Feature selection methods for machine learning-based docking prediction of Indonesian medicinal plant compounds and HIV-1 protease

***Abstract*—This work evaluates usage feature selection methods to reduce the number of features required to predict docking results between Indonesian medicinal plant compounds and HIV protease. Two feature selection methods, Recursive Feature Elimination (RFE) and Wrapper Method (WM), are trained with a dataset of 7,330 samples and 667 features from PubChem Bioassay and DUD-E decoys. To evaluate the selected features, a dataset of 368 Indonesian herbal chemical compounds labeled by manually docking to PDB HIV-1 protease is used to benchmark the performance of linear SVM classifier using different sets of features. Our experiments show that a set of 471 features selected by RFE and 249 by WM achieve a reduction of classification time by 4.0 and 8.2 seconds respectively. Although the accuracy and sensitivity are also increased by 8% and 16%, no meaningful improvement observed for precision and specificity.**

I. Introduction

The evolution of viruses can make them resistant to existing drugs. One of the most famous cases is HIV (Human Immunodeficiency Virus), which caused AIDS (Acquired Immunodeficiency Syndrome), which has been a global issue for years. HIV possesses high drug resistance due to its high replication and mutation abilities. Since drug discovery is a very complicated, expensive, and time-consuming, curing AIDS, and other illness caused by evolving virus become very challenging[1].

To discover new drugs, first, one needs to find a set of chemical compound candidates by observing the reaction to drug target in the lab. This process is usually called high-throughput screening (HTS). Despite its importance, this process is considered inefficient and expensive because most of the chemical compounds consumed in the experiments. One way to make this process more efficient is by reducing the number of compounds that need to be tested in the lab by performing virtual screening beforehand [2]. By having the number of lab experiments reduced, ultimately it will reduce the overall time and cost needed in drug discovery [3].

Virtual screening applies computer algorithms to find chemical compounds that have a high probability of reaction to the drug's target. One of its approaches is Ligand-Based Screening (LBS), where new candidates are chosen based on their structural or characteristic similarity to known drug's chemical compounds. This implies that the LBS approach relies on previous drug discovery results, which usually obtained using HTVS (High Throughput Virtual Sreening) such as PubChem BioAssay [4], CHEMBL [5], PubChem Compound [6] and ZINC [7].

Since LBS is also a pattern matching problem, supervised learning algorithms can be used to classify chemical compounds using a database of known drug descriptions as a training dataset. The number of features required to describe each compound also affects the performance of both supervised and unsupervised learning algorithms. This phenomenon is usually addressed as the curse of dimensionality [8]. Two techniques commonly applied to solve this phenomenon are feature extraction and feature selection. While the first one extracts or processes existing features to get a set of new ones, the last one selects a subset of features from the existing ones. This work focuses on observing the performance of two feature selection methods, SVM Recursive Feature Elimination (SVM-RFE) and Wrapper Method (WM), to select a subset of features from Indonesian herbal chemical compounds that react to HIV-1 protease.

II. Related Work

Related research in virtual screening used a method that consists of two phases: First, machine learning-based LBS is used to select potential chemical compound candidates, and second, molecular docking is done with between potential candidates and drug's target [9]. Since molecular docking requires a lot of computational resources, high LBS precision is required to improve efficiency. On the other hand, low recall or sensitivity causes potential candidates excluded [3]. This work shows that the Support Vector Machine (SVM) performs well to classify potential candidates in LBS. Using this as the basis, we explore the usage of feature selections to improve SVM performance in LBS.

The molecular descriptor is a numerical value representing chemical information encoded within a symbolic representation of a molecule. This numerical value can also be obtained by some standardized experiments on a molecule [10]. At least 701 types of molecular descriptors can be extracted from a chemical compound. Therefore, it is difficult to analyze manually all correlations between descriptors [3]. In machine learning-based LBS, not all molecular descriptors directly affect the result of classification. For instance, the number of Bromine (Br) atom is always 0 for every compound in the PubChem BioAssay database. There are even around 500 descriptors behaving in such a way in the same database. Therefore, it is also recommended to reduce the number of features by using techniques like Feature Selection [3].

Feature selection can improve the accuracy of a classification task, and also improves its efficiency by reducing computational costs. On top of that, it can give a better understanding of the resulted model, as suggested by another related research [8]. However, it should also be noted that improvement given by the application of feature selection is depending on the type of data. Hence, the result of its application may vary between datasets[8]. To anticipate this, our experiments use datasets from two different sources: public source (PubChem BioAssay + DUD-E) and Indonesian Herbal DB.

III. Dataset

In order to test the effectiveness and efficiency of the feature selections, two datasets are used. The first one is a combination between extracted molecular descriptors from PubChem BioAssay HIV-1 inhibitor [4] and DUD-E decoy chemical compounds[11]. The second one is built by extracting descriptors from Indonesian Herbal DB, a database of molecular structure from local medicinal plants [12], and labeling each of them based on manual docking results with HIV-1 inhibitor using Autodock [13].

The first dataset consists of 7,330 samples: 3,665 compounds labeled as positive, and 3,665 decoys as unfavorable. The positives come from AID 162030, AID 160444, and AID 83109 compounds which target HIV-1 protease (GI:75593047), which are also used in related research[1]. These compounds are part of the PubChem BioAssay database published by the National Center for Biotechnology Information (NCBI). The negative samples are decoy compounds that do not target the HIV-1 protease. They are part of Database of Useful Decoys - Enhanced (DUD-E) which are provided by Shoichet Laboratory in the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF) [11]. OpenBabel[14] is used to extract Molfile (MOL2) from the original structure-data file (SDF) from PubChem BioAssay and DUD-E, then PaDel Descriptor[10] is used to extract 667 molecular descriptors listed in Table I for each compound. Through these processes, we acquired a balanced dataset for supervised learning.

Our primary dataset is made of molecular descriptors from Indonesian Herbal DB which is produced by Faculty of Pharmacy, Universitas Indonesia [12]. The descriptors are also extracted from the original Molfile (MOL2) with Padel Descriptor using the same configuration as the first dataset to obtain the same set of descriptors. Since Indonesian Herbal DB is just a collection of chemical compounds 3D structural data, docking simulations with HIV-1 need to be done in order to know which of them are positive and negative samples. Docking simulation with Autodock is done for 368 compounds from Indonesian Herbal DB against HIV-1 protein from Protein Data Bank (PDB) where 357 are positive and 11 negatives. The XYZ coordinate used in the docking simulation is 5.192, -4.557, 14.799, the dimension of the grid box is a 50x50x50 unit, and maximum energy evaluation is set to 1,000,000.

TABLE I
Molecule descriptors extracted using Padel Descriptor

Descriptor Type	Number	Class
ALOGP	3	2D
APol	1	2D
Aromaticatomscounts	1	2D
Aromaticbondscount	1	2D
Atomcount	13	2D
Autocorrelation(charge)	5	2D
Autocorrelation(mass)	5	2D
Autocorrelation(polarizability)	5	2D
BCUT	6	2D
Boundcount	5	2D
BPol	1	2D
Carbontypes	9	2D
Chichain	10	2D
Chicluster	8	2D
Chipath	16	2D
Chipathcluster	6	2D
Eccentricconnectivityindex	1	2D
Atomtypeelectrotopologicalstate	482	2D
Fragmentcomplexity	1	2D
Hbondacceptorcount	1	2D
Hbonddