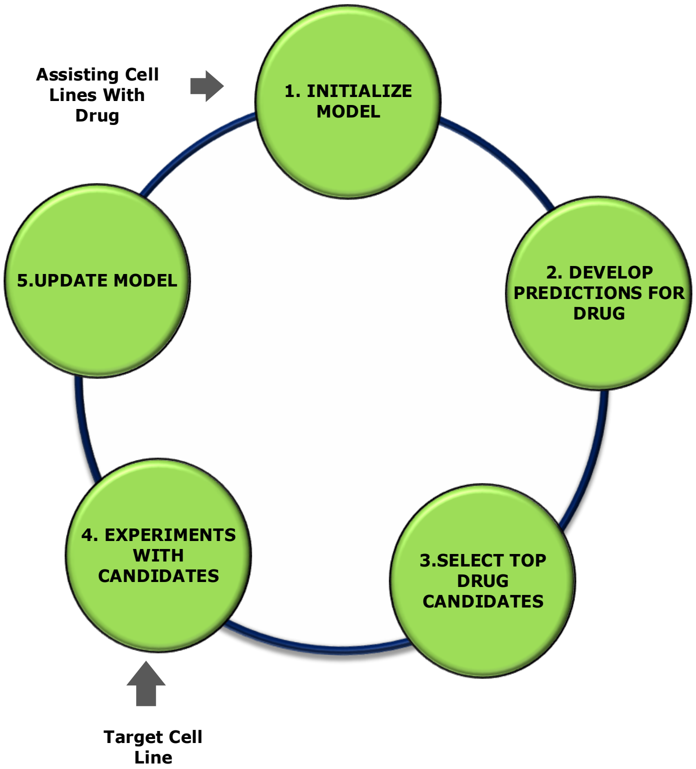
Assisting Cell Lines With Drug
1. INITIALIZE MODEL
2. DEVELOP PREDICTIONS FOR DRUG
3.SELECT TOP DRUG CANDIDATES
4. EXPERIMENTS WITH CANDIDATES
5.UPDATE MODEL
Target Cell Line

Fig.1. Activity of Drug Discovery

The field of biomedicine is enriched with numerous activities in the research in terms of medical applications. Although the expansion of novel drugs leaves a huge chink, it is expected to get better in present era of biomedicine [5].But several factors such as the expenditure and time-consumption are always a hindrance to the discovery of new novel drugs in biomedicine. Research and development in drug discovery pipeline should discover noteworthy effectiveness gains, if the business has to keep producing novel drugs in zone to determine the drug targets interactions of cell line [6]. It has assisted for the testing of the preclinical safety of recently created pharmaceutical drugs using silico, in vivo, and in vitro before being regulated in people [7].

Artificial Intelligent experts have previously dealt with fundamental difficulties such as outstandingly minimizing rate, occasion, as well as employing experiments during the beginning phases of drug discovery process [8]. In a significant number of these studies, a correlation with Machine Learning(ML) techniques has been made and these studies exhibited that Deep Learning accomplishes comparable or preferred execution over other ML methods for various properties together with forecast of biological movement, Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) properties as well as physico-chemical parameters [9]. Deep Learning methods have also shown to accomplish prediction in little molecular drug detection and progress [10].

Numerous tools and techniques are provided by machine learning approach for the discovery and making decisions when the input is abundant [32]. A number of algorithms such as support vector machines(SVM), artificial neural networks(ANN) and convolutional neural networks(CNN) have been utilized to provide drug responses to various drugs. The proposed deep learning hybrid model in this paper is capable enough to predict the drug responses on cell lines with extremely elevated accuracy, and exactly opt for a good number of positive result-oriented drugs for a new cell line.

The CNN and XGBoost algorithms are considered to be super classifiers[31]. Individually when used CNN and XGBoost have several drawbacks [31]. The proposed model is a hybrid of CNN-XGBoost, which can be utilized to predict the cell line responses to various newly discovered drugs.

II. RELATED WORK

The Machine learning techniques are applicable for all stages of drug discovery to improve the invention in an effective way [11]. Recently, Machine learning technology along with active learning approaches are considered for acquiring effective drugs and to enhance the predictive accuracy on drug responses on cell line at rapidly increasing pace in domains of bioinformatics and drug discovery [12]. Based on this, the author presented the Random Forest (RF) to be successful for predicting protein interactions in [13] [14]. In addition, the active learning strategies were utilized to achieve the superior exactness via predicting the useful protein sets for labeling. The algorithm was applied to choose candidate protein-pairs whose interface condition and its exactness of prediction is processed in terms of precision, F score and Recall using experiments. This technique is able to facilitate the reduction in the expense and endeavor constructing the human protein interaction, by generously diminishing the various new in-vitro experiments requisite to decide explicit p-p interaction pairs. Addressing the difficulty of multi-classification for compounds, these techniques were unsuccessful in identifying drugs that can interact on cell line.

Deep Neural Network (DNN) was utilized as a sensible Quantitative Structure-Activity Relationship (QSAR) technique [15], and can easily outperform Random Forest in most experiments. DNNs preserve more data and makes enhanced prospective predictions than RF on a set of large diverse QSAR data sets that were retrieved from Merck's drug discovery effort. While the magnitude of the change in Coefficient of Determination (R^2) comparative with RF appeared small in some datasets, hence, they proved that this model is superior to RF. The statement disclosed that the performance of DNN changes relying upon the activation function utilized and the network architecture (number of hidden layers as well as number of neurons in each layer).

Due to this, prediction accuracy was not obtained in sufficient level in drug discovery process. A type of deep neural network which could be satisfactorily be used in classification of images is the Convolutional Neural Network(CNN)[30].Numerous application of CNN are evident in AlexNet, VGGNet and ResNet.

For prediction of drug-target interactions dependent on kernelized matrix, factorization is presented in [16] with active learning approaches. It was shown that the prediction accuracy was attained for drug-target interactions with minimum number of experiments. Most probably, the regression model for predicting correctness of replicated active learning that succeeds the phase could be improved. The researchers endeavor to establish the best standards to discontinue the drug detection progression as it nears the final stage that could reasonably be expected regardless of the way that did not aspire to pick the best sensible drugs as fast as possible. Subsequently, for the estimation of reasoning in an abrupt approach, the authors build up the Computer-Aided Drug Design Based on Active Learning in [17]. The researchers addressed the issue of discovering active compounds at different phases of the drug discovery process using the Support Vector Machines [18]. SVM was adequately applied to different active learning issues and it exhibited that choosing unlabeled model in accurate way utilizing hyper plane yielded forgo results. Furthermore, this model is not adequate to compete with accurate prediction in drug discovery in terms of time consumption and accuracy [19]. A combination of LASSO regressions with DNN model for extracting the features for protein-specific and a drug-specific feature separately was also employed [19]. Supplementary to this, the author offered the lasso method to upgrade the prediction of sensitivity of drug dependent on active learning. It is indicated that the LASSO-DNN model has superior advantage in prediction of drug target interactions through selected features. Nevertheless, a lasso regression advent was not upheld to choose the sensitivity of the drug in a gratifying demeanor [20]. Moreover, the model did not commence to comprehend when it has minimum two different labels in training data. Attempting to subdue the cold start issue for linear model, authors introduced a CNN based active learning representation that can quicken the process of drug discovery in [21]. The result demonstrated very less in classification than in regression. The explanation for this is the presence of significantly more data contained in regression. Nonetheless, the classification did not completely express its potential and a linear method to predict for drugs and cell lines individually and by using the average scores, the most promised drugs are preferred. We aim to propose a predictive model to improve the prediction through classification and regression utilizing XGBoost with Convolutional Neural Network. The objective focused is to predict an applicable drug and furthermore classify the labeled drug viably.

III. CLASSIFICATION

In drug discovery, very accurate prediction is needed for creating new target cell line based on assisting cell lines and precisely selecting the most promising drug for a new cell line. However, the process being more time consuming with lot of number of experiments being conducted and then selecting the sensitive drug with low value becomes a major challenge. These drawbacks could be minimized by increasing the accuracy in classification and reducing the Root Mean Square Error in regression. The efficient prediction accuracy is evaluated using following metrics:

- Accuracy
- Precision
- Recall
- F1 Score
- Coefficient of Determination (R^2)
- Root Mean Square Error (RMSE)

A. Accuracy

With the intension of achieving the better prediction, computing of accuracy of classification is sustainable and it is a main performance metric. The high accuracy shows that the proposed model is acceptable in classification based on confusion matrix. The accuracy3 is calculated using the following formula:

$$Accuracy = \frac{True\ Positive\ + TrueNegative}{True\ Positive + False\ Positive + \ False\ Negative\ + True\ Negative} \quad (1)$$

B. Precision

It is the one of the metric to find the classification accuracy based on denoting the correct positive prediction. Precision is determined as the ratio of correctly predicted positive value (True Positive) to the total of correctly predicted positive value and False Positive which occurs when the actual class differs with the predicted class.

$$Precision\ = \frac{True\ Positive}{True\ Positive\ + False\ Positive} \quad (2)$$

C. Recall

The metric of Recall is indicated by the missed positive prediction. It is also called likewise sensitivity. The ratio of True Positive Rate (TPR) to the sum of True positive (TP) and False Negative (FN) is Recall. The value of 0.1 is obtained when the model does not produce value of False Negative.

$$Recall\ = \frac{True\ Positive}{True\ Positive\ + False\ Negative} \quad (3)$$

D. F1 Score

F1 Score is a common metric to evaluate the classification accuracy based on Precision and Recall functions. It is also called as F- measure. It is used to find unbalanced classification in accuracy. Perfect score of F1 score is one.

$$F1 = \frac{2\, Precision * Recall}{Precision + Recall} \qquad (4)$$

E. *Coefficient of Determination(R^2)*

The coefficient of determination which is also known as R-squared (or R^2) is the output of regression analysis. It is used to expose the linear relationship between dependent and independent variables within range of 0.0 and 1.0. For perfect fit in regression analysis, value of 1 can be predicted. The definition of coefficient of perseverance is the square of the correlation (r) amid predicted values and actual values.

$$R^2 or\, r^2 = \frac{\sum(\widehat{bi} - \bar{b})^2}{\sum bi - \bar{b})^2} \qquad (5)$$

where, bi as the observed values of the dependent variable, $\bar{b}$ denotes its mean, and $\widehat{bi}$ indicates fitted value.

F. *Root Mean Square Error(RMSE)*

Root Mean Square Error (RMSE) is the standard deviation to measure the error of prediction. It has direct relationship with correlation coefficient. Lower value of RMSE shows that our proposed model consists of low error. RMS Error is calculated using below formula,

$$RMSE = \sqrt{1 - r^2 SD_b} \qquad (6)$$

where, r^2 indicates correlation coefficient and SD_b denotes standard deviation of b.

IV. PROPOSED MODEL

The research proposes an ensemble novel active learning drug-cell line interactions method. Here SMILES notations as input data of drug and cell line are taken. Two channels of Convolutional Neural Network (CNN) are used that is needed for drug and cell lines, where Convolutional Neural Network performs as a feature extractor [22] to spontaneously obtain features from the cell line. The final layer of fully connected (FC) output layer in CNN is changed to XGBoost. It performs as classifier to perform the regression and classification analysis and predict the new target cell line. The proposed model is displayed in Fig.2.

In the first stage, the symbols are collected from SMILES dataset and drug SMILES are converted into a two dimensional Boolean matrix $S * C$, where S indicates count of symbol and C denotes longest length of SMILES among all drugs. Hence, binary conversion is achieved for processing the SMILES. For cell lines, cell mutation states are converted into one dimensional Boolean vector or binary vector. Thus, SMILES and features of cell lines are converted into one hot format encoding. During the second stage, separate CNN is utilized to select the features from drug and cell lines of one hot format. The drug for CNN network branch is then applied to the 1D convolution operation with 28 channels and within one single channel 1D convolution operation is done for cell line respectively through CNN network branch containing three layer of CNN such as Convolution, Pooling and RELU. Then, by training the Convolutional Neural Network model along with output layer of Fully Connected (FC), the training error is fixed. The process is continued until all the errors are fixed completely. The feature extraction occurs in this phase. After the CNN model is trained, FC layer of Convolutional Neural Network is changed to XGBoost to do regression and classification analysis in final stage. The predicted features and testing data are fed into XGBoost classifier as input feature vector. It can perform the regression and classification analysis to enhance the prediction accuracy in new target cell. In order to obtain new target cell lines, the sensitivity of drug is predicted using regression analysis and also sensitivity and

