Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2022;14(2):70-80

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH DOI: 10.5336/biostatic.2022-88704

Activity Classification of Small Drug Molecules Using Deep Neural Networks and Classical Machine Learning Models

Derin Sinir Ağları ve Klasik Makine Öğrenimi Modelleri Kullanılarak Küçük İlaç Moleküllerinin Aktivite Sınıflandırması

Hatice KANBERİZ[a], Selçuk KORKMAZ[a], Necdet SÜT[a]

[a]Department of Biostatistics and Medical Informatics, Trakya University Faculty of Medicine, Edirne, Türkiye

ABSTRACT Objective: The main goal in the early phase of drug discovery studies is to detect small drug molecules that show activity against a specific receptor. For this purpose, small drug molecules are classified as actives or inactives by performing high-throughput screening (HTS) experiments. The datasets obtained from these experiments are uploaded to the PubChem database. This database contains more than one million bioassays that are obtained through HTS experiments. Alternatively, classification models can be developed using datasets in the PubChem database. **Material and Methods:** In this study, we obtained 5 datasets with different degrees of imbalance structure from the PubChem database. We trained these datasets using deep neural networks (DNN) for the classification of small drug molecules as actives or inactives. The test set performances of DNN models were compared with the support vector machines (SVM) and random forest (RF) algorithms. **Results:** The DNN achieved better balanced accuracy (minimum-maximum: 0.764-0.865), recall (minimum-maximum: 0.630-0.823), F1-score (minimum-maximum: 0.496-0.843) and Matthews correlation coefficient (minimum-maximum: 0.439-0.721) compared to the SVM and RF. **Conclusion:** Our results showed that the DNN is a well-performed machine learning algorithm that can be in the early phase of drug discovery studies since it performs better than traditional machine learning algorithms in the case of imbalanced class structures.

Keywords: Deep learning; imbalanced classes; support vector machines; random forest; virtual screening

ÖZET Amaç: İlaç keşif çalışmalarının erken evresindeki temel amaç, belirli bir reseptöre karşı aktivite gösteren küçük ilaç moleküllerini tespit etmektir. Bu amaçla küçük ilaç molekülleri, yüksek verimli tarama [high-throughput screening (HTS)] deneyleri gerçekleştirilerek aktif veya inaktif olarak sınıflandırılır. Bu deneylerden elde edilen veri setleri PubChem veri tabanına yüklenir. Bu veri tabanı, HTS deneyleri yoluyla elde edilen 1 milyondan fazla biyo-tahlil veri setini içerir. Alternatif olarak, PubChem veri tabanındaki veri kümeleri kullanılarak sınıflandırma modelleri de geliştirilebilir. **Gereç ve Yöntemler:** Bu çalışmada, PubChem veri tabanından farklı derecelerde dengesizlik yapısına sahip 5 adet veri seti elde ettik. Bu veri setlerini, küçük ilaç moleküllerinin aktif veya inaktif olarak sınıflandırılması için derin sinir ağları [deep neural networks (DNN)] kullanarak eğittik. DNN modellerinin test seti performansları, destek vektör makineleri [support vector machines (SVM)] ve rastgele orman [random forest (RF)] algoritmaları ile karşılaştırılmıştır. **Bulgular:** DNN modeli, dengeli doğruluk oranı (en küçük-en büyük: 0.764-0.865), duyarlılık (en küçük-en büyük: 0.630-0.823), F1-skoru (en küçük-en büyük: 0.496-0.843) ve Matthews korelasyon katsayısı (en küçük-en büyük: 0.439-0.721) açısından SVM ve RF'den daha iyi performans göstermiştir. **Sonuç:** Sonuçlarımız DNN'nin dengesiz sınıf yapıları durumunda, klasik makine öğrenimi algoritmalarından daha iyi performans gösterdiğini, bu nedenle ilaç keşif çalışmalarının erken aşamasında iyi performans gösterebilen bir makine öğrenimi algoritması olduğunu ortaya koymuştur.

Anahtar kelimeler: Derin öğrenme; dengesiz sınıflar; destek vektör makineleri; rastgele orman; sanal tarama

Correspondence: Selçuk KORKMAZ
Department of Biostatistics and Medical Informatics, Trakya University Faculty of Medicine, Edirne, Türkiye
E-mail: selcukorkmaz@gmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 04 Feb 2022 ***Received in revised form:*** 01 Jun 2022 ***Accepted:*** 06 Jun 2022 ***Available online:*** 08 Jun 2022

