International Journal of Chemical and Biochemical Sciences
(ISSN 2226-9614)

Journal Home page: www.iscientific.org/Journal.html

Identification of DPP-4 Inhibitor Active Compounds Using Machine Learning Classification

Oky Hermansyah, Suci Rahmawati, Camelia Dwi Putri Masrijal, Rose Intan Permasari, Samwilson Slamet*

Diploma III Pharmacy Program, Faculty of Mathematics and Natural Sciences, University of Bengkulu, Indonesia.

Abstract

The development of drug discovery is currently reaching an era where it is not limited to laboratories with solvents and instrument equipment but is developing towards computing by utilizing artificial intelligence and big data which can currently be accessed easily. Several machine learning models can be used for virtual screening of various drug targets including DPP-4 inhibitors, this class of drugs is increasingly used in the treatment of type 2 diabetes because in addition to increasing insulin secretion from the pancreas it can also neutralize body weight. The purpose of this study is to create an artificial intelligence model using a classification machine learning algorithm, and utilize the model to find DPP-4 inhibitor compounds from US FDA-registered drug chemical compounds, so that the clinical trial process for potential compounds will be easier because the safety profile of the drug is already known. The data used for modeling is a dataset of research results on the DPP-4 enzyme in the ChEMBL database which amounts to 5098 data, which is then partitioned as much as 80% of the data for the training set and 20% of the data for the test set. From the results of research using 3 machine learning models, namely XGBoost tree ensemble, Random Forest, and Support Vector Machine, the best model was obtained from the results of internal validation with 10 times cross validation and external validation obtained the highest accuracy of 81.64%, namely the XGBoost Tree Ensemble model. This model is used for screening on 2096 drug data that has been registered by the US FDA. The results after similarity testing with the modeling database, DUDE Decoys and Lipinsky's Rule of 5 obtained 29 compounds that have the potential to inhibit DPP-4. These compounds can be further developed with invitro or invivo studies to find new lead compounds in anti-diabetic therapy for DPP-4 inhibitors.

Keywords: DPP-4, Virtual Screening, Machine Learning, Knime

Full-length article **Corresponding authors*, e-mail: Oky.hermansyah@Unib.ac.id

1. Introduction

In the current era of industrial revolution 4.0, technological developments are rapidly penetrating all fields, including the health sector. Drug discovery is a highly developed field due to the development of chemical library databases such as PubChem, which contains more than one hundred million compounds; various research databases such as ChEMBL, which is systematically arranged with very complete macromolecular targets; and several other databases such as Molport and Drug3D. This development is also followed by discoveries related to artificial intelligence that can be utilized in drug discovery [1, 2]. The use of machine learning algorithms such as Deep Learning, XGBoost Tree Ensemble, Random Forest, and Support Vector Machine for the search for new drugs has been widely reported. Even in 2012, a Kaggle competition was conducted in collaboration with the Merck pharmaceutical industry to utilize deep learning for drug activity prediction [3]. During the COVID-19 pandemic, artificial intelligence was also instrumental in screening the activity of new compounds in order to find active compounds that could inhibit COVID-19 disease [4].

In medicinal chemistry, there are two commonly used methods to predict the activity of compounds for drugs, namely ligand-base drug discovery (LBDD) and structure-base drug discovery (SBDD). In LBDD, for example, in Quantitative Structure and Activity Relationship (QSAR), research data is used, and then a relationship between the molecular structure of chemical compounds and their activity is modeled. The well-known conventional QSAR is the Hansch model, but along with the development of big data and AI, machine learning models are being made that are able to predict activity both in the form of classification (active, inactive, toxic, or non-toxic) and regression (inhibitory activity values such as pIC50 and ki) [2, 5]. Machine learning classification models are growing rapidly due to their better accuracy compared to regression models, which are very difficult to achieve a coefficient of determination above 0.8. This QSAR approach with artificial intelligence allows the process to be done automatically and in a shorter time for analysis with data containing thousands or even millions of chemical compounds [6, 7].

The classification machine learning models we use are Random Forest, XGBost Tree Ensemble, and Support Vector Machine. These models are widely used today to make predictions of the activity of various chemical compounds against drug targets as well as predictions in other fields outside of health. For the target that we developed is DPP-4 inhibitors. DPP-4 is an enzyme that inhibits the work of GLP1 and GIP hormones to respond to insulin release

when blood sugar increases, so that insulin cannot be secreted. By inhibiting this enzyme, insulin can be secreted, without increasing the incidence of hypoglycemia, currently DPP-4 inhibitor antidiabetic drugs are increasingly replacing the role of sulfonyl urea because it has better glycemic control in the treatment of diabetic patients [8] and has the advantage that it can be used in obese patients, and is safe for kidney and heart patients [9]. From the results of a study of patients receiving sulfonyl urea drugs, it was reported that the proportion of patients who experienced at least one hypoglycemic episode was 45% in patients taking sulfonylureas, and experienced an average weight gain of 4 and 2 kg respectively. This weight gain can increase insulin resistance and ultimately worsen the disease [10]. In addition, some of the current DPP-4 class drugs such as sitagliptin and vildagliptin show serious side effects such as pancreatitis [11]. From the above explanation, the discovery of new DPP-4 inhibitor drugs that are safe for this class of drugs is increasingly attractive, to find the ideal drug with the least possible side effects.

By utilizing artificial intelligence using a classification machine learning model, it is hoped that a classification model with high accuracy can be produced so that it can be used for virtual screening, predicting thousands or even millions of new compounds whose potential against the DPP-4 enzyme is unknown. The predicted potential compounds can later be used for further studies such as molecular docking, in vitro, or in vivo studies. Furthermore, it is hoped that these potential compounds can become lead compounds for new DPP-4 inhibitors.

The chemical compound data that we will screen is a data set from e-Drug3D that contains chemical compounds that have been registered by the USA FDA for circulation, so the resulting screening compounds will be easier to obtain and test further because data related to the chemical compound is already available, such as safety, toxicity, and maximum dose.

2. Materials and Methods

This research was conducted in silico using an AMD Ryzen 5 computer and 16 GB of RAM. The software used for HKSA modeling with machine learning is KNIME v.5, which is an opensource application for data mining and various cheminformatics activities on one screen[12]. The database used for modeling with machine learning is the database from ChEMBL with the homo sapiens DPP-4 target and pIC50 activity dataset downloaded at the web URL address: https://www.ebi.ac.uk/ [13, 14].

2.1. Machine learning modeling

The data set for modeling comes from data on DPP-4 inhibitory activity from various studies in the ChEMBL database. The data set is curated, then partitioned into 80% for the training set and 20% for the test set. Machine learning algorithms that will be developed are supervised learning classification algorithms, including Random Forest, XGBoost Tree Ensemble, and Support Vector Machine. The model is optimized to get the optimal machine learning algorithm parameters, and then the model's performance is trained with 80% of the data set. The model is then externally validated with a 20% test set to see its performance on unseen data (data outside of modeling) (Shown in Figure 1) [15, 16].

2.2. Model Evaluation

To see the predictive ability of the model, internal validation with 10-fold cross-validation of the training data and external validation of the test data were performed. Furthermore, all models were evaluated with:

- True Positive (TP) is an active compound that is correctly predicted as an active compound.
- True Negative (TN) is an inactive compound that is correctly predicted as an inactive compound.
- False Positive (FP) is an inactive compound that is incorrectly predicted as an active compound.
- False Negative (FN) is an active compound that is predicted incorrectly and classified as an inactive compound.

In addition, the model is also evaluated with several parameters, such as sensitivity to determine the ability of the model to detect active compounds, specificity for inactive compounds, and accuracy for all compounds. Recall (RE) = sensitivity (SE), which is the proportion (percentage) of active compounds predicted to be active; specificity (SP), which is the proportion (percentage) of inactive compounds predicted to be inactive; and accuracy, which is the proportion (percentage) of chemical compounds correctly classified. F-measures refer to the harmonic mean of recall and precision, where recall refers to the real prediction accuracy and precision defines the accuracy of the predicted class [17, 18].

2.3. Virtual Screening

The chemical compounds that will be screened are datasets from e-Drug3D chem, which is data on drugs that have been approved by the USA FDA for circulation. This data contains approximately 2056 molecular structures downloaded from https://chemoinfo.ipmc.cnrs.fr [19].

3. Results and Discussions

From the ChEMBEL database, there are around 5098 research results on targets, which are then curated against missing value data, which includes molecular structure and pIC50 value. The final result is 5094 research results, which are curated again against the same data, and then the latest year is selected. Finally, 4198 molecules of DPP-4 enzyme research data are obtained, and the molecular structure data in the form of smiles is converted into images and desalted with the RDKit salt stripper node. From the ChEMBEL database, there are around 5098 research results on targets, which are then curated against missing value data, which includes molecular structure and pIC50 value. The final result is 5094 research results, which are curated again against the same data, and then the latest year is selected. Finally, 4198 molecules of DPP-4 enzyme research data are obtained, and the molecular structure data in the form of smiles is converted into images and desalted with the RDKit salt stripper node, as shown in Figure 2.

A total of 3592 research data points were then calculated as descriptors using the CDK Fingerprint node; the fingerprint used was the extended connectivity fingerprint (ECFP) of 6 1024 bits. previously used the CDK Descriptor and RDKit Descriptor nodes to calculate molecular descriptors such as SlogP, NumLipinski HBA, NumLipinski HBD, Element Count, Bond Count, and several other descriptors related to molecular properties. However, the

accuracy generated with the CDK Molecular Properties node is the highest at 74.5%, compared to the RDKit Descriptor node at 77.4% (shown in Figure 3). The best accuracy was obtained using the 1024-bit ECFP6 fingerprint, which reached 80%.

Various studies have used the ECFP6 fingerprint as a variable including research on JAK2 protein inhibitors, prediction of binding to estrogen receptors, and many related studies that use the ECFP6 fingerprint in developing machine learning models [20, 21]. After calculating the fingerprint of all data, the data is then classified into 2 classes of compounds, namely active compounds (pIC50 $\geq$ 7.5) and inactive compounds (pIC50 < 7.5). The division of this compound class produces a maximum accuracy of 80.79%. The division of this compound class is based on various previous studies; generally, the cut-off for active compounds is pIC50> 7.5. The pIC50 value of 7.5 is equivalent to an IC50 value of 500 nM [22]. The data was partitioned into two parts: a training set and a test set, with a ratio of 80% (2873 datasets): 20% (719 datasets). The training set was used for optimization and model training, and the test set was used for external model validation [15, 23].

There are three machine learning algorithms that we developed for virtual screening, namely Random Forest, XGBoost Tree Ensemble, and Support Vector Regression. All three are supervised learning classification models. These three machine learning models are currently widely used in analyzing large data sets for predicting the activity of active compounds. The three models work in different ways, Random Forest is a supervised learning algorithm that integrates multiple trees through the idea of Bagging. The bootstrap method is used to extract training sample sets from the original sample data, and the corresponding decision tree model is trained for each training set. Finally, all the base classifiers are voted, and the one with the most votes is the final category. XGBoost, short for extreme gradient boosting, is a boosting algorithm. Both XGBoost and random forest are integration algorithms based on decision trees. Unlike the Bagging algorithm, the Boosting algorithm builds up weak learnings one by one, accumulating multiple weak learnings through continuous iteration [24]. Support Vector Machine (SVM) is a popular method for classification analysis. It is a supervised learning approach that selects a small number of significant boundary samples from all classes and creates a linear discriminant function that divides them as widely as possible. SVM separates groups into different categories using a multi-dimensional hyperplane [25].

The first step to finding a predictive model is parameter optimization to get the model with the highest accuracy. For data with a pIC50 cutoff of 7.5, the parameter values in the xgboost Tree Ensemble that are searched are nRounds and maxdepth; the highest accuracy is obtained at nRounds 975 and maxdepth 10. For other parameters, we set the default. In the random forest model, the optimized parameters are nModel and tree depth. The optimal nModel value is 577 and Tree Depth is 25, and the optimized support vector machine parameters are sigma and pinalty; the sigma value that gives the highest accuracy is 0.324 and the pinalty value is 17.974. After obtaining the optimal parameters, the next training is carried out on the model with 10-fold cross-validation. From the training model, the best accuracy is obtained on the XGBoost tree ensamble model of 80.79%. The following results are in Table 1 and table 2.

3.1. Model Evaluation

In the external validation of the model against data outside the modeling of 719 test sets, the XGBoost Tree ensemble algorithm produces the highest accuracy, of 0.8164 (82,64) or 81.64% and cohen's kappa 0.617 in figure 4. From several models that have been evaluated based on the results of internal and external validation, the XGBoost Tree Ensemble model produces the highest accuracy among random forest and support vector machine models. This best model will be used for virtual screening of the molecular structure database from e-Drug3D.

3.2. Virtual Screening of The Database

The first step in virtual screening begins with reading the e-Drug3d database, which contains a drug that has been registered by the US FDA in 2056 molecules, with the node mol2 reader. This was followed by data curation by removing missing Velue fragment molecules and salt data. The results of this process left approximately 11 molecules. Next, the fingerprint calculation was carried out with the CDK fingerprint node. The fingerprint chosen is the same as the modeling, namely ECFP 6, 1024 bits. After that, the screening process was carried out with the XGBoost Tree Ensemble model using the parameters and data sets from the previous modeling results. The results of 2011 molecules obtained compounds that have the potential to inhibit the DPP-4 enzyme in as many as 117 molecules, and these molecules were tested for similarity with the ChEMBL database and the DUDE Decoys database [26, 27]. To ensure that these potential molecules have not been studied on the DPP-4 enzyme before.

This process resulted in 93 potential molecules. The last step is Lipinski's Rule-of-Five test to see if the structure of this potential molecule meets the rules for being absorbed and circulated in the blood so that it can reach the target [28]. 29 compounds meet Lipinski's Rule of Five molecules have no more than 5 H-bond donors and no more than 10 H-bond acceptors; the molecular weight is no more than 500; and the logP is no more than 5. Some of the potential DPP-4 inhibitor compounds that we found include: ZOLPIDEM is a sedative-hypnotic currently used to treat insomnia, and PRASUGREL is a medicine to prevent blood clots. ACALABRUTINIB is used to treat mantle cell lymphoma (a type of non-Hodgkin's lymphoma) in adults. Baricitinib is an anti-inflammatory drug usually used to treat rheumatoid arthritis.

TRICLABENDAZOLE is a medicine for diseases caused by Fasciola hepatica or Fasciola gigantica worm infections. Some of these compounds are used for antibacterial and even cancer treatment, so benefit and side effect assessments are a consideration. However, some drugs can be optimized, such as those with anti-inflammatory and anti-parasitic properties. In fact, some of these drugs were also reported to have activity against infections caused by COVID-19, and some were even added to therapies such as Zulfidem [29, 30]. Potential compounds as DPP-4 inhibitors from the screening results are shown in Figure 5.

These screening compounds are only predictions, despite their high accuracy and the results of other statistical parameters that meet the validation test. These predicted results can still be further developed into lead compounds by conducting in silico tests again with molecular docking or

molecular dynamics and in silico studies in experimental animals.

Table 1: Model training results with the highest accuracy at a pIC50 cutoff of 7.5 based on class statistics.

Statistical parameters	**Active**	**Inactive**
Active (Predicted)	860	257
Inactive (Predicted)	295	1461
True Positives	860	1461
False Positives	257	295
True Negatives	1461	860
False Negatives	295	257
Precision	74,46%	85,04%
Sensitivity	76,99%	83,20%
Specificity	74,46%	74,46%
F-Measure	75,70%	84,11%

Table 2: Model training results with the highest accuracy at a pIC50 cutoff of 7.5 based on overall statistics

Overall Accuracy	Overall Error	Cohen's kappa (k)	Correctly Classifield	Inncorrectly Classifiede
80,79%	19,21%	0,598	2321	552

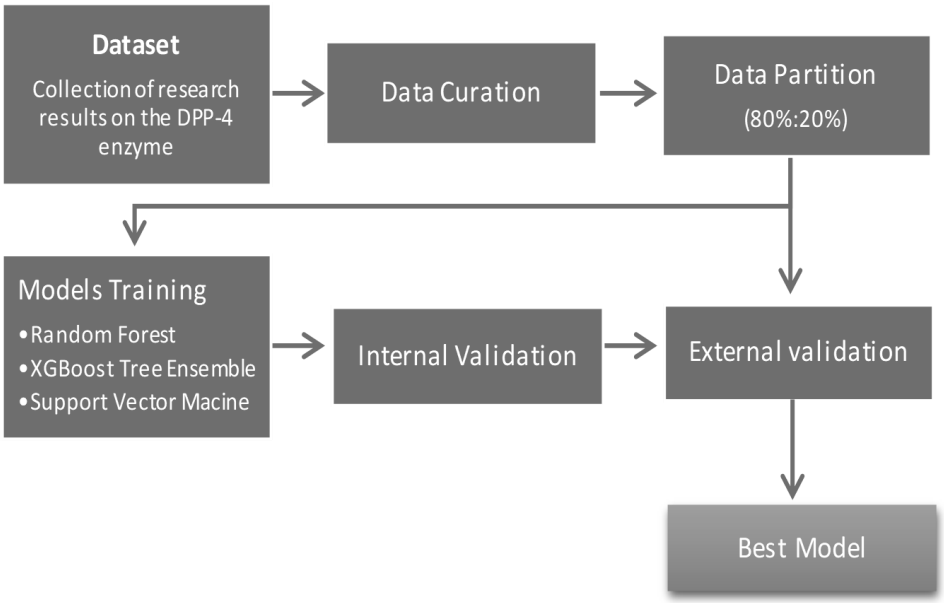


Figure 1: Overview of steps to create a prediction model

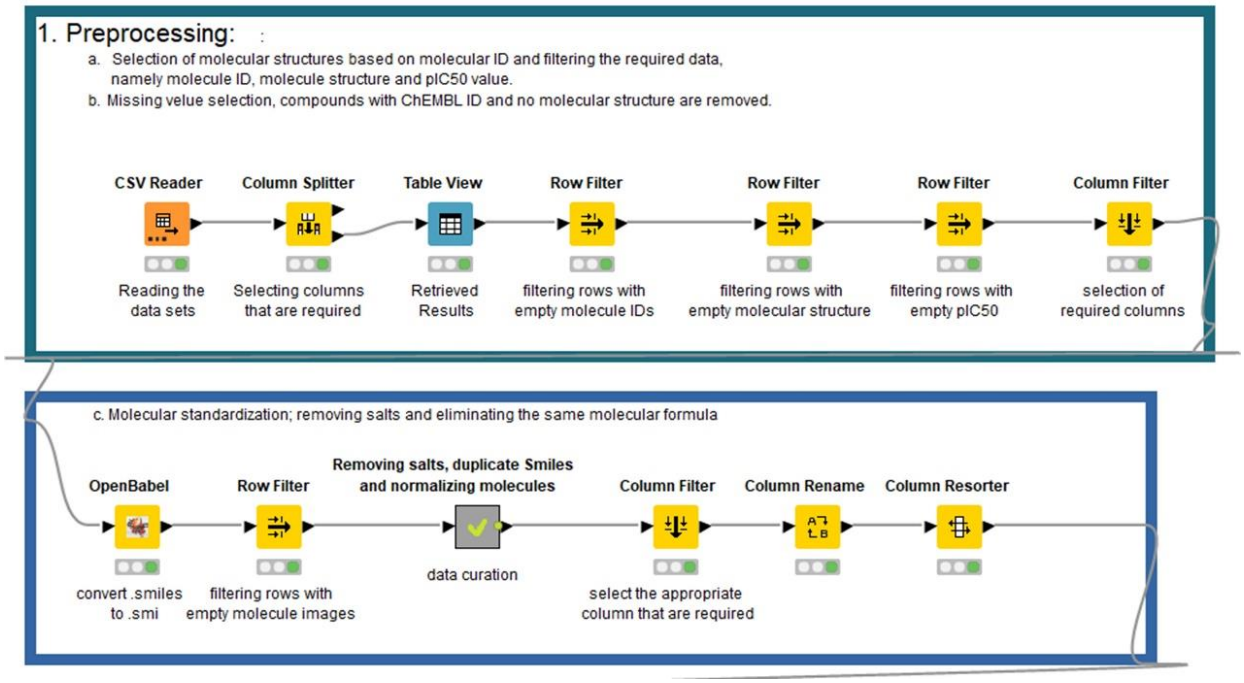
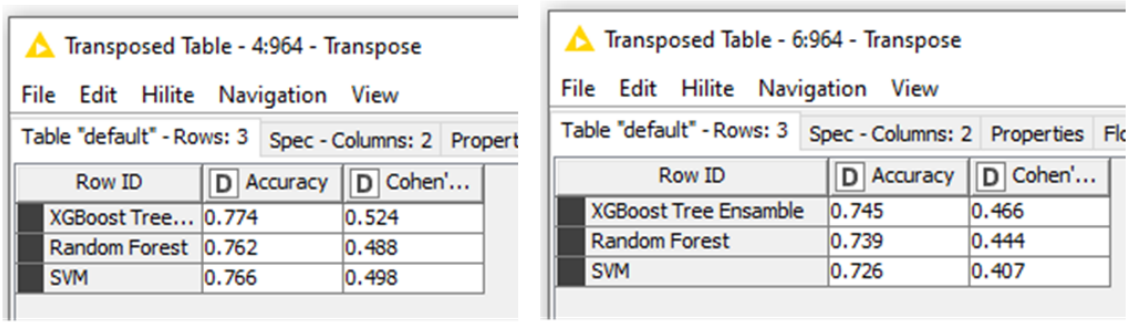


Figure 2: Nodes for data curation with RDKit



(a) (b)

Figure 3: Accuracy with descriptors (a) nodes from the Molecular Properties CDK and (b) nodes from the RDKit Descriptor

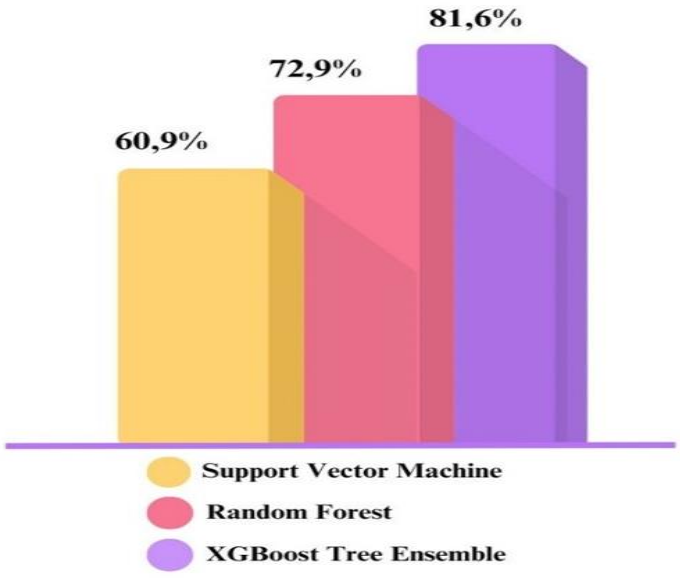


Figure 4: Accuracy of some models on the test set

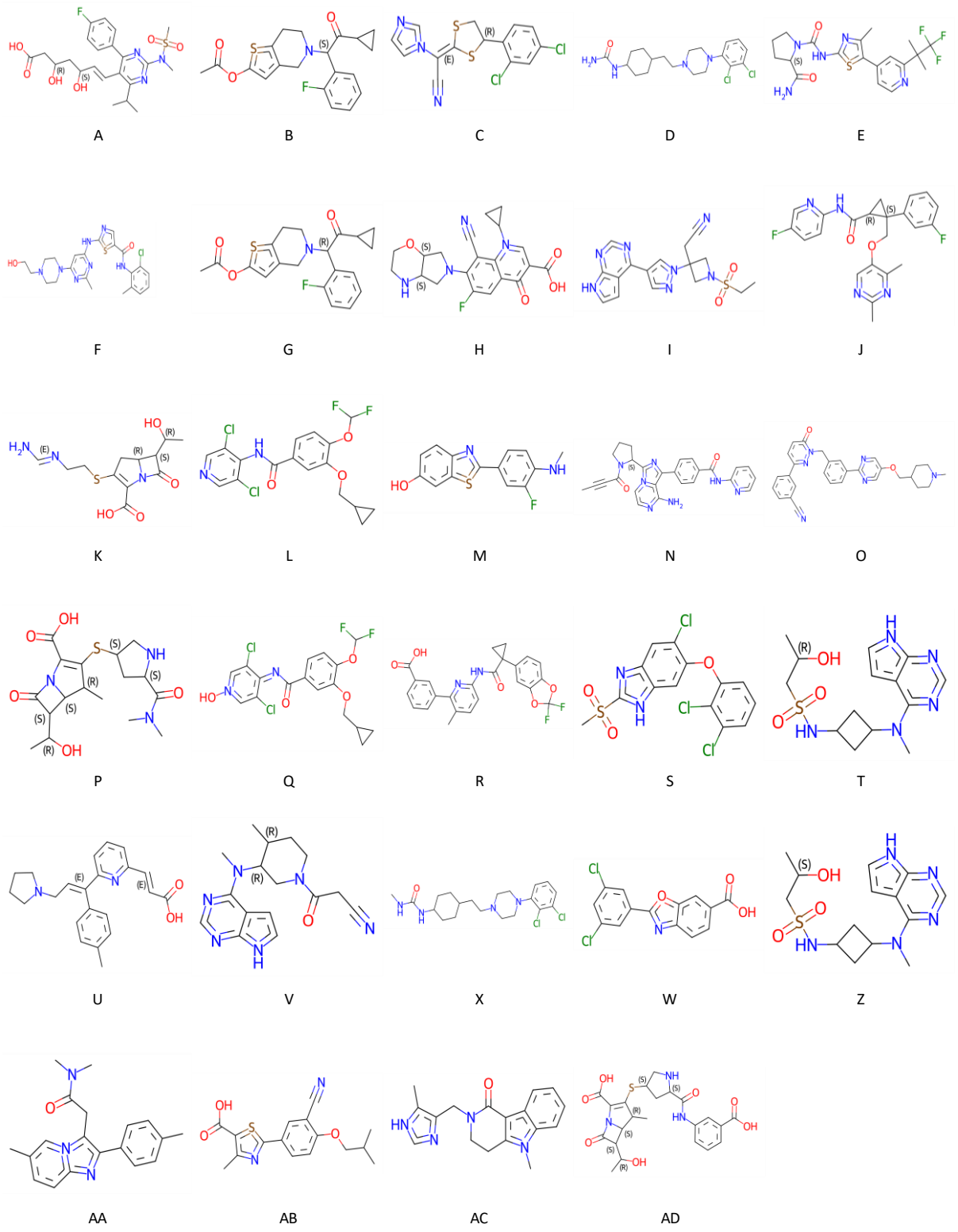


Figure 5: Molecular structures of potential compounds as DPP-4 inhibitors from virtual screening

4. Conclusions

However, potential compounds from the test results will be quickly developed into DPP-4 inhibitor class antidiabetic therapies because the drug profile has been previously known from the US Pharmacopeia (USP). The machine learning model we used for modeling DPP-4 inhibitors is XGBoost tree ensemble because, based on the results of internal and external validation, the highest accuracy was obtained, and other parameters such as sensitivity, recall, specificity, accuracy, F-measure, and precision were met compared to other machine learning models such as Random Forest and Support Vector Machine. In screening the database from e-Drug3D, which is a drug that has been registered by the US FDA for as many as 2056 molecules, 29 potential compounds were obtained that can be further developed into DPP-4 inhibitor lead compounds through in vitro and in silico tests.

Acknowledgements

The authors would like to thank the Faculty of Mathematics and Natural Sciences of Bengkulu University for fully funding this research through a research grant scheme called "Pembinaan Penelitian" in 2023 (1974/UN30.12/HK/2023), and all institutions and individuals who have provided moral and material support during the implementation of the research**.**

Declaration of Interest Statement

The authors declare that they have no conflict of interests.

Availability of data and material

The data supporting this research is openly available at https://bit.ly/3YVE4JB

References

[1] K.-K. Mak, M. R. Pichika. (2019). Artificial intelligence in drug development: present status and future prospects. Drug discovery today. 24 (3): 773-780.