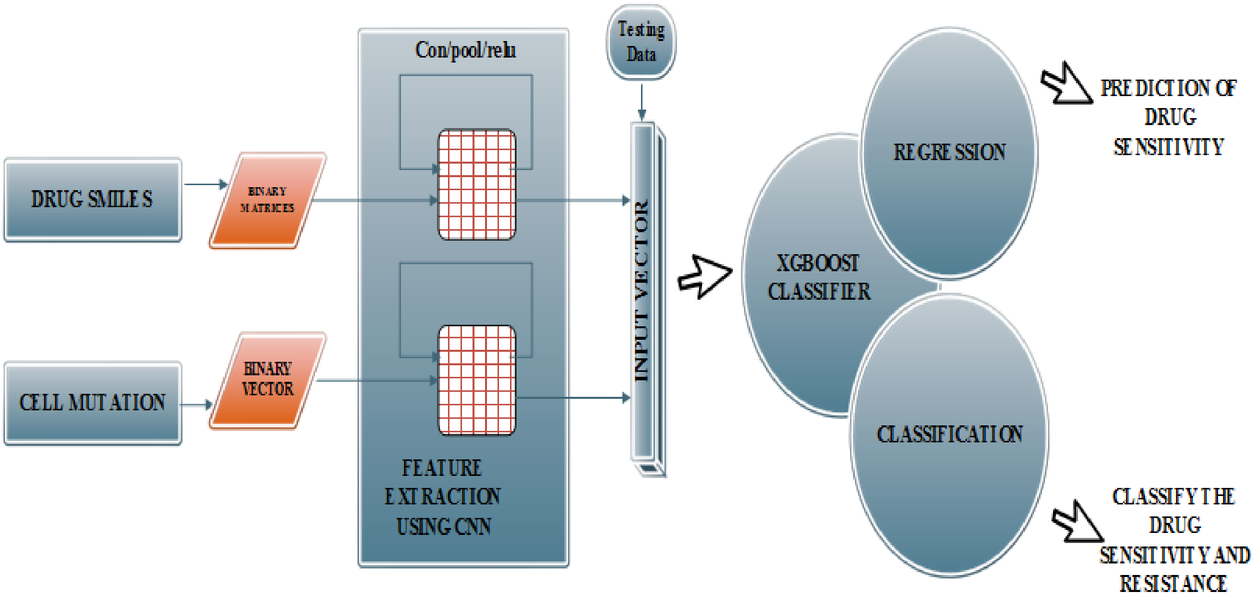

Fig.2. The Proposed Model

resistance of drug are classified using classification analysis in a successful manner. Hence, the proposed model is able to discover redundant useful drug with limited count of experiments.

A. *Feature Extraction Using CNN*

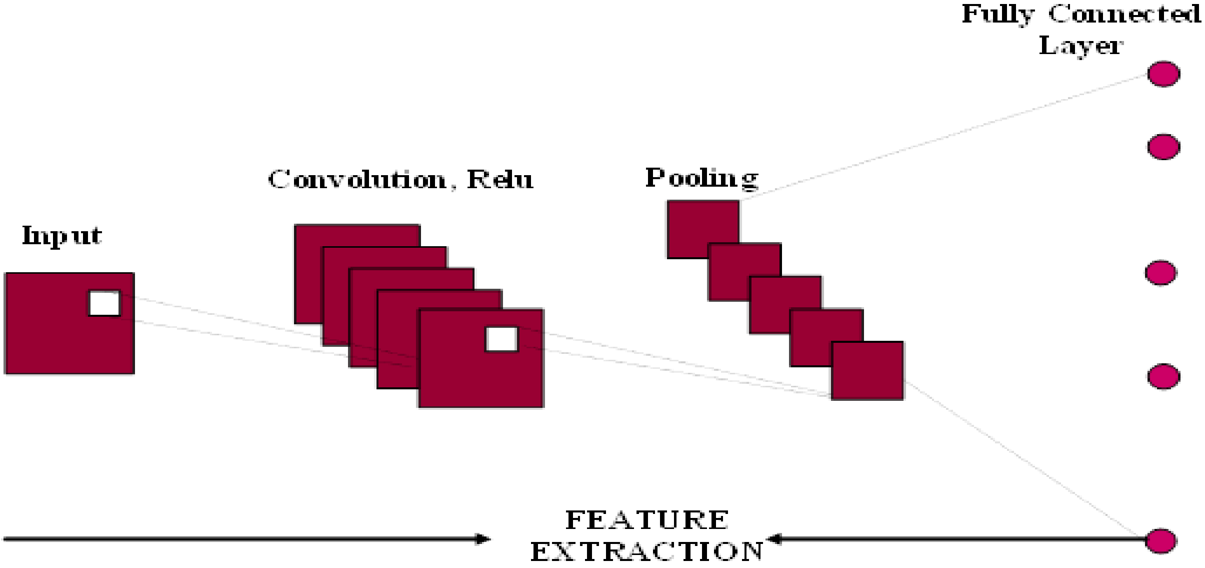

Fig.3. Convolutional Neural Network Extraction

Feature Extraction is one of the popular methods to extract the information as features between the drug targets in drug discovery. It is used to enhance the prediction accuracy precisely. By avoiding more features than samples, the efficient feature extraction can be achieved. Deep learning method of CNN is powerful enough to realize the high-level features with neural networks. Thus, the feature extraction is done by Convolutional Neural Network with the ability to learn of high-level representations of data in our model.

1. *Convolutional Neural Network (CNN)*

In order to obtain the performance of regression and classification, the professor Yann LeCun introduced the Convolutional neural network (CNN) with the help of his colleagues in University of Toronto. CNN is a neural network which perfectly classifies images[30]. It consisted of the Convolutional Layer, Pooling Layer, Rectified Linear Unit (ReLU) and Fully Connected Layer (FC). The training model of feature extraction is achieved through the Convolutional Layer, Pooling Layer and Rectified Linear Unit as shown in Fig. 3 and the Fully Connected layer accomplishes the classification performances. CNN is utilized to minimize the intricacy of network arrangement and the count of parameters via the concepts of responsive fields, sub-sampling (pooling) and weight sharing. For this reason, we have taken two convolutional neural networks with regards to the drug and the cell line.

The purposeful features are extracted from the drugs and the cell line using the two Convolutional Networks. For cell line, one-dimensional Convolutional activity is performed using the single channel. Along with 28 channels, 1D convolution is performed for drugs in CNN. This method figures out how to extract features from sequences of examination and how to map the internal features to various activity types. Accordingly, the convolution operation of CNN is applied to the path of the SMILES. The path of symbols from SMILES is considered like various channels. Notwithstanding the quantity of channels, the excess hyper-parameters for the each CNN are equivalent.

2. *Convolutional Layer*

In Convolutional Neural Network architecture, the core building block of Convolutional layer are the hidden layers which are the cornerstones. CNN is used to perform the challenged computational lifting through making them convolved with square grid of weights and input. Filter process is done in the Convolutional Layer for extracting the features. The spatial size of the output volume is calculated as follow,

$$\frac{V - N + 2P}{S} + 1 \tag{7}$$

where, V indicates the function of the input volume size, N denotes size of the neurons of Convolutional Layer based on field, S means the stride and P indicates the volume of zero padding utilized on the border.

Convolution layer are provided mathematically in the form as below,

$$CON_p^l(i,j) = \sigma \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} x(i-u, j, -v) \times \theta_{p,q}^l(u,v) + b_q^l \tag{8}$$

where CON is a convolution layer, l denotes the layer, p, q indicates the map indices for present and following layers, respectively, i, j are the indices of row and column based on feature map. σ is the activation method, x represent image or activation map, θ and b denotes kernel and bias.

3. *Pooling Layer*

In Convolutional Network Architecture, the pooling layer is inserted irregularly amid sequential Convolutional Layers. In order to minimize the count of parameters and measurement in the network, this layer performs subsample process on input images. Moreover, dimension of feature from the convolutional layer are limited by activating the down sample method. The max pooling is done using the max operation through resize and spatially on every depth slice of the input individually. Utilizing the zero padding method, the down-sampling is done for input size using the filter kernel as below way,

$$\frac{V-N}{S}+1 \tag{9}$$

The pooling layer are provided mathematically in form as below,

$$POLL_p^l\,(i,j) = max \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} CON_p^l(i-u, j-v) \tag{10}$$

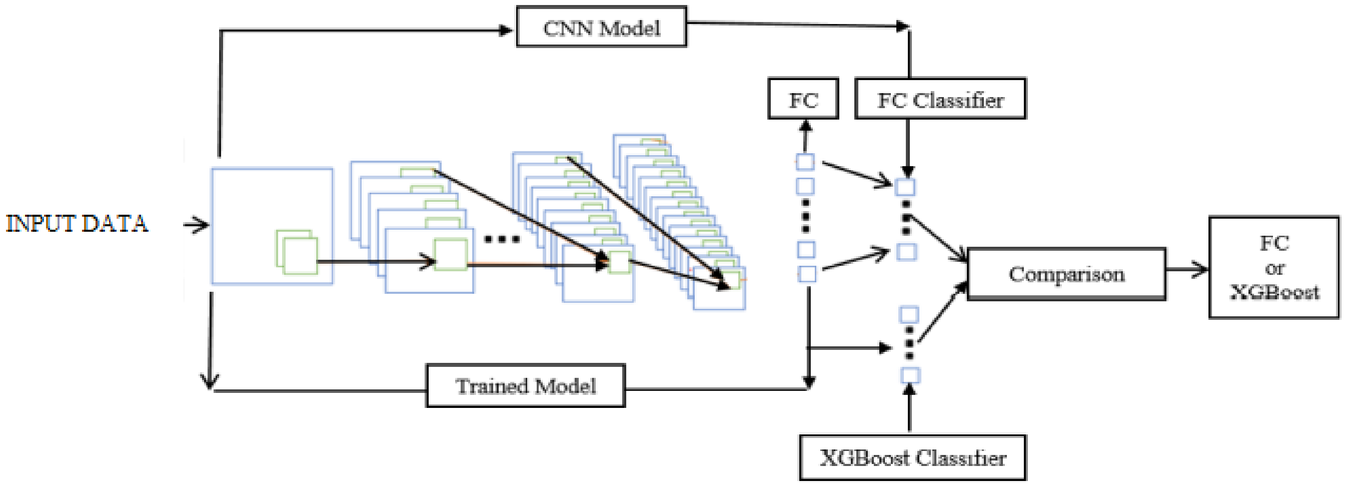

Fig 4.Fully Connected and XGBoost Classifier

4. *Rectified Linear Unit*

To use the non-linearity, the rectifier activation function is utilized against for linear activation function in the network. According to this statement, to eliminate the any negative value, the nonlinear activation function of ReLu (Rectified Linear Unit) is applicable.

5. *Fully Connected Layer*

The fully connected (FC) layer acts as trainable classifier where their activations could be calculated by means of a matrix multiplication in the course of a bias offset. The main contrast between the FC layer and convolution layer is that the neurons in the convolution layer are associated to a neighborhood area in the input and that a considerable lot of the neurons in a convolution volume share parameters. In any case, the neurons in the two layers still compute dot products, so their functional form is indistinguishable [23]. The fully connected layer takes the input from the convolution layer for classification intension. The most intuitive way to achieve correct labeling is to prolong the loss of cross-entropy. The binary cross entropy loss (BCE) over sigmoid activation has shown better performance when applied to multi-label tasks rather than using the cross-entropy loss function. The loss of binary cross-entropy is,

$$min -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{L} y_{i,j}\log\left(\sigma\left(f_{ij}\right)\right)+\left(1-y_{ij}\right)\log(1-\sigma\left(f_{ij}\right))) \tag{11}$$

where,

$$\sigma\ (x) = \frac{1}{1+\ e^{-x}}$$

There is a convolution operator in each of the three layers, with the length of the filter and the length of the stride being 7, 1 respectively, and the max pooling with zero padding. The pooling zone and stage size is 3. Separately, the channel numbers for the three CNN layers are 40, 80, and 60. The FC layer serves as a training model and can minimize the error and mark correct features from drug and cell line. The Fig 4 depicts a model using fully connected layer and XGBoost used for classifying in CNN. The Fig 5 depicts a 8 layer network that comprises of 6 convolutional layers and two fully connected layers. A max-pooling layer and drop out layer follows 2 convolutional and ReLU layer. Following the max-pooling layer is a fully connected layer. The fully connected layer holds 512 neurons. The final classification is performed by the last fully connected layer, which consists of 502 neurons.

S_1 x S_2 is the size of the RGB image matrix which is the input.L_1,L_2,L_3,L_4,L_5 and L_6 are feature maps employed to calculate the features. Feature map neurons is connected to nine inputs and a 3x3 convolutional filtering kernel is defined. The kernel is shared by each neuron on the feature map and weights are connected.L_7 is a a hidden layer and L_8 is the output layer to the XGBoost.

B. *XGBOOST Classifier*

Boosting method was introduced for weak learner (base classifier) and on applying it the learner changed into superior learner based on assigning weights modulation to the learns. In light of this idea, numerous enhancements have been presented to enhance boosting. Gradient Boosting, AdaBoost[24], ARCing, and XGBoost algorithms [25] are the options of the customized boosting algorithms. Corresponding the approach of gradient boosting, it includes essentially three stages.

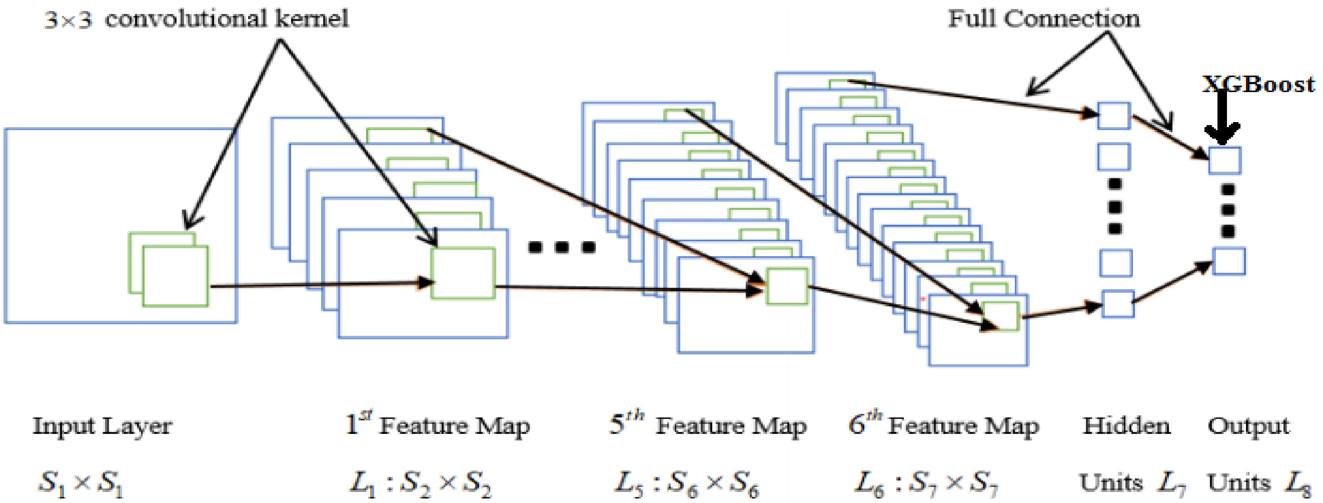

Fig5. Trained CNN used as input to XGBoost Classifier

First, it recognizes the appropriate differentiable loss function for a known issue. The advantage of the gradient boosting model is that you don't need to extract the latest algorithms for different loss functions. It is necessary that a reasonable loss function is selected and then integrated into the gradient boosting process. Second, a decision tree is chosen as a weak learner in gradient boosting and it is developed to make the predictions. Third, the weak learner's loss functions are predicted by generating an addictive model. This method of adding the trees occurs one by one. Then, the output generated in the new tree is applied to the output of the previous trees series to increase the model's final output. This process will proceed until desired loss function value is obtained.

XGBoost is Extreme Gradient Boosting which at its core has gradient boost. Nonetheless, the process of combining weak learners is not done individually in the other gradient boosting algorithms. Thus, it takes a multi-threaded method where the machine's CPU core is fully evaluated owing to higher speed and performance. It is a scalable and effective implementation of the gradient boosting concept as in [26]. It provides a powerful linear model solution and algorithm for tree-learning. It provides multiple objective functions namely regression, ranking, and classification. XGBoost also supports evaluation function and customized objective function. In addition, there is a sparse conscious implementation that involves automated conduct of losing data values, after that blocking structure to enable tree building parallelization, and continuous training process so that a previously configured model on new data can be further enhanced.

Tree boosting is a form of machine learning which is very powerful and commonly used. Tree boosting is a method to learn to increase the classification ability of poor classifiers by recursively introducing new decision trees to the decision trees ensembles. XGBoost is a regression tree with the same rules for judgment as the decision tree. The inside nodes represent values for test attributes in the regression tree, and the leaf nodes with scores are a decision. The XGBoost is a tree-ensemble approach that includes multiple classifications and regression trees. [27].

For a set of data with n classes and m features $D = \{x_i, y_i)(|D|) = x_i \in \mathbb{R}^{n \times m}, y_i \in \mathbb{R}^n$ the arithmetical illustration of the tree ensemble model is represented as,

$$\hat{yi} = \sum_{t=1}^{t} h_t(x_i), h_t \in R \qquad (12)$$

where k seems to be the number of trees, h is just a function in the functional space R, and R is now the set of decision of regression and classification trees. The prediction of a tree boosting to a $(x_i, y_i,)$ is then provided by,

$$\hat{yi} = \sum_{t=1}^{M} h_t(x_i) \tag{13}$$

where $h_t(x_i) = w_q(x_i)$is the prediction of the t-th decision tree containing leaf weights w_q on a data point x_i, and M denotes the number of members in the ensemble. XGBoost has the same gradient boosting principle as that of the gradient tree boosting algorithm, but does make a slight improvement on the normalized goal. The predictive function of decision trees h_t can therefore be discovered by reducing the objective function,

$$F_1(\theta) = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{t=1}^{M} \Omega(h_t) \tag{14}$$

The former $\sum_{i=1}^{N} l(y_i, \hat{y}_i)$ is a differentiable loss function that tests whether the model is appropriate for data set for training. The last $\sum_{t=1}^{M} \Omega(h_t)$ is an object punishing the model's complexity. The second term does penalize the model's sophistication. As the model's complexity increases, it will subtract the corresponding performance. This has the function of preventing over-fitting of the model. In XGBoost the complexity of regularization can be defined as,

$$\Omega(h) = \gamma K + \frac{1}{2}\lambda \sum_{j=1}^{K} \omega_j^{\,2} \tag{15}$$

where γ denotes gamma parameter, K indicates number of leaves, λ represents L2 regularization term on weights in the model, Note that this regularization penalizes all large weights on the leaf nodes (similar to L2-regularization) and has huge partitions and, ω seems to be vector score on leaves and $\gamma \ and \ \lambda$ imply constant coefficient. As described above, the tree boosts the ensemble of decision trees iterative manner, and then the prediction of k-the iteration can be represented as,

$$\hat{y}_i^{(k)} = \sum_{j=1}^{k} \hat{y}_i^{(k-1)} + h_t(x_i) \tag{16}$$

The objective function ( ) may be updated at step k as,

$$F_1(\theta) \ or \ Obj^{(t)} = \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(k-1)}\right) + \sum_{t=1}^{M} \Omega(h_t) \tag{17}$$

Gradient boosting is a system designed for regression and classification in machine learning. It exercises its model in a preservative approach. At every instance, a new tree is added, and the new score is equal to the corresponding score plus the score of a new tree. The gradient boosting process utilizes second but rather-of Taylor expansion to boost the loss function and removes the constant term to achieve the simplest objective. The concluding objective function is exposed,

$$Obj^{(t)} = \sum_{i=1}^{n} [g_i \omega_q(x_i) + \frac{1}{2} h_i \omega_q^2(x_i)] + \gamma K + \frac{1}{2}\lambda \sum_{j=1}^{K} \omega_j^{\,2} \tag{18}$$

$$= \sum_{j=1}^{K} [(\sum_{i \in I_j} g_i)\omega_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right)\omega_j^2] + \gamma K \tag{19}$$

where $g_i = \vartheta \hat{y}_i^{(k-1)} l(y_i, \hat{y}_i^{(k-1)}=$ is the loss function's primary gradient statistics and $h_i = \vartheta^2_{\hat{y}_i^{(k-1)}} l(y_i, \hat{y}_i^{(k-1)}$ is the second one. Decision of the optimal weight $\omega_j$of leaf j meant for a permanent tree structure, q(x) have obtained by the below equation,

$$\omega_j^{obj} = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{20}$$

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{\frac{1}{2}\sum_{j=1}^{K} \sum_{i \in I_j} g_i^{\ 2}}{\sum_{i \in I_j} h_i + \lambda} + \gamma K \tag{21}$$

The tree measures quality of the structure as in – sample performance of h_k and finding the decision tree which is minimizing the value. On the other hand, discovering and denoting all probable tree which preserves and minimizes the quality of the structure of the tree is impractical. Instead, an approximate greedy algorithm that works to optimize one tree level at a time by attempting to find optimal data fragments, leading to a tree with a local minimum of the tree structure standard, which is then applied to ensemble. The below Eq.22 is used over and over again for the estimation of split candidates to achieve the leaf node score.

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda}\right] - \gamma \tag{22}$$

C. *Hybrid CNN-XGB Model*

In [28], CNN is used as a feature extractor along with SVM classifier, which is not sufficient to extract the features as well as to perform classification. Especially, while extracting the features from sequence data in drug discovery it is unsuitable. To enhance the classification accuracy, the gradient boosting algorithm of XGBoost is used [29]. It is also called as integrated learning algorithm. Based on this, the hybrid CNN-XGB model was developed by removing the FC layer in CNN and adding the XGBoost classifier instead. The outputs from the fully connected layer in CNN network are estimated and the probability for input sample for another one are selected. Here, the activation function is used to consider the each output probability. Moreover, the trainable weights of previously hidden layers linear combination of the outputs is given as an input of the activation function. Taking the hidden layers performance standards can also be used as features for each and every classifier. Initially, the structured and labeled

SMILES Notation and unlabeled samples are given to the input layer, and the CNN model with the FC output layer is trained until to fix training error in training process. Here, validation, and testing datasets are predicted. In the end of the training process in Convolutional Neural Network, the FC layer in CNN is eliminated. Then, both predicted values of training and testing datasets are converted into input feature vector through the mathematically. This input feature vector is provided the input for the XGBoost model. The training time is depending upon the count of classes in dataset because XGBoost develops number of label based on its tree. By using number of iterations, XGBoost classifier is trained. The classification of sensitive and resistance in drug is achieved through the predicted values. The regression method is used to predict sensitive drug and also calculate the error amid truth sensitive drug and predicted sensitive drug.

V. Result

The prediction accuracy can be increased by active classification and regression method. The performance analysis of classification and regression is used to find effective model to predict the accuracy in target cell line. Thus, the performance of our hybrid model is compared with existing model such as Neural and Lasso for the tasks of regression and classification and also we can finalize that which one discover the good drugs for the target cell lines in fast manner. The database of SMILES are downloaded from Library of the Integrated Network-based Cellular Signatures (LINCS) and analysis the experiments in efficient way.

A. *Classification Analysis*

An efficient classification is needed for predicting the useful accumulated number of sensitive drugs. In this way, the performance of accurate classification provides exact prediction and also finds out the new target cell efficiently. Some metrics are used to compute the accuracy between true sensitive drug and predicted sensitive drug and also provide the desirable output such as sensitive and resistance in following manner.

$$\text{Precision} = \frac{\sum x_i \epsilon E_j \frac{TP}{TP_{xi} + FP_{xi}}}{|E_j|} \qquad (23)$$

Table 1. Accuracy Comparison of CNN-XGBoost

Method	Precision	Recall	F1	ACC
Lasso	0.84	0.82	2.46	86%
CNN	0.91	0.89	2.67	92%
CNN-XGB	0.99	0.98	2.94	98%

$$\text{Recall} = \frac{\sum x_i \epsilon A_j \frac{TP}{TP_{xi} + FN_{xi}}}{|A_j|} \qquad (24)$$

$$F1 = \frac{2 \times Precision \times recall}{recall + Precision} \qquad (25)$$

where A_j and E_j are the set of sensitive drug for true IC_{50} and the number of sensitive drug for predicted IC_{50} respectively and xi indicates all target cell in testing set.

By using below formula,

$$ACC = \frac{TP}{TP + FP + FN} \qquad (26)$$

The Precision, Recall, F1 score and ACC for all target cell are shown in Table 1 for given dataset. As seen in Table 1, the CNN-XGBoost value exceeds for the ACC and F1 score of all of the existing models and also outstanding performance in drug sensitivity prediction accuracy is 98%. Further, the metrics are used to find the prediction accuracy. Therefore, in terms of precision and recall, CNN-XGBoost has the best results of all the methods.

The Accuracy of Lasso, CNN and CNNXBG is plotted in percentage form in graph analysis in below Fig 6

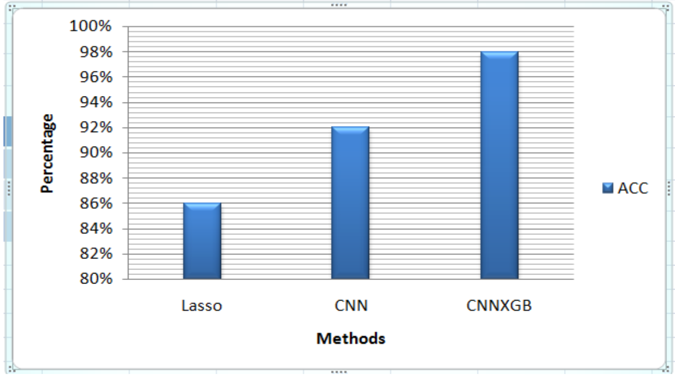

Fig.6. Graphical Analysis of Classification Accuracy

The proposed model achieved high accuracy percentage for classification. In this way, it should be enhance prediction accuracy in drug discovery.

