Prediction of Drug-likeness of Central Nervous System Drug Candidates Using a Feed-Forward Neural Network Based on Chemical Structure

Yi-Gao Yuan, Xiao Wang*

School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, 210023, Jiangsu, China

Email: wangxiao@nju.edu.cn

Abstract: Modern medical science has been greatly advanced by the development of new drugs, despite the fact that the process of developing new drugs is costly and time-consuming. An accurate prediction method for the drug-likeness in the early stage of drug discovery is highly desirable, as it will facilitate the discovery process and reduce the overall cost and eventually contribute to human well-being. Based on a central nervous system (CNS) drug dataset, we constructed an artificial neural network (NN) to predict the CNS drug-likeness of a given bioactive compound. We first constructed a simple feed-forward neural network, to learn and predict the possible correlations between twelve physiochemical properties and the CNS drug-likeness. The accuracy of prediction has reached 80%, which has been improved from previous reports. We further constructed a neural network based on chemical structure, and the accuracy has increased to 86%. The successful prediction of the CNS drug-likeness renders this NN a powerful tool for virtual drug screening.

KEYWORDS: Central Nervous System, Drug-likeness, Drug Screening, Neural Network, Artificial Intelligence

Introduction:

Modern medicine has advanced rapidly that many drugs are approved and marketed every year. However, the cost of developing a new drug is still prodigious. Reducing the cost in the early stage of drug discovery is important for reducing the overall cost. Traditional guides to assist the hit and lead identification include the classical Lipinski's Rule of Five [1], and recently scientists have been using an increasing number of computational methods for such prediction. In specific, when targeting the central nervous system (CNS) drugs, multiparameter optimization methods have been developed, which include CNS multiparameter optimization (MPO)[2] and probabilistic MPO method [3]. In a broader sense where all potential drugs are considered, quantitative estimation of drug-likeness (QED) was proposed by Hopkins et al. [4] to assess whether a compound is likely to be druggable or not. The general form of these approaches is formatted as follow:

$$score = \Sigma w_k f(x_k);$$

where x_k is the physicochemical properties, f(x) is the function that maps x_k into a weightless score, w_k is the weight factor which measures the importance of each physicochemical property. These multiparameter optimization methods are moderately accurate, with an accuracy of approximately 70%. Therefore, there is still much space for improvement. In this report, we present a new prediction method for the CNS drug-likeliness by the implementation of artificial neural network, with which the accuracy can be further improved.

The artificial neural networks (ANN) are widely used in function approximation, classification and data processing. A typical ANN has three components: architecture, activation function, and learning rule. The common types of ANN include feed-forward networks, recurrent networks, and reinforcement networks. In this work, we will demonstrate the implementation of a feed-forward network (FFN) [5] to improve the accuracy of assessing the drug-likeness of a chemical compound. An FFN consists of an input layer, a certain number of hidden layers, and an output layer. The input layer receives values of the independent variables. The hidden layers process the input information based on an activation function and pass on the values to the next layer. The output layer receives values from the hidden layers and gives an output. The learning process of FNN is achieved by supervised learning using a training sample set and backpropagation-gradient descendent algorithm. The learning process is monitored by a loss function, which represents the deviation from the true answer. A dataset from the work by Gunaydin [3] was selected to be reanalyzed, which contains 665 marketed orally available drugs with 14 corresponding physicochemical properties. In the dataset, 299 drugs were reported to be capable of penetrating the brain-blood barrier which is denoted as “CNS drugs”. To simplify the analysis, we herein redefine the CNS drug-likeness as the ability of crossing the brain-blood barrier.

Methods:

The FFN for this study consisted of an input layer with 128 neurons and a rectified linear unit (ReLU) activation function. The second layer had 32 neurons, and the third layer had 16 neurons. The output layer used sigmoid as the activation function. Between each of the two layers, a dropout layer with a factor of 0.5 was added to prevent overfitting. The supervised learning used Adam optimizer to ensure minimal loss. The training dataset was adopted from Gunaydin’s approach [3], tabulated and imported. Further data processing, including cleaning null values and normalization, was performed before the data was fed into the neuron network.

```python
def cns_nn_simple():
    model = keras.Sequential([
                keras.layers.Dense(128,activation=tf.nn.relu, input_shape=[len(normed_train_data.keys())]),
                keras.layers.Dropout(0.5),
                keras.layers.Dense(32,activation=tf.nn.relu),
                keras.layers.Dropout(0.5),
                keras.layers.Dense(16,activation=tf.nn.relu),
                keras.layers.Dropout(0.5),
                keras.layers.Dense(1,activation=tf.nn.sigmoid)
            ])


    model.compile(
                    optimizer='adam',
                    loss='binary_crossentropy',
                    metrics=['accuracy','binary_crossentropy']
                    )

    #model.summary()

    return model
```

Figure 1. Python code for constructing the simple feed-forward network using TensorFlow.