[2] J. Panteleev, H. Gao, L. Jia, (2018). Recent applications of machine learning in medicinal chemistry. Bioorganic & medicinal chemistry letters. 28 (17): 2807-2815.

[3] T. Unterthiner, A. Mayr, G. Klambauer, M. Steijaert, J. K. Wegner, H. Ceulemans, S. Hochreiter. (2014). Deep learning as an opportunity in virtual screening. in Proceedings of the deep learning workshop at NIPS. 27: 1-9.

[4] S. Mohanty, M. H. A. Rashid, M. Mridul, C. Mohanty, S. Swayamsiddha. (2020). Application of Artificial Intelligence in COVID-19 drug repurposing. Diabetes & Metabolic Syndrome: Clinical Research & Reviews. 14(5): 1027-1031.

[5] A. Lavecchia. (2015). Machine-learning approaches in drug discovery: methods and applications. Drug discovery today. 20(3): 318-331.

[6] S. Kausar, A. O. Falcao. (2018). An automated framework for QSAR model building. Journal of cheminformatics.10(1):1-23.

[7] Y. C. Lo, S. E. Rensi, W. Torng, R. B. Altman. (2018). Machine learning in chemoinformatics and drug discovery. Drug discovery today. 23(8): 1538-1546.

[8] G. P. Fadini, D. Bottigliengo, F. D'Angelo, F. Cavalot, A. C. Bossi, G. Zatti, I. Baldi, A. Avogaro. (2018). Comparative effectiveness of DPP-4 inhibitors versus sulfonylurea for the treatment of type 2 diabetes in routine clinical practice: a retrospective multicenter real-world study. Diabetes Therapy. 9: 1477-1490.

[9] B. Gallwitz. (2019). Clinical use of DPP-4 inhibitors. Frontiers in endocrinology. 10: 389.

[10] S. R. Aravind. (2014). Adding a DPP-4 inhibitor to metformin therapy may be safer than you think. Taylor & Francis. 30: 791-794.

[11] F. Yang, Y. Dong, B. Li, B. Ning, Q. Zhao. (2022). Pancreatic safety of DPP-4 inhibitors in type 2 diabetes millitus: A protocol for systematic review and network meta-analysis. Medicine. 101: 17.

[12] R. Silipo. (2018). KNIME Beginner's Luck. Zurich - Switzerland: KNIME Press. 248-248.