The new drug development is both a costly and time-consuming process. A new drug development study may take over 15 years and maybe spent over a billion dollars. Fortunately, the time and cost of new drug development studies considerably decreased with the rational drug design approach. This approach aims to find the 3-dimensional structure of a receptor (protein) that is thought to cause disease and reveal its active region. Thus, the drug molecule to be synthesized will have the highest chance to bind to this active region of the receptor. Before starting a drug discovery study, it is necessary to screen a huge number of chemical molecules. For this purpose, a high-throughput screening (HTS) experiment is performed to identify the activities of chemical molecules against a particular receptor or enzyme.[1] The molecules that show activity as a result of the HTS experiment are called lead compounds. These lead compounds are then optimized and passed on to pre-clinical trials (i.e., animal experiments). After the pre-clinical trials are successfully completed, the clinical trials phase begins. These trials consist of 3 stages: Phase I, Phase II, and Phase III. Upon successfully completing Phase III, the new drug is submitted for the approval of the regulatory agency (such as Food and Drug Administration or *European Medicines Agency*). Finally, the approved new drug is ready to be placed on the market.

Alternatively, a computational method, called virtual screening (VS), can be used to screen drug molecules. In the VS, a large number of molecules in chemical libraries are computationally screened against a specific receptor or enzyme, and molecules that are predicted to well bind to this specific receptor or enzyme are experimentally tested.[2] In the last 2 decades, various machine learning methods have been extensively used for classification or activity prediction purposes. One of the first studies where machine learning was used in drug discovery was carried out by Sadowski and Kubinyi.[3] In this study, artificial neural networks (ANN) were used for the classification of drug molecules. Other studies compared performances of support vector machines (SVM) and ANN algorithms, and revealed that SVM outperformed ANN.[4,5] Korkmaz et al. investigated the effects of different feature selection methods on the SVM performance.[6] In another study, Korkmaz et al. used 23 different machine learning algorithms and developed a web-tool to classify drug molecules using the ten best-performed methods.[7] Other methods including, random forest (RF), naive Bayes (NB), Bayesian neural networks and k-nearest neighbors (kNN) are also used to classify active and inactive molecules in drug discovery studies.[8,9]

Recently, deep neural networks (DNN) have performed well in drug discovery studies. A DNN model is used to estimate quantitative structure-activity relationships and performed better than the RF model.[10] In another study, the DNN model was used as a multitask learning approach to estimate compound toxicity.[11] Similarly, multitask learning using the DNN is applied to numerous data sets [PubChem BioAssay (PCBA), maximum unbiased validation (MUV), A Database of Useful Decoys: Enhanced (DUD-E), The Toxicology in the 21st Century (Tox21)].[12] Other studies focused on comparing the performance of deep learning models with other classical machine learning algorithms, including SVM, RF, NB, and kNN.[13,14] Korkmaz explored the factors affecting the DNN performance and found that the learning rate, batch size, and level of imbalance significantly affect the performance of the model.[15] In another study, Korkmaz found that balancing the classes prior to the training increased the performance of the DNN.[16]

Since VS is a computational filter that reduces the size of the chemical library to be experimentally screened, it can reduce the time and cost of finding lead compounds compared to the HTS method. Today, experimentally obtained data on drug molecules using the HTS method are uploaded to freely accessible databases. One of the largest databases containing drug molecules is the PubChem database. However, the bioassays in the PubChem, which were obtained by the HTS experiment, are usually imbalanced. That means the number of inactive molecules is significantly larger than the number of active molecules. It is well known that this imbalanced nature of datasets negatively affects the classification performance of traditional machine learning algorithms.

Here, we obtained 5 different data sets with different levels of imbalance structure from the PubChem database. Traditional machine learning methods, including SVM and RF, are frequently used in the literature

to classify small drug molecules. However, these algorithms cannot perform well in imbalanced data structures. Recently, DNN has performed well in many areas and outperformed traditional machine learning methods such as SVM and RF. In this study, we aimed to train the DNN model using imbalanced datasets obtained from the PubChem database and compare its test set performances with SVM and RF.

MATERIAL AND METHODS

DEEP NEURAL NETWORKS

DNN contain multiple nonlinear hidden layers between input and output layers.[17] Using these hidden layers and optimizable parameters, the DNN can learn high-dimensional relationships.[18] The training process of a DNN algorithm can be summarized as follows: (i) First, random values are used to set the weights. (ii) Then, the DNN predictions and true classes are compared to calculate the loss score. (iii) Next, the backpropagation algorithm is used to calculate the derivatives by applying the chain rule.[19,20] (iv) Using an optimization algorithm (i.e., gradient descent algorithm) and the resulting derivatives, the weights are modified slightly in the correct direction to reduce the loss. (v) Finally, the steps between ii and iv are iteratively repeated to minimize the loss score. As a result, a DNN model with the lowest loss score, which makes the predictions as close to true classes as possible, is trained.[21]