Results and Discussion:

The supervised learning process was monitored by the loss and cross-entropy. To verify that overfitting was not happening, we used a validation dataset which was completely different from the training dataset. Figure 2 shows how the loss and cross-entropy changed as the supervised learning proceeded. The output layer of FFN employed a sigmoid activation function, which gave a value between 0 and 1. We took the output value and compared it with the cutoff value to define whether a compound can be considered as a CNS drug or not.

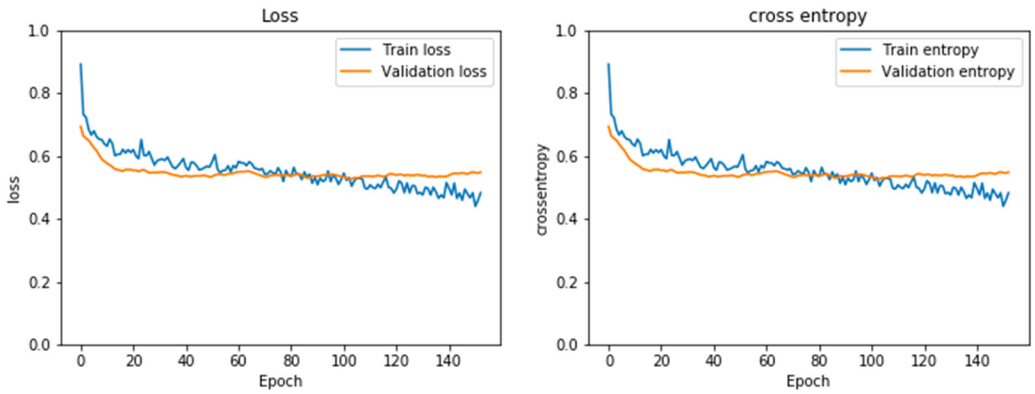

Figure 2. The loss of the training dataset and the validation dataset as a function of learning iteration. The cross-entropy is shown on the right. Both diagrams show that the neuron network is not overfitting because the training dataset has a similar loss compared to the validation dataset.

To visualize the capability of the FFN model, confusion matrixes and receiver operating curves (ROC) were utilized. Four quadrants, which represented the true-positive (TP), false-positive (FP), true-negative (TN) and false-negative (FN) predictions, were quantified in the confusion matrix. The ROC curve adopted the false-positive rate (FPR) as the x-axis and the true-positive rate (TPR) as the y-axis. The perfect model should be at (0,1), which means no false prediction since FPR was defined as FP/(FP+TN). Figure 3 shows the confusion matrix and the ROC for the simple FFN model. The true positive set, also known as the hit set, had 217 drugs. And the true negative set, also known as the correct rejection set, had 317 drugs. This gave the accuracy of the simple FFN model as (TP+TN)/Total = 80.5%. Type I error, the false positive set, had 49 drugs. And Type II error, the false negative set, had 80 drugs. Therefore, the true-positive rate (TPR) = TP/(TP+FN) = 73.1%, and the false-positive rate (FPR) = FP/(FP+TN) = 13.4%. Based on the high accuracy, the high TPR and the low FPR, it can be concluded that the simple FFN model gave a satisfactory performance, with enhanced accuracy from previous models. The ROC curve in Figure 3B gives a clear illustration of where the cutoff value should be, which is marked with a green dot.