The regression result for the proposed hybrid model, final error of the Root Mean Square Error (RMSE) is calculated among the truth drug and the predicted drug are predicted through the metric of Coefficient of determination R^2. The result of regression comparison is given in tabular form of Table 2.

Table.2. Comparison Table for Accuracy of Regression

Method	R^2	RMSE
Lasso	0.59	1.956
CNN	0.83	1.679
CNN-XGB	0.95	1.022

The lower value of RMSE is used to enhance the performance of accuracy of prediction in certain level. As shown in the Table 2, the Coefficient of determination R^2 is enhanced and also Root Mean Square Error is reduced in a precisely manner in our proposed model and also prediction accuracy in high level is achieved. The graphical representation is given in Fig. 5 for comparison of accuracy of regression.

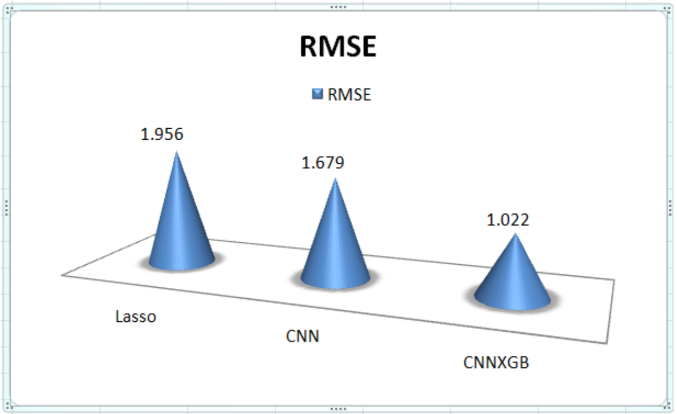

Fig.7. Graphical Analysis of Regression Accuracy

From Fig.6 and Fig.7, we can understand that our proposed model provides better accuracy in terms of regression and classification than other existing methods. It tends to give the confidence to predict the useful sensitive drug of IC50 with low cost, time and experiments in active drug discovery in outstanding way.

VI. CONCLUSION

This research presents an Ensemble Neural Network based active learning framework to enhance the prediction accuracy to select a more sensitive drug. In drug discovery, an accurate prediction is needed for creating new target cell line. In the drug discovery process, the time taken is plentiful with the numerous experiments being conducted and finally selecting the sensitive drug with low value is a major challenge. To overcome this disadvantage, a Hybrid CNN-XGB model to achieve high prediction accuracy through feature extraction, classification and regression. First, the SMILE datasets and Cell Mutation are given as an input. Hence, binary conversion is achieved for process the SMILES dataset. For cell lines, cell mutation states are converted into one dimensional Boolean vector or binary vector. Then, the converted one hot format encoding is processed and each CNN is utilized to pick out the features from drug and cell lines. Further, the classification is done using the selected features and testing data. Here, drugs sensitivity is predicted and classified by using XGBoost Classifier. Finally, the performance of the proposed hybrid model is compared with existing model such as Neural and Lasso for the tasks of regression and classification.

For comparison of results attained by the proposed model , certain methods proposed by various researchers have been referenced as shown in Table 3.

Table 3..Datasets used for Comparison of Results

Algorithms	**Accuracy %**
CNN	97.6
SVM	97.16
GoogleNet	99.83
XGBoost	99.56
CNN-XGBoost(proposed)	99.84

The comparison result shows proposed model outperforms counterparts well in terms of RMSE and classification accuracy. In future enhancement, this work can focus on more sensitive drugs datasets and also we can enhance regression performance by using recurrent neural network. In addition, we can use any other feature extraction methods to minimize the redundant features.

In the Hybrid model, the CNN is employed as a trainable automatic feature extractor. The CNN extracts features from raw data. This is followed by the feeding of these features to XGBoost, that performs the recognition of the needed details. The aim of the research was to improve and provide a better aid in drug discovery. The criteria employed for this was the recognition accuracy and how much reliable the performance would be .Using the hybrid model enhanced the automatic extraction of features from raw data, a combination of qualities of both CNN and XGBoost algorithm were merged in the hybrid and promising accuracy was achieved. Moreover, the proposed model helped to overcome the limitations of CNN and XGBoost individually with a combined hybrid.

A lot of research is currently being done on the hybrid CNN-XGBoost model. Even though the model has been successful in providing better aid in drug discovery, fine-tuning the structure and the associated parameters could improve the performance. The kernel functions along with the size of input layers could be modified to attain a better result.

References

[1] Stephenson, Natalie, et al. "Survey of machine learning techniques in drug discovery." Current drug metabolism 20.3 (2019): 185-193.

[2] Van Norman, Gail A. "Drugs, devices, and the FDA: part 1: an overview of approval processes for drugs." JACC: Basic to Translational Science 1.3 (2016): 170-179.

[3] T Issa, Naiem, et al. "Drug metabolism in preclinical drug development: a survey of the discovery process, toxicology, and computational tools." Current drug metabolism 18.6 (2017): 556-565.

[4] Ma, Junshui, et al. "Deep neural nets as a method for quantitative structure–activity relationships." Journal of chemical information and modeling 55.2 (2015): 263-274.

[5] Rifaioglu, Ahmet Sureyya, et al. "Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases." Briefings in bioinformatics 20.5 (2019): 1878-1912.

[6] Wen, Ming, et al. "Deep-learning-based drug–target interaction prediction." Journal of proteome research 16.4 (2017): 1401-1409.

[7] Hodos, Rachel A., et al. "In silico methods for drug repurposing and pharmacology." Wiley Interdisciplinary Reviews: Systems Biology and Medicine 8.3 (2016): 186-210.

[8] Lamberti, Mary Jo, et al. "A study on the application and use of artificial intelligence to support drug development." Clinical therapeutics 41.8 (2019): 1414-1426.

[9] Lavecchia, Antonio. "Deep learning in drug discovery: opportunities, challenges and future prospects." Drug discovery today 24.10 (2019): 2017-2032.

[10] Jing, Yankang, et al. "Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era." The AAPS journal 20.3 (2018): 58.