[13] O. Méndez-Lucio, B. Baillif, D. A. Clevert, D. Rouquié, J. Wichard (2020). De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. Nature communications. 11(1): 10.

[14] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, E. Félix, M. P. Magariños, J. F. P, Mutowo., Nowotka, P. (2019). ChEMBL: towards direct deposition of bioassay data. Nucleic acids research. 47(D1): D930-D940.

[15] O. Hermansyah, A. Bustamam. A. Yanuar. (2021). Virtual screening of dipeptidyl peptidase-4 inhibitors using quantitative structure–activity relationship-based artificial intelligence and molecular docking of hit compounds. Computational Biology and Chemistry. 95: 107597

[16] S. Syarofina, A. Bustamam, A. Yanuar, D. Sarwinda, O. Hermansyah (2020). Cluster analysis in prediction of biological activity and molecular structure relationship of dipeptidyl peptidase-4 inhibitors for the type two diabetes mellitus treatment. AIP Conference Proceedings.

[17] K. Roy, S. Kar, and R. N. Das, (2015). A primer on QSAR/QSPR modeling: fundamental concepts. Springer.

[18] N. Singh, S. Chaudhury, R. Liu, M. D. M. AbdulHameed, G. Tawa, and A. Wallqvist. (2012). QSAR classification model for antibacterial compounds and its use in virtual screening. Journal of chemical information and modeling. 52(10): 2559-2569.

[19] E. Pihan, L. Colliandre, J. F. Guichou, D. Douguet. (2012). e-Drug3D: 3D structure collections dedicated to drug repurposing and fragment-based drug design. Bioinformatics. 28(11): 1540-1541.

[20] S. Belenahalli Shekarappa, S. Kandagalla, J. Lee, (2023). Development of machine learning models based on molecular fingerprints for selection of small molecule inhibitors against JAK2 protein," Journal of Computational Chemistry. 44(16): 1493-1504.

[21] D. P. Russo, K. M. Zorn, A. M. Clark, H. Zhu, S. Ekins. (2018). Comparing multiple machine learning algorithms and metrics for estrogen

receptor binding prediction. Molecular pharmaceutics.15(100) :4361-4370.

[22] C. Selvaraj, S. K. Tripathi, K. K. Reddy, S. K. Singh (2011). Tool development for Prediction of pIC50 values from the IC50 values-A pIC50 value calculator. Current Trends in Biotechnology and Pharmacy. 5(2): 1104-1109.

[23] S. Simeon and N. Jongkon. (2019). Construction of quantitative structure activity relationship (QSAR) Models to predict potency of structurally diversed janus kinase 2 inhibitors. Molecules, 24(23): 4393.

[24] H. Chen, N. Wang, X. Du, K. Mei, Y. Zhou, G. Cai. (2023). Classification Prediction of Breast Cancer Based on Machine Learning. Computational Intelligence and Neuroscience. 2023: 6530719. 2023/01/11 2023, doi: 10.1155/2023/6530719.

[25] M. M. Hassan, M. M. Hassan, F. Yasmin, M. A. R. Khan, S. Zaman, Galibuzzaman, K. K. Islam, A. K. Bairagi. (2023), comparative assessment of machine learning algorithms with the Least Absolute Shrinkage and Selection Operator for breast cancer detection and prediction. Decision Analytics Journal, 7:100245.

[26] N. Huang, B. K. Shoichet, J. J. Irwin, (2006). Benchmarking sets for molecular docking, Journal of medicinal chemistry. 49(23): 6789-6801.

[27] M. M. Mysinger, M. Carchia, J. J. Irwin, B. K. Shoichet. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. Journal of medicinal chemistry. 55(14): 6582-6594.

[28] D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward, K. D. Kopple. (2022). Molecular properties that influence the oral bioavailability of drug candidates. Journal of medicinal chemistry, 45(12): 2615-2623.

[29] D. S. Wishart D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, J. T. Sajed, D. Johnson. C. Li, Z. Sayeeda. (2018). DrugBank 5.0: a major update to the DrugBank database for. Nucleic acids research. 46(D1). D1074-D1082

[30] T. C. Maigoda, R. Refdanita, F. Firdaus, A. Wahyudi, A. Rizal, J. Jumyati. (2022). Anti-Microbial Activity and Diabetic Wound Healing Effects of Psidium Guava and Melastoma Malabathricum Linn Leaves Gel in Rats (Sprague Dawley). Proceeding B-ICON. 1(1): 417-426.