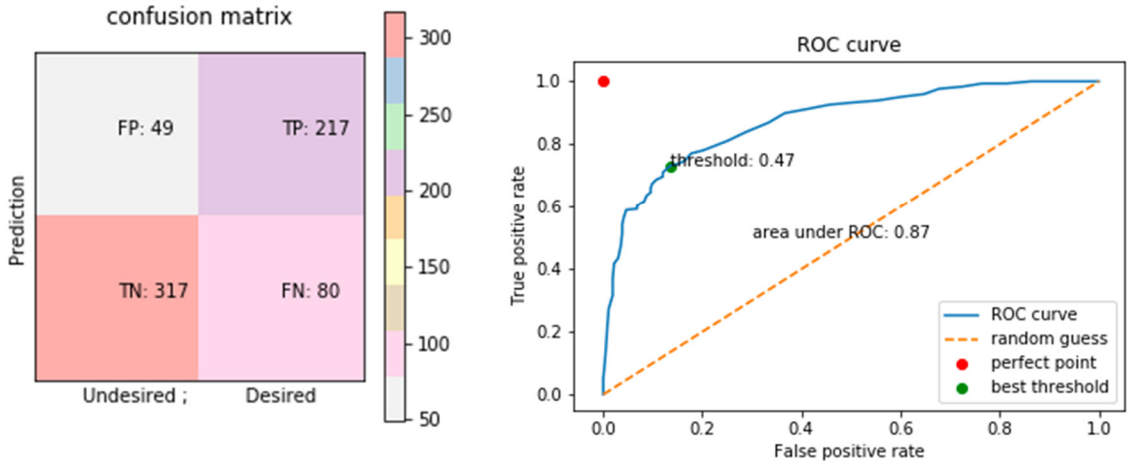

Figure 3. (A) Confusion matrix showing a prediction accuracy of (TP+TN)/Total = 80.5%. (B) Receiver operating curve showing that the area under ROC is 0.87, and the best threshold is 0.47.

To further evaluate the applicability of the FFN model, we took advantage of an open-source chemical informatics python package called RDKit [6], which is capable of generating the physicochemical properties of any molecule based on its chemical formula, the SMILES [7]. Using the RDkit package, we generated 45 different physicochemical properties of each inputted SMILES, and then fed the values into the input layer of the FFN constructed previously. The loss function and cross-entropy performance are shown in Figure 4. Both diagrams demonstrate that the neuron network was not overfitting because the training dataset had

a similar loss compared to the validation dataset. Provided that the FFN model did not overfit, the prediction based on the new FFN model could be even more accurate than the initial FFN model.

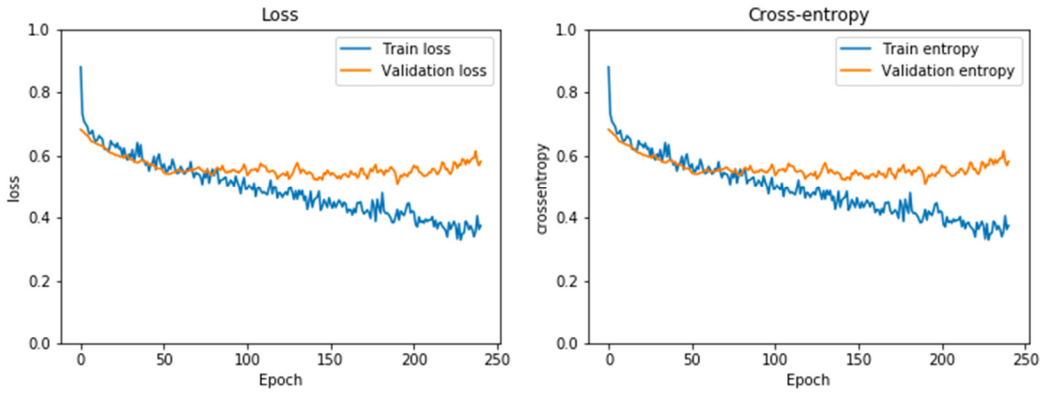

Figure 4. The loss of the training dataset and the validation dataset as a function of learning iteration. The cross-entropy is shown on the right.

Figure 5 shows the confusion matrix and ROC curve for the new FFN model. From the confusion matrix, TP had 260 drugs, and TN had 313 drugs. The FP and FN had 53 and 39 drugs, respectively. Gratifyingly, the accuracy reached 86.4%, and the TPR reached 87% while FPR was 14.5%. The ROC curve had an optimal cutoff value of 0.43, and the area under ROC reached 0.91. Therefore, this new FFN model exhibited better performance as compared to the initial FFN model.