[11] Vamathevan, Jessica, et al. "Applications of machine learning in drug discovery and development." Nature Reviews Drug Discovery 18.6 (2019): 463-477.

[12] Sverchkov, Yuriy, and Mark Craven. "A review of active learning approaches to experimental design for uncovering biological networks." PLoS computational biology 13.6 (2017): e1005466.

[13] Lang, Tobias, et al. "Feasibility of active machine learning for multiclass compound classification." Journal of chemical information and modeling 56.1 (2016): 12-20.

[14] Reker, Daniel, et al. "Active learning for computational chemogenomics." Future medicinal chemistry 9.4 (2017): 381-402.

[15] Ma, Junshui, et al. "Deep neural nets as a method for quantitative structure–activity relationships." Journal of chemical information and modeling 55.2 (2015): 263-274.

[16] Temerinac-Ott, Maja, Armaghan W. Naik, and Robert F. Murphy. "Deciding when to stop: efficient experimentation to learn to predict drug-target interactions." BMC bioinformatics 16.1 (2015): 213.

[17] Kangas, Joshua D., Armaghan W. Naik, and Robert F. Murphy. "Efficient discovery of responses of proteins to compounds using active learning." BMC bioinformatics 15.1 (2014): 1 11.

[18] Copur, Mert, Buse Melis Ozyildirim, and Turgay Ibrikci. "Image Classification of Aerial Images Using CNN-SVM." 2018 Innovations in Intelligent Systems and Applications Conference (ASYU). IEEE, 2018.

[19] You, Jiaying, Robert D. McLeod, and Pingzhao Hu. "Predicting drug-target interaction network using deep learning model." Computational Biology and Chemistry 80 (2019): 90-101.

[20] Liu, Pengfei, and Kwong-Sak Leung. "Accelerating Drug Discovery Using Convolution Neural Network Based Active Learning." TENCON 2018-2018 IEEE Region 10 Conference. IEEE, 2018.

[21] Kangas, Joshua D., Armaghan W. Naik, and Robert F. Murphy. "Efficient discovery of responses of proteins to compounds using active learning." BMC bioinformatics 15.1 (2014): 1-11.

[22] Fu, Lei, et al. "Convolution Neural Network with Active Learning for Information Extraction of Enterprise Announcements." CCF International Conference on Natural Language Processing and Chinese Computing. Springer, Cham, 2018.

[23] Karpathy, Andrej. "Cs231n convolutional neural networks for visual recognition." Neural networks 1.1 (2016).

[24] An, Tae-Ki, and Moon-Hyun Kim. "A new diverse AdaBoost classifier." 2010 International Conference on Artificial Intelligence and Computational Intelligence. Vol. 1. IEEE, 2010.

[25] Ramraj, S., et al. "Experimenting XGBoost algorithm for prediction and classification of different datasets." International Journal of Control Theory and Applications 9 (2016): 651-662.

[26] Wang, Chen, Chengyuan Deng, and Suzhen Wang. "Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost." Pattern Recognition Letters (2020).

[27] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[28] Elleuch, Mohamed, Rania Maalej, and Monji Kherallah. "A new design based-SVM of the CNN classifier architecture with dropout for offline Arabic handwritten recognition." Procedia Computer Science 80 (2016): 1712-1723.

[29] Chen, Tianqi, et al. "Xgboost: extreme gradient boosting." R package version 0.4-2 (2015): 1-4.

[30] Paisit Khanarsa, Arthorn Luangsodsa, Krung Sinapiromsaran, Self-Identification ResNet-ARIMA Forecasting Model, WSEAS Transactions on Systems and Control, ISSN / E-ISSN: 1991-8763 / 2224-2856, Volume 15, 2020, Art. #21, pp. 196-211

[31] Halefom Tekle Weldegebriel a , Han Liu b (Member IEEE), Anwar Ul Haq a , Emmanuel Bugingo a , and Defu Zhang, A New Hybrid Convolutional Neural Network and eXtreme Gradient Boosting Classifier for Recognizing Handwritten Ethiopian Characters,IEEE Access,Vol xx,2019

[32] Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer & Shanrong Zhao, Applications of machine learning in drug discovery and development, Nature Reviews Drug Discovery,Vol 18, pages463–477 (2019).

Mr.Mukesh Madanan is a Senior Lecturer of Computer Science at Dhofar University,Oman. He completed his Bachelors in Technology in Computer Science and Engineering from M.G.University, India and went on to complete his M.Sc.in Software Engineering from the University of Portsmouth,UK. He is currently pursuing PhD in Information & Communication Technology from UNITEN, Malaysia. His areas of research include Machine Learning, Deep Learning, Robotics, Software Methodologies, IoT and Computer Networks.

Dr. Biju Theruvil Sayed is a higher educational academician with a total experience of 25+ years in the field of Computer

Science/Computer Engineering/Management Information Systems/Information Technology. As of the current date, he has been involved in teaching and learning processes for an approximate of 12000+ candidates. He is currently associated as a Chairperson and Assistant Professor in Computer Science at Dhofar University in Salalah, Oman. In addition, he is a Certified Training Professional and his current research area and interests are in Education Management, Education Assessment, Knowledge Management Systems, Expert Systems.

Dr.Nurul Akhmal Mohd Zulkefli is an Assistant Professor at Department of Computer Science, College of Arts & Applied Sciences, Dhofar University, Oman. She received he PhD in Information Technology(Expert Decision System) from Universiti Teknologi PETRONAS, Malaysia in 2019. Her research interests are expert decision systems, Big Data, Knowledge Management system, Software Engineering and IT in Education

Mrs.Nitha C.Velayudhan completed B.Tech in Information Technology and M.E in computer science and engineering. She has more than 14 years of teaching experience as lecturer and Assistant professor in field of computer science and engineering at various engineering colleges inside and outside Kerala .Currently pursuing PhD from Noorul Islam Centre for Higher Education Kumaracoil, Thuckalay, Kanyakumari, and Tamil Nadu, India. Her current research focus on elimination of sybil attack in vehicular ad-hoc networks operating in an urban environment. She had published papers and attended conference as the part of elimination of sybil attack in vehicular ad-hoc networks operating in an urban environment.

Contribution of individual authors to the creation of a scientific article

Mukesh Madanan carried out the background study and literature review and the analysis of the proposed research. He evaluated the proposed model and he was responsible for the Simulation and Statistics.
Biju.T.Sayed was responsible for coordinating and the verifying the simulation results of the proposed model.
Nurul Akhmal has implemented the CNN-XGB model.
Nitha C.Velayudhan has organized and executed the experiments.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)