International Journal of Applied Pharmaceutics

ISSN - 0975 - 7058 Vol 12, Special Issue 1, 2020

Research Article

APPLICATION OF THE MACHINE AND DEEP LEARNING METHODS FOR THE CLASSIFICATION OF CANNABINOID- AND CATHINONE-DERIVED COMPOUNDS

WIDYA DWI ARYATI, MUHAMMAD SIDDIQ WINARKO, GERRY MAY SUSANTO, ARRY YANUAR*

Laboratory of Biomedical Computation and Drug Design, Faculty of Pharmacy, Universitas Indonesia, Depok, West Java, Indonesia.
Email: arry.yanuar@ui.ac.id

Received: 02 December 2019, Revised and Accepted: 03 January 2020

ABSTRACT

Objective: New psychoactive substances (NPS) have been rapidly developed to avoid legal entanglement. In 2013–2018, the number of cathinone-derived compounds increased from 30 to 89. In 2016, of 56 NPS compounds, 21 were identified as cannabinoid-derived; only 43 were regulated in the narcotics law. Artificial intelligence, such as machine and deep learning, is a method of data processing and object recognition, including human poses and image classifications.

Methods: Herein, the machine and deep learning methods for cathinone- and cannabinoid-derived compound classification were compared using pharmacophore modeling as the reference method. For classifying cathinone-derived compounds, the structure was transformed into fingerprints, which was used as a learning parameter for the machine and deep learning methods. Contrarily, the physicochemical properties and fingerprint shape were utilized as learning materials for the deep learning method to classify the cannabinoid-derived substances.

Results: Consequently, in the cathinone-derived compound classification, the deep learning method produced the accuracy and Cohen kappa values of 0.9932 and 0.992, respectively. Furthermore, such values in the pharmacophore modeling method were higher than those in the machine learning method (0.911 and 0.708 vs. 0.718 and 0.673, respectively). In the cannabinoid-derived compound classification, the deep learning method with the fingerprint form had the highest accuracy and Cohen kappa values (0.9904 and 0.9876). Such values in this method with the descriptor form were higher than those in the pharmacophore modeling method (0.8958 and 0.8622 vs. 0.68 and 0.396, respectively).

Conclusion: The deep learning method has the potential in the NPS classification.

Keywords: Deep learning, Cannabinoid, Cathinone, Pharmacophore modeling, Psychoactive substance.

 DOI: http://dx.doi.org/10.22159/ijap.2020.v12s1.FF005

INTRODUCTION

New psychoactive substances (NPS) are frequently misused psychoactive compounds that mimic psychoactive drugs. Some of these are the cathinone and cannabinoid derivatives, which are often modified to avoid legal entanglement [1-5]. Due to the rapid development of NPS, new methods are made to quickly identify them as one of the compounds regulated in a country.

In silico research can support the conventional one, which requires a relatively long time. Using this approach, a new cannabinoid ligand has been discovered. Accordingly, this *in silico* approach can determine the similarities in a compound by comparing its physicochemical properties and quickly analyze the structural similarity in large numbers. In addition, the fingerprint modeling can be used to identify the similarities in both 2- and 3-dimensional structures [6]. The pharmacophore modeling can also search for structural similarity [7].

In this study, the machine and deep learning methods for the cathinone- and cannabinoid-derived compound classification were compared, and the pharmacophore modeling was used as a reference method. In the classification of the cathinone-derived compounds, the structure was transformed into fingerprints and this form was used as a learning parameter for the machine and deep learning methods. On the other hand, in the cannabinoid-derived compound classification, the physicochemical properties and fingerprint shape were used as the learning parameters for the deep learning method, which were expected to be an alternative approach in the classification process of compounds.

METHODS

Tools

The computers with Intel® i7 950 processor (CPU), Nvidia® GeForce GTX 680 graphics processor, and 24-GB DDR3 Random Access Memory with the Windows 10 Pro operating system were used in this study. Furthermore, the programs included Knime 3.5.1 [8], MarvinSketch 18.13, and LigandScout 4.2.