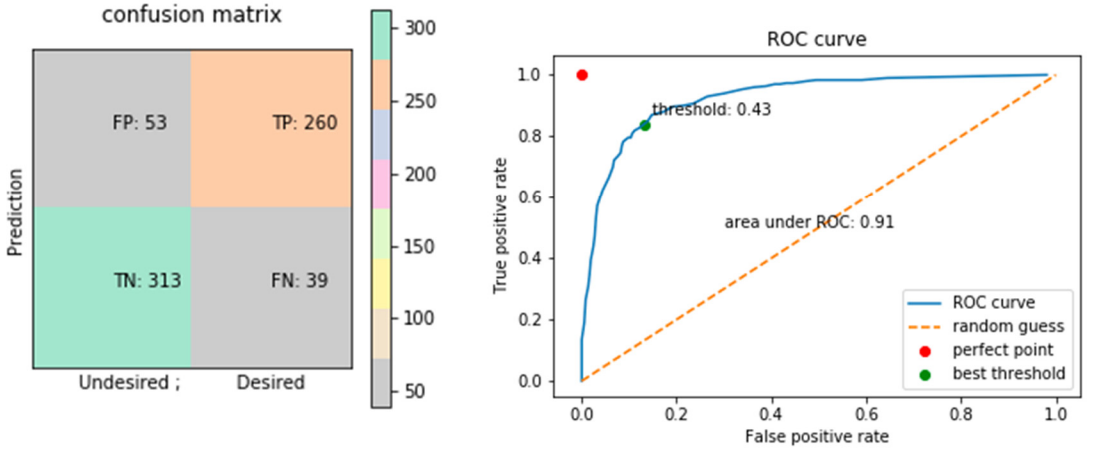

Figure 5. (A) Confusion matrix showing a prediction accuracy of (TP+TN)/Total = 86.4%. (B) Receiver operating curve showing that the area under ROC is 0.91, and the best threshold is 0.43.

It is worth noting that achieving a high accuracy does not eliminate false prediction. Six representative drugs with false-positive classification are shown in Figure 6. Estradiol, for example, is a steroid hormone and proven to be able to penetrate the brain-blood barrier. However, the dataset placed it in the wrong category as a false-positive. For other false predictions, we believe that they mainly derived from the error of the method.

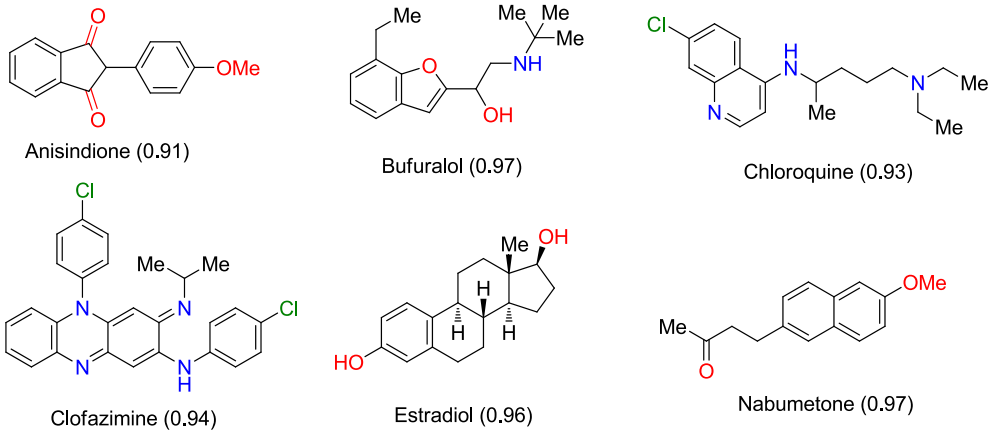

Figure 6. Examples of drugs with false-positive predictions. The corresponding prediction score is in the parenthesis.

Conclusion

In this work, feed-forward networks were implemented to predict the CNS-drug-likeness. The accuracy has reached 80% using a simple FFN and further improved to 86% by involving more physicochemical properties, which were more satisfactory than previous reports. Although the current size of the dataset is too limited to release the full potential of neural networks in processing large datasets, the primary objective of verifying the possibility to utilize neural networks in predicting CNS drug-likeness of an organic small molecule, has been achieved. We hope that these methods can serve as an applicable set of protocols for virtual drug screening. The investigation of larger datasets of CNS and other drugs using this FFN method is currently on-going in our lab, in the hope of realizing a broader applicability of NN for virtual CNS drug screening in the early stage of drug discovery.

References:

[1] LIPINSKI C A, LOMBARDO F, DOMINY B W, et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings 1[J]. Advanced Drug Delivery Reviews, 2012, 64(1– 3): 4–17.
[2] WAGER T T, XINJUN H, VERHOEST P R, et al. Moving beyond rules: the development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of druglike properties[J]. ACS Chemical Neuroscience, 2010, 1(6) : 435.
[3] GUNAYDIN H. Probabilistic Approach to Generating MPOs and Its Application as a Scoring Function for CNS Drugs. [J]. ACS Medicinal Chemistry Letters, 2015, 7(1) : acsmedchemlett.5b00390.
[4] BICKERTON G R, PAOLINI G V, BESNARD J, et al. Quantifying the chemical beauty of drugs [J]. Nature Chemistry, 2012, 4(2) : 90 – 98.
[5] HORNIK K, STINCHCOMBE M, WHITE H. Multilayer feedforward networks are universal approximators [J]. Neural networks, 1989, 2(5) : 359 – 366.
[6] RDKit: Open-source cheminformatics; http://www.rdkit.org
[7] WEININGERD.SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules [J]. Journal of chemical information and computer sciences, 1988, 28(1): 31 – 36.