SUPPORT VECTOR MACHINES

SVM is a classification algorithm explicitly designed to solve the classification problem. The SVM has a very high generalization performance, and it can handle high-dimensional data.[6] The SVM is developed by Cortes and Vapnik based on statistical learning theory and the principle of minimizing structural risk.[22] The SVM algorithm aims to find a hyperplane that can optimally separate the distance between support vectors belonging to different classes.[22] Suppose the classification problem cannot be solved linearly. In that case, the SVM tries to find the optimal margin between classes in a high dimensional space by mapping the nonlinear sample space to a high dimension where the samples can be separated linearly.[6]

RANDOM FOREST

RF is a decision tree-based machine learning method developed by Breiman.[23] The RF is an ensemble learning method developed for both classification and regression. The RF can be used in high-dimensional complex data structures. In the RF, trees are created with the Classification and Regression Tree algorithm, and these trees are not pruned. During the creation of each tree, a random subset of variables is chosen among the input variables, and those with the best split results are selected. This algorithm uses the information gain or Gini index to decide which variables should be used to split the data. The trees are evaluated separately, and the final classification result is calculated by the majority vote of the estimates obtained by each three.

DATA SETS

We used 5 data sets that are obtained from the PubChem. The number of active, inactive, and toal molecules in each data set are given in Table 1.

TABLE 1: The characteristics of the bioassay data sets used in the study.

Data set	Number of active molecules	Number of inactive molecules	Total number of molecules	Active/inactive ratio
AID652178	178	897	1,075	1:5
AID1053187	420	1,172	1,592	1:3
AID1053196	231	2,058	2,289	1:9
AID1159608	624	637	1,261	1:1
AID115909	717	1,070	1,787	1:1.5

1) AID652178: This bioassay was created for a transmembrane domain receptor (GQ-linked G protein-coupled receptors M1 Muscarinic receptor), which significantly affects the treatment of Alzheimer's disease and schizophrenia-associated cognitive degeneration. This bioassay data set includes a total of 1,075 compounds (178 actives and 897 inactives).

2) AID1053187: This bioassay data were generated for the muscarinic M1 receptor and included results from different HTS experiments (AID628, AID677, AID859, AID860). This bioassay data set consists of 1,592 compounds (420 actives and 1,172 inactives).

3) AID1053196: This bioassay was developed for choline transporter inhibitors and included results from different HTS experiments (AID488975, AID493221, AID504840, AID588401, AID493222, AID602208, AID49322). This bioassay data set contains a total of 2,289 compounds (231 actives and 2,058 inactives).

4) AID1159608: This bioassay data was created for antagonists of the neuropeptide Y receptor Y2 and included results obtained from different HTS experiments (AID793, AID1257, AID1256, AID1279, AID1272, AID2210, AID2212, AID2224). This bioassay data set includes a total of 1,261 compounds (624 actives and 637 inactives).

5) AID115909: This bioassay data were generated for esters and lactones of phenolic amino carboxylic acids as prodrugs for iron chelation. This bioassay data set consists of 1,787 compounds (717 actives and 1,070 inactives).

PERFORMANCE MEASURES

The following performance measures are calculated to evaluate the test set performance of the trained machine learning algorithms.

Balanced accuracy: Traditional accuracy measure is not appropriate when the classes is not balanced. Balanced accuracy, on the other hand, is a useful measure when there is high imbalance between classes.

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (1)$$

Recall: This measure is the rate of actual positives among all positives.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

Precision: This measure is the rate of actual positives among observations classified as positives.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

F1-score: This measure is the harmonic mean of precision and recall. This measure is beneficial in the case of imbalanced classes.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Matthews correlation coefficient (MCC): MCC measures the correlation between observed and estimated classes. MCC value equals -1 means a total disagreement between observed, while MCC value equals +1 shows a perfect prediction. A value of 0, on the other hand, indicates the random estimation.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

where TP=true positives, TN=true negatives, FP=false positives, and FN=false negatives.

DATA PRE-PROCESSING AND MODEL BUILDING

It is necessary to calculate the molecular descriptors to analyze the bioassay acquired by the HTS method using machine learning algorithms. PaDEL is free and open-source software developed by Yap.[24] In this study, we computed 2,757 descriptors using the PaDEL software for each dataset. Then, we removed missing values, empty columns and zero variance variables. Subsequently, the number of variables was reduced to 1,348. Each data set is randomly divided into 2 parts as 80% training and 20% test set. Furthermore, 10% of the training set randomly selected as a validation set during the training procedure. The z-score transformation was applied to the training sets, and the test sets were standardized according to these. A 10 fold cross-validation was applied for parameter optimization of SVM and RF. For SVM, sigma and cost parameters are determined as 0.0004 and 4, and for RF, the number of randomly selected predictors is founded as 2, and the number of decision trees is founded as 500. For the DNN model, we used the rectified linear unit function as an activation function for input and hidden layers, and the sigmoid activation function for the output layer. Moreover, we applied batch normalization in each layer in order to improve the performance and stability of the network. We used a binary loss function and the Adam method for stochastic optimization. Parameters of the network were optimized using the loss score of the validation sets. We created a network with four hidden layers (the first layer: 1,024 nodes, the second layer: 2,048 nodes, the third layer: 1,500 nodes, and the fourth layer: 128 nodes). To avoid overfitting, we used 20% dropout rate. We used Python v-3.7.3 for DNN model training and R v-3.6.1 for SVM and RF. For the model building of SVM and RF, we used the wrapper package caret in R. The randomForest package is used to build the RF models, while kernlab and e1071 packages are used to build the SVM models. We used the keras library to build model for DNN, and the scikit-learn library to evaluate model performance metrics.

RESULTS

We used 5 HTS datasets to train DNN, SVM, and RF algorithms, and the performance of each algorithm was tested on the same test set. The performances of the algorithms were compared using balanced accuracy, precision, recall, MCC and F1-score.

Performance measures of DNN, SVM and RF for the AID652178 dataset were calculated (Table 2). This data set had an imbalanced class structure, where the number of inactives was approximately 5 times greater than the number of actives. The DNN outperformed SVM and RF (balanced accuracies were 0.767, 0.526, and 0.540, respectively). Similarly, the DNN achieved better performances regarding recall (0.686), MCC (0.464) and F1-score (0.558). However, the RF and SVM showed better performances than the DNN in terms of precision (0.750, 0.667, and 0.471, respectively).

TABLE 2: Test set performances of DNN, SVM and RF algorithms for AID652178.

Measures	DNN	SVM	RF
Balanced accuracy	0.767	0.526	0.540
Recall	0.686	0.057	0.086
Precision	0.471	0.667	0.750
F1 score	0.558	0.105	0.154
MCC	0.464	0.162	0.219

DNN: Deep neural networks; SVM: Support vector machines; RF: Random forest; MCC: Matthews correlation coefficient.

Performance measures of the DNN, SVM and RF algorithms for the AID1053187 dataset were calculated (Table 3). This data set also has an imbalanced class structure, where the number of inactives is approximately three times greater than the number of actives. The DNN algorithm outperformed the SVM

and RF regarding all measures. The DNN showed the best balanced accuracy with 0.865, best recall with 0.809, best precision with 0.782, best F1-score with 0.795, and best MCC with 0.721.

TABLE 3: Test set performances of DNN, SVM, and RF algorithms for AID1053187.

Measures	DNN	SVM	RF
Balanced accuracy	0.865	0.556	0.765
Recall	0.809	0.155	0.619
Precision	0.782	0.565	0.712
F1 score	0.795	0.243	0.663
MCC	0.721	0.191	0.555

DNN: Deep neural networks; SVM: Support vector machines; RF: Random forest; MCC: Matthews correlation coefficient.

Performance measures of the DNN, SVM, and RF algorithms for the AID1053196 dataset were calculated (Table 4). This was the most imbalanced bioassay used in the study. In this bioassay, the number of inactive molecules was approximately nine times greater than the number of active molecules. Once again, the DNN algorithm achieved better performances than the SVM and RF in all metrics, except precision. The DNN yielded the best balanced accuracy with 0.764, best recall with 0.630, best F1-score with 0.496, and best MCC with 0.439.

TABLE 4: Test set performances of DNN, SVM, and RF algorithms for AID1053196.

Measures	DNN	SVM	RF
Balanced accuracy	0.764	0.544	0.544
Recall	0.630	0.087	0.087
Precision	0.409	1.000	1.000
F1 score	0.496	0.160	0.160
MCC	0.439	0.281	0.281

DNN: Deep neural networks; SVM: Support vector machines; RF: Random forest; MCC: Matthews correlation coefficient.

Performance measures of the DNN, SVM and RF algorithms for the AID1159608 dataset were calculated (Table 5). This bioassay was the only dataset, which includes balanced classes. The DNN algorithm yielded better performances than the SVM and RF in terms of all performance measures in this balanced dataset. The DNN delivered the best balanced accuracy with 0.849, best recall with 0.823, best precision with 0.864, best F1-score with 0.843, and best MCC with 0.698.

TABLE 5: Test set performances of DNN, SVM, and RF algorithms for AID1159608.

Measures	DNN	SVM	RF
Balanced accuracy	0.849	0.625	0.645
Recall	0.823	0.565	0.621
Precision	0.864	0.637	0.647
F1 score	0.843	0.598	0.634
MCC	0.698	0.252	0.291

DNN: Deep neural networks; SVM: Support vector machines; RF: Random forest; MCC: Matthews correlation coefficient.

Performance measures of the DNN, SVM and RF algorithms for the AID1159609 dataset were calculated (Table 6). The number of inactive molecules is approximately 1.5 times the number of active

molecules in this bioassay. The DNN resulted in better performances than the SVM and RF regarding all metrics. The DNN gave the best balanced accuracy with 0.846, best recall with 0.805, best precision with 0.827, best F1-score with 0.816, and best MCC with 0.696.

TABLE 6: Test set performances of DNN, SVM, and RF algorithms for AID1159609.

Measures	DNN	SVM	RF
Balanced accuracy	0.846	0.566	0.575
Recall	0.805	0.259	0.266
Precision	0.827	0.578	0.603
F1 score	0.816	0.358	0.369
MCC	0.696	0.169	0.192

DNN: Deep neural networks; SVM: Support vector machines; RF: Random forest; MCC: Matthews correlation coefficient.

The performances of the algorithms were also compared by considering the level of imbalance. It was observed that the DNN was the best performing algorithm in all imbalance structures in terms of balanced accuracy. The performances of the SVM and RF algorithms were found similar except for AID1053187. The results for the balanced accuracy were given in Figure 1.

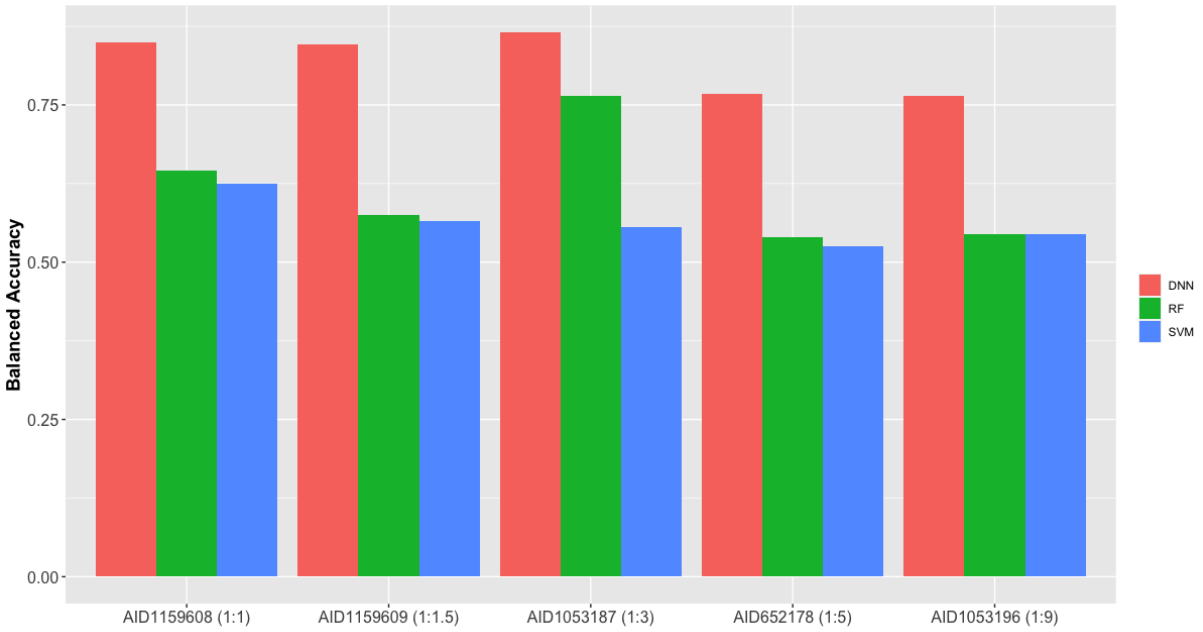

FIGURE 1: Comparison of DNN, SVM, and RF performances regarding balanced accuracy.

DNN: Deep neural networks; SVM: Support vector machines; RF: Random forest.

Similar outcomes were obtained for F1-score and MCC. The DNN was the best performing algorithm in all levels of imbalance in terms of F1 score and MCC (Figure 2 and Figure 3). The performances of the SVM and RF were found similar, except for the AID1053187, where RF performed better than SVM.

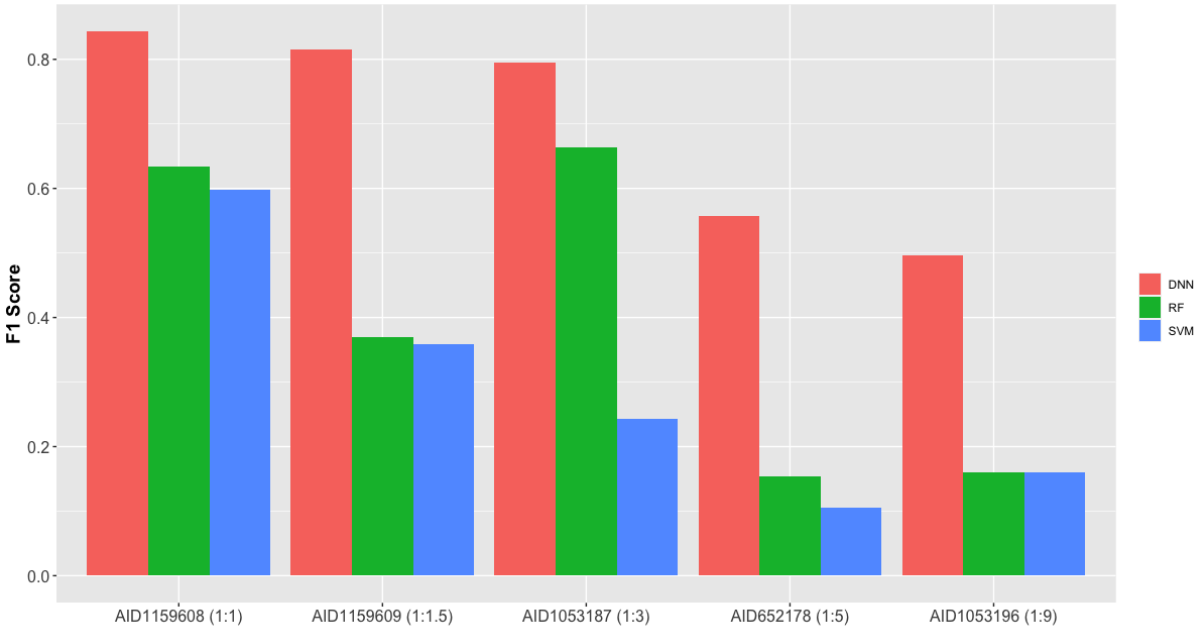

FIGURE 2: Comparison of DNN, SVM, and RF performances regarding F1-score.

DNN: Deep neural networks; SVM: Support vector machines; RF: Random forest.

Although the DNN showed the best performance in all imbalanced data structures regarding balanced accuracy, MCC, and F1 score, it is observed that the network performance disrupts as the level of imbalance increases.

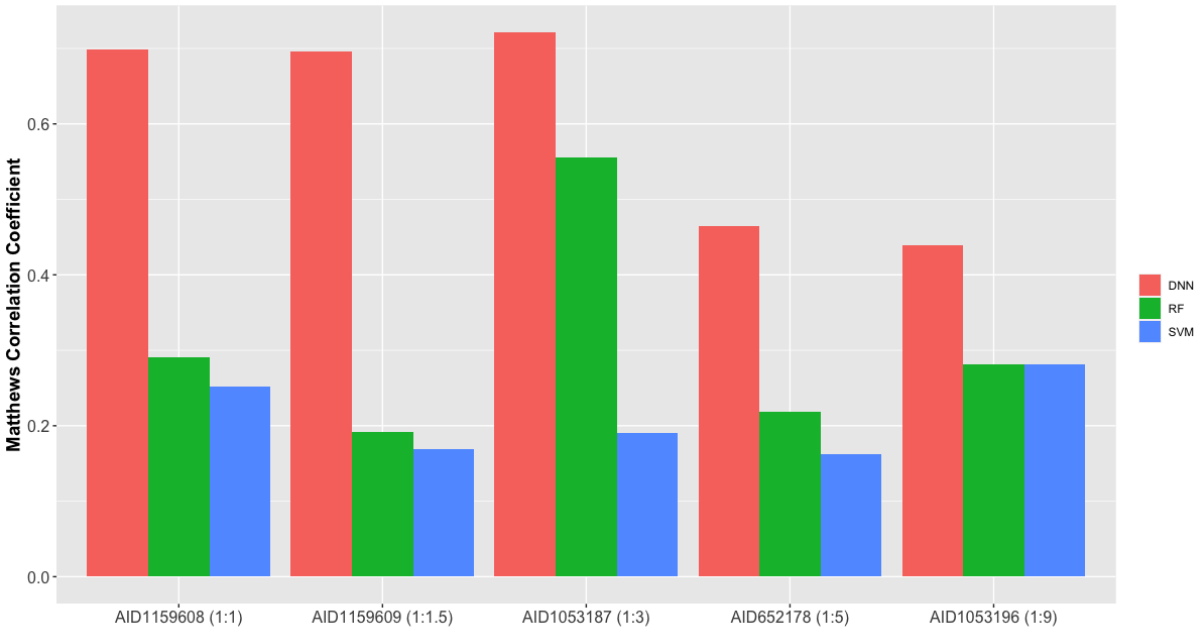

FIGURE 3: Comparison of DNN, SVM, and RF performances in terms of Matthews correlation coefficient.

DNN: Deep neural networks; SVM: Support vector machines; RF: Random forest.

DISCUSSION

Machine learning methods have been extensively investigated for the activity classification of small drug molecules. Sadowski and Kubinyi developed a classification model using the ANN algorithm with 169,331 non-drug molecules from the Available Chemicals Directory (ACD) and 38,416 drug molecules from the World Drug Index (WDI).[3] The accuracy rate of the ACD was 83% (non-drug molecules), and the accuracy rate of the WDI was 77% (drug molecules). Byvatov et al. compared the performances of SVM and ANN algorithms using 4,998 drugs and 4,210 non-drug molecules.[4] They found that the SVM (82% accuracy and 0.63 MCC) outperformed the ANN (80% accuracy and 0.58 MCC). In another study, Zernov et al. compared performances of SVM and ANN using 15,000 drug and 15,000 non-drug molecules.[5] Once again, SVM performed better than different ANN models (accuracy rate of the multi-layer sensor was 72.52%, the modular feed-forward network was 70.92%, and the generalized feed-forward network was 69.85%) with an accuracy of 75.15%. Korkmaz et al. aimed to distinguish between drug and non-drug compounds using the SVM with various feature selection methods.[6] They used 311 drug and 320 non-drug molecules for the training set, and 98 drug and 118 non-drug molecules for the test set. As a result, the test set's accuracy rates were between 76% and 81%. In another study, Korkmaz et al. investigated performances of 23 machine learning models using the same dataset.[7] They reported accuracy rate between 68% and 79%, sensitivity between 81% and 92%, positive predictive value between 60% and 72%, F1 score between 0.72 and 0.79, MCC between 0.42 and 0.59.

Recently, DNN has been used in drug classification studies and outperformed traditional machine learning methods. Ma et al. found that the DNN showed a better performance than the RF for the development of quantitative structure-activity relationships models.[10] Mayr et al. utilized DNN for the toxicity prediction.[11] Ramsundar et al. performed a multitask learning with the DNN using different publicly available datasets (PCBA, DUD-E, Tox21 and MUV).[12] Koutsoukas et al. used seven different bioactivity datasets (ChEMBL205, CHEMBL301, CHEMBL240, CHEMBL219, CHEMBL244, CHEMBL218) to compare the DNN with other classical machine learning algorithms.[13] They reported that the DNN outperformed all traditional machine learning methods. Another study conducted by Lenselink et al. showed that the DNN performed better than NB, RF, SVM, and logistic regression.[14]

In this study, 5 HTS bioassays with distinct levels of imbalance were used. Then, the performance of the DNN compared with classical machine learning algorithms, including SVM and RF. The balanced accuracy for the DNN was found between 0.764 and 0.865, whereas the balanced accuracy rates for the SVM and RF were found between 0.526-0.625 and 0.540-0.765, respectively. Similar results obtained for recall (0.630-0.823 for the DNN, 0.057-0.560 for the SVM and 0.086-0.619 for the RF), F1-score (0.496-0.843 for the DNN, 0.160-0.598 for the SVM and 0.160-0.663 for the RF), and MCC (0.439-0.721 for the DNN, 0.162-0.281 for the SVM and 0.192-0.555 for the RF). This study showed that the DNN performed better than the SVM and RF in terms of all performance measures, except precision in AID652178 and AID1053196. Even though the DNN performed better than the SVM and RF based on balanced accuracy, F1 score, and MCC in all datasets, it was observed that the performance of the DNN decreased as the level of class imbalance increased.

Although traditional machine learning algorithms broadly used in the drug classification problem, the datasets used in these studies were generally well balanced. However, in practice, it is well known that the number of active compounds is significantly smaller than the number of inactive compounds. Our results clearly showed that traditional machine learning algorithms (i.e., SVM and RF) were negatively affected by the class imbalance. The DNN, on the other hand, can mitigate this effect to a degree. Therefore, other methods, such as data balancing methods, should be incorporated into the DNN to overcome the class imbalance problem.

CONCLUSION

Drug development is a challenging, costly and time-consuming process. The rational drug design approach has been used to reduce the time and cost of drug development studies. Bioassays acquired by the HTS method are stored in an online database, called PubChem. Thus, machine learning methods can be trained, and high-performed models can be developed to detect active molecules using the PubChem. Although traditional machine learning algorithms have been used for drug development studies for a long time, the data sets used to train these algorithms are generally balanced. However, the actual datasets in the PubChem are imbalanced. The nature of these datasets (i.e, imbalanced classes) negatively affects the performance of classical machine learning algorithms. Recently, the deep learning has performed well in many areas, including drug discovery.

In this study, we trained DNN algorithms using 5 different imbalanced HTS datasets. We also compared the performance of the DNN with traditional machine learning algorithms, SVM and RF. We observed that the DNN outperformed the SVM and RF. The results of this study suggest that the PubChem database is a useful resource for training machine learning models. Moreover, the DNN is a well-performed algorithm that can be used to classify small drug compounds in an imbalanced class problem.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Selçuk Korkmaz, Hatice Kanberiz;* ***Design:*** *Selçuk Korkmaz, Hatice Kanberiz;* ***Control/Supervision:*** *Selçuk Korkmaz, Necdet Süt;* ***Data Collection and/or Processing:*** *Hatice Kanberiz, Selçuk Korkmaz;* ***Analysis and/or Interpretation:*** *Hatice Kanberiz, Selçuk Korkmaz;* ***Literature Review:*** *Hatice Kanberiz, Selçuk Korkmaz;* ***Writing the Article:*** *Hatice Kanberiz, Selçuk Korkmaz;* ***Critical Review:*** *Hatice Kanberiz, Selçuk Korkmaz, Necdet Süt;* ***References and Fundings:*** *Hatice Kanberiz, Selçuk Korkmaz, Necdet Süt.*

REFERENCES

1. Broach JR, Thorner J. High-throughput screening for drug discovery. Nature. 1996;384(6604 Suppl):14-6. [PubMed]
2. Shoichet BK. Virtual screening of chemical libraries. Nature. 2004;432(7019):862-5. [Crossref] [PubMed] [PMC]
3. Sadowski J, Kubinyi H. A scoring scheme for discriminating between drugs and nondrugs. J Med Chem. 1998;41(18):3325-9. [Crossref] [PubMed]
4. Byvatov E, Fechner U, Sadowski J, Schneider G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. J Chem Inf Comput Sci. 2003;43(6):1882-9. [Crossref] [PubMed]
5. Zernov VV, Balakin KV, Ivaschenko AA, Savchuk NP, Pletnev IV. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. J Chem Inf Comput Sci. 2003;43(6):2048-56. [Crossref] [PubMed]
6. Korkmaz S, Zararsiz G, Goksuluk D. Drug/nondrug classification using Support Vector Machines with various feature selection strategies. Comput Methods Programs Biomed. 2014;117(2):51-60. [Crossref] [PubMed]
7. Korkmaz S, Zararsiz G, Goksuluk D. MLViS: A web tool for machine learning-based virtual screening in early-phase of drug discovery and development. PLoS One. 2015;10(4):e0124600. [Crossref] [PubMed] [PMC]
8. Fang J, Yang R, Gao L, Zhou D, Yang S, Liu AL, et al. Predictions of BuChE inhibitors using support vector machine and naive Bayesian classification techniques in drug discovery. J Chem Inf Model. 2013;53(11):3009-20. [Crossref] [PubMed]
9. Ehrman TM, Barlow DJ, Hylands PJ. Virtual screening of Chinese herbs with random forest. J Chem Inf Model. 2007;47(2):264-78. [Crossref] [PubMed]
10. Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. Deep neural nets as a method for quantitative structure-activity relationships. J Chem Inf Model. 2015;55(2):263-74. [Crossref] [PubMed]
11. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: toxicity prediction using deep learning. Front Env Sci-Switz. 2016;3:80. [Crossref]

12. Ramsundar B, Kearnes S, Riley P, Webster D, Konerding D, Pande V. Massively multitask networks for drug discovery. arXiv preprint. 2015. [Link]
13. Koutsoukas A, Monaghan KJ, Li X, Huan J. Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. J Cheminform. 2017;9(1):42. [Crossref] [PubMed] [PMC]
14. Lenselink EB, ten Dijke N, Bongers B, Papadatos G, van Vlijmen HWT, Kowalczyk W, et al. Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. J Cheminform. 2017;9(1):45. [Crossref] [PubMed] [PMC]
15. Korkmaz S. Small drug molecule classification using deep neural networks. Türkiye Klinikleri J Health Sci. 2019;11(2):93-101. [Crossref]
16. Korkmaz S. Deep learning-based imbalanced data classification for drug discovery. J Chem Inf Model. 2020;60(9):4180-90. [Crossref] [PubMed]
17. Larochelle H, Bengio Y, Louradour J, Lamblin P. Exploring strategies for training deep neural networks. J Mach Learn Res. 2009;10:1-40. [Link]
18. Patterson J, Gibson A. Deep Learning: A Practitioner's Approach. 1st ed. USA: O'Reilly Media, Inc.; 2017.
19. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533-6. [Crossref]
20. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44. [Crossref] [PubMed]
21. Chollet F. Deep Learning with Python. 1st ed. Shelter Island: Manning Publications Company; 2017.
22. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273-97. [Crossref]
23. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. [Crossref]
24. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466-74. [Crossref] [PubMed]