Materials

The database consisted of 360 2D NPS structures. In addition, the structures referring to PubChem were drawn using MarvinSketch and stored in 2D forms.

Classification of the cannabinoid- and cathinone-derived compounds using the pharmacophore modeling method

The structure of the NPS tested (cathinone and cannabinoid) was divided into two sets of compounds using the clustering method based on the structural similarity, and these were used as a training set to create a pharmacophore model. Compounds other than the training set were used as the test set. Subsequently, the pharmacophore model produced by the training set was tested against the test set. Then, the prediction results were used to determine the accuracy and Cohen kappa values.

Classification of the cathinone-derived compounds using the machine learning method

The structure of NPS compounds was drawn using MarvinSketch and stored in *.smi format. Then, using the fingerprint calculator in the Knime program, this was converted into a fingerprint in binary number

format. The variation was done on the number of bits (i.e., 512, 1024, and 2048), and the iteration was fixed at 10. Each experiment was carried out 5 times. Subsequently, the structure was labeled according to the NPS group conducted in the study. The fingerprint clustering in the Knime program was used for the classification and the results were shown by the accuracy and Cohen kappa values.

Classification of the cathinone-derived compounds using the deep learning method

Using the fingerprint calculator in the Knime program, the NPS structure in *.smi format was converted into a fingerprint in binary number format. Then, the compounds were separated into training and test sets. The former set was used as a learning material and for the prediction of the latter set. The variation was done on the number of bits (i.e., 1024, 2048, and 4096). Moreover, the iteration and dense layer in the learning parameters were fixed at 10 and 2, respectively. The variations in these parameters were carried out on the number of epochs (i.e., 5, 100, and 250). Each experiment was repeated 5 times, and the classification results were shown by the accuracy and Cohen kappa values.

Classification of the cannabinoid-derived compounds using the deep learning method

Two learning materials were used to classify the cannabinoid-derived compounds in the deep learning method: The fingerprint format and physicochemical property descriptors performed with the Knime program.

Changing structural image into a binary form using the fingerprint calculator in the Knime program did classification with fingerprint formats. Then, the compounds were separated into training and test sets. The former set was used as a learning material and for the prediction of the latter set. The variations were made on the number of bits (i.e., 512 and 1024). Moreover, the iteration and dense layer in the learning parameters were fixed at 50 and 2, respectively. The variations in these parameters were carried out on the number of epochs (i.e., 5, 10, and 50). Each experiment was repeated 5 times, and the classification results were shown by the accuracy and Cohen kappa values.

With the physicochemical property descriptors as the learning material, the classification was done by calculating the compound descriptor values using the descriptor calculation. The descriptors used in this study included the van der Waals surface area values based on the compound log p and the compound's atomic partial values, amounts of cyclic nitrogen and acyclic oxygen, and numbers of acyclic duplicates, nodules divided by two rings altogether, and ends divided by two rings altogether. Furthermore, the compounds were separated into training and test sets. The former set was used as a learning material and for the prediction of the latter set. Furthermore, the iteration and dense layers in the learning parameters were fixed at 50 and 2, respectively. The variations in these parameters were carried out on the number of epochs (i.e., 5, 10, and 50). Each experiment was repeated 5 times, and the classification results were shown by the accuracy and Cohen kappa values.

RESULTS AND DISCUSSION

Classification of the cannabinoid- and cathinone-derived compounds using the pharmacophore modeling method

A total of 127 test sets and 223 decoy compounds were used for the cannabinoid derivative classification. The prediction results of the pharmacophore modeling method for this classification are presented in Table 1.

For the classification of the cathinone-derived compounds, 44 test set and 271 decoy compounds were used. The prediction results of the pharmacophore modeling method for this classification are shown in Table 2.

Consequently, the results of the classification of the cannabinoid-derived compounds using the pharmacophore modeling method were poor. The Cohen kappa and accuracy values were 0.396 and 0.68, respectively. This was due to the diverse structure of the ligand used in this experiment, resulting in more varied data. The more identical the data in a group, the better the accuracy obtained [9].

On the contrary, the classification of the cathinone-derived compounds using the pharmacophore modeling method had a good sensitivity result, which was 1, indicating that this predicted an active compound more accurately. This method also had a specificity value of 0.897, denoting that most decoy compounds can be predicted correctly. However, some compounds were still predicted as false positives. In addition, its accuracy and Cohen kappa values were 0.911 and 0.708, respectively, signifying that this method had a strong acceptance level [10].

Both results indicated the differences in the acceptability values of the two methods. The pharmacophore modeling did not distinguish the active and decoy compounds well in the cannabinoid derivative classification but adequately differentiated them in the classification of the cathinone derivatives.

Classification of the cathinone-derived compounds using the machine learning method

The results of the cathinone derivative classification are shown in Figs. 1 and 2.

In this method, clustering was done based on the fingerprint similarity in the binary number form. The greater the data processed, the lower the accuracy of the prediction [11]. This was shown as the increase in the number of fingerprint bits reduced the accuracy and Cohen kappa values.

The best results in this method included a group with a strong acceptability value, with an average Cohen kappa value of 0.637 [10].

Table 1: Prediction results of the cannabinoid-derived classification using pharmacophore modeling method

S. No.	Database prediction	Active	Decoy
1.	Active	114	13
2.	Decoy	99	124

Table 2: Prediction results of the cathinone-derived classification using pharmacophore modeling method

S. No.	Database prediction	Active	Decoy
1.	Active	44	28
2.	Decoy	0	243

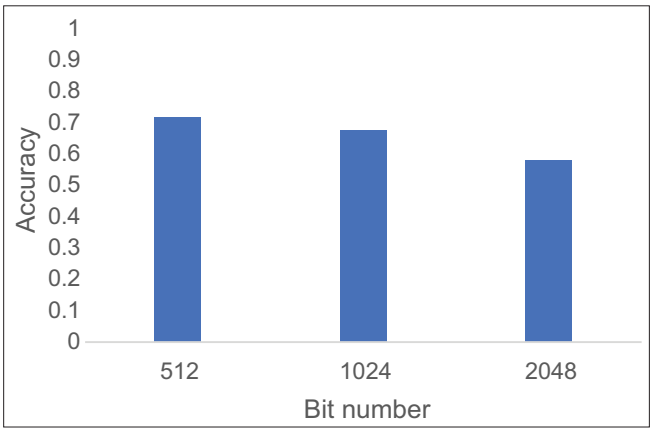

Fig. 1: Difference in the accuracy value to the number of bits

However, these results were not supported with good precision since the data were too varied.

Classification of the cathinone- and cannabinoid-derived compounds using the deep learning method

To classify the cathinone-derived compounds, a method with the bit number variations of 1024, 2048, and 4096 was used. The bit number used was the Morgan fingerprint type. This generates bits from tracking each atomic molecule with unique properties such as donors, acceptors, aromatic rings, halogens, and charges. Thus, this fingerprint is often used to search for molecular similarities. The results of the cathinone derivative classification are shown in Figs. 3 and 4, whereas those of the cannabinoid derivative one are presented in Figs. 5 and 6.

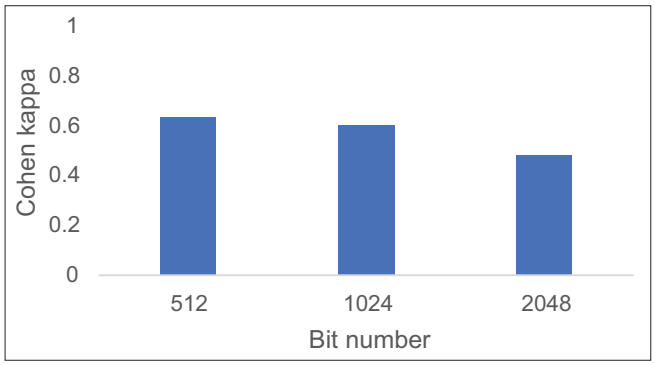

Fig. 2: Differences in the Cohen kappa value to the number of bits

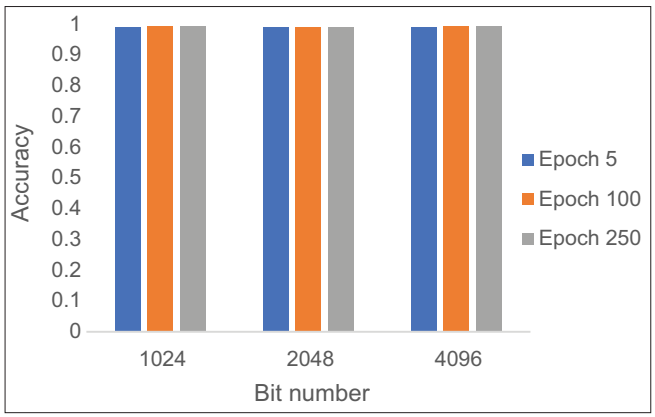

Fig. 3: Difference in the accuracy value to the bit number and epoch

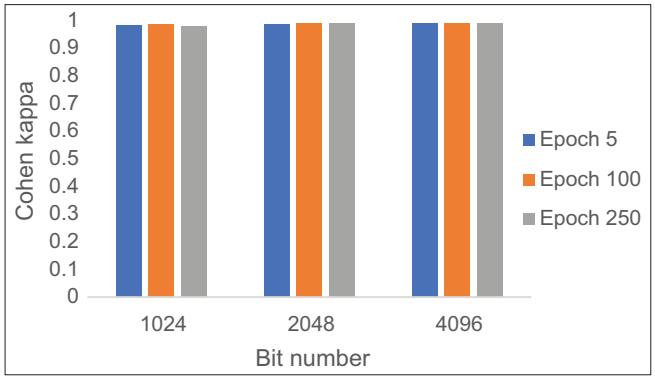

Fig. 4: Difference in the Cohen kappa value to the bit number and epoch

The variations of the bit number and epoch were done in this study. According to the research conducted by Gulli and Pal [12], the increased accuracy of the training and test sets occurs along with an increased epoch number. As shown in Fig. 4, the bit number 1024 had high accuracy (0.99272 for epoch 5) and Cohen kappa values. Increasing the epoch using the same bit number did not necessarily affect the accuracy or Cohen kappa values. However, the consistency of prediction results was noted: Five experiments produced the relative standard deviation (RSD) of 0.24% for the bit number 1024 and epoch 250. Moreover, the RSD value for the method with the bit number 1024 and epoch 5 was 0.32% and that with epoch 100 was 0.32%. Likewise, for the classification of the cannabinoid-derived compounds, the RSD value for the method with the bit number 512 and epoch 5 was 0.27% and that with epoch 50 was 0.17%. This was not in accordance with the results reported by Gulli and Pal [12]; therefore, it can be caused by the different deep learning methods used.

Increasing the bit number in the same epoch did not necessarily affect the accuracy or Cohen kappa values. In Fig. 5, no significant difference in the accuracy or Cohen kappa values was noted with the increased bits. Moreover, in epoch 5, the bit number 512 had an accuracy value of 0.9902, while 1024 had 0.9900. The addition of bits actually decreased the consistency of the prediction results, which can be seen from the RSD value on epoch 50 for 512 bits by 0.17% and 1024 bits by 0.51%.

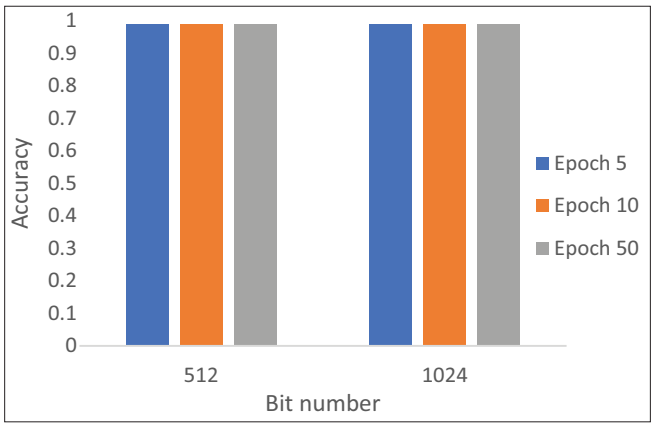

Fig. 5: Difference in the accuracy value to the bit number and epoch

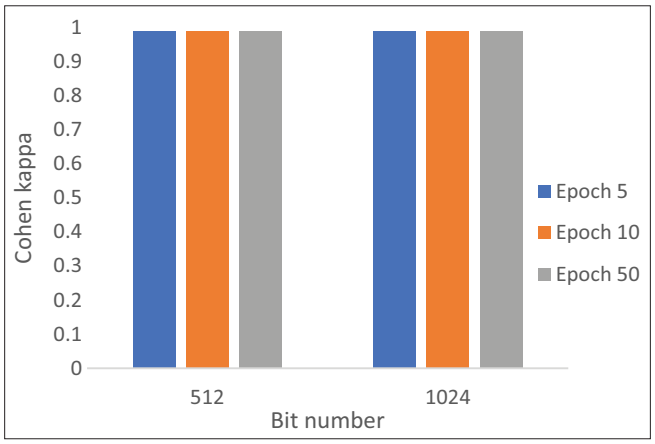

Fig. 6: Difference in the Cohen kappa value to the bit number and epoch

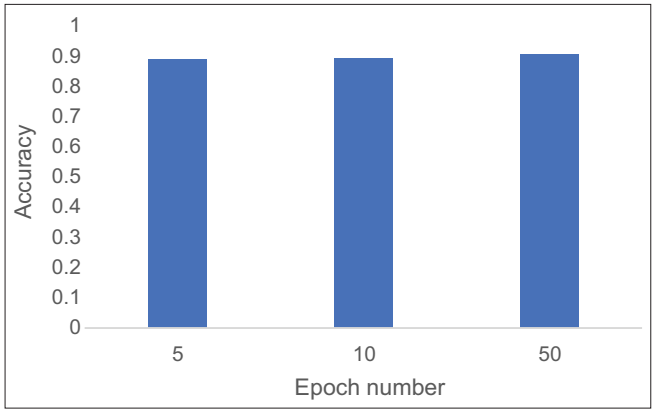

Fig. 7: Difference in the accuracy value to the epoch number

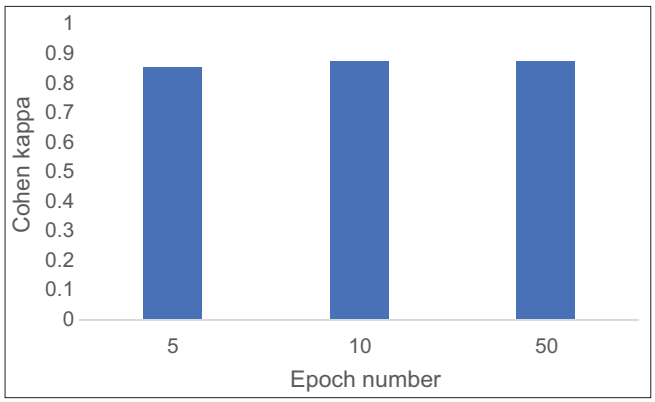

Fig. 8: Difference in the Cohen kappa value to the epoch number

Classification of the cannabinoid-derived compounds using the deep learning method with the descriptor

The results of the classification of the cannabinoid-derived compounds using the deep learning method with the descriptor as a learning material are shown in Figs. 7 and 8.

In this method, the increased number of epochs did not significantly affect the accuracy or Cohen kappa values. This also has poor consistency as seen from its highest RSD value of 6%. Thus, the results of the deep learning method with the descriptor were considered to be less consistent [13] because the descriptor parameters used as the learning material were less specific compared with those with the fingerprint method. In the deep learning method with the descriptor, the learning material relied only on the physicochemical values, so the data used to produce the learning algorithm were less specific and not absolute. Consequently, it showed less consistent results compared with the deep learning method with fingerprints, which used binary numbers that had absolute values, thereby producing an excellent learning material. Another factor that caused the validation value to be worse was the lacking number of training sets, and thus the learners were unable to predict all data correctly [14,15].

CONCLUSION

Based on the results, compared with the machine learning method, the pharmacophore modeling was better in classifying the cathinone-derived compounds. However, for the classification of the cannabinoid-derived compounds, the deep learning method was superior to the pharmacophore modeling one. In addition, this method with the descriptor learning materials was better than the pharmacophore modeling one for the cannabinoid derivative classification. Furthermore, among the three methods, the deep learning one with the fingerprint learning material was the best for the classification of the cannabinoid and cathinone derivatives.

ACKNOWLEDGMENTS

This study was financially supported by Directorate of Research and Community Enggagement (DRPM), Universitas Indonesia, via PITTA 2018.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

REFERENCES

1. Baumann MH. Awash in a sea of "bath salts": Implications for biomedical research and public health. Addiction 2014;109:1577-9.
2. Shadiq GF. Law enforcement of new psychoactive subtances narcotics crime based on law number 35 year 2009 concerning narcotics [Penegakan hukum terhadap tindak pidana narkotika new psychoactive subtances berdasarkan undang-undang nomor 35 tahun 2009 tentang narkotika]. J Wawasan Yuridika 2017;1:35-53.
3. Scientific Working Group for the Analysis of the Seized Drugs. Available form: http://www.swgdrug.org. [Last accessed 2019 Aug 08].
4. Badan Narkotika Nasional. Prevention and Eradication of Drug Abuse and Illicit Trafficking [Pencegahan dan pemberantasan penyalahgunaan dan peredaran gelap narkoba]. Vol. 4. Jakarta: Badan Narkotika Nasional; 2016.
5. Sumitha SK, Pattammady VS, Sambathkumar R. Pharmacology of noval cannabinoids. Int J Pharm Pharm Sci 2019;12:1-5.
6. Eckert H, Bajorath J. Molecular similarity analysis in virtual screening: Foundations, limitations and novel approaches. Drug Discov Today 2007;12:225-33.
7. Wolber G, Langer T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. J Chem Inf Model 2005;45:160-9.
8. KNIME-Open for Innovation. Knime.com. Available form: https://www.knime.com. [Last accessed 2019 Aug 08].
9. Cho J, Lee K, Shin E, Choy G, Do S. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? 2015;ArXiv:1511.06348.
10. Viera AJ, Garrett JM. Understanding interobserver agreement: The kappa statistic. Fam Med 2005;37:360-3.
11. Freitas AA, Lavington SH. Mining Very Large Databases with Parallel Processing. Boston: Kluwer Academic Publishers; 2000.
12. Gulli A, Pal S. Deep Learning with Keras. Birmingham: Packt Publishing; 2017.
13. Simonsen KL, Churchill GA, Aquadro CF. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 1995;141:413-29.
14. Bennett ER, Clausen J, Linkov E, Linkov I. Predicting physical properties of emerging compounds with limited physical and chemical data: QSAR model uncertainty and applicability to military munitions. Chemosphere 2009;77:1412-8.
15. Polamreddy P, Vishwakarma V, Mahto MK. Combinatorial pharmacophore modeling and atom based 3D QSAR studies of benzothiadiazines as HCV-NS5B inhibitors. Int J Pharm Pharm Sci 2018;10:43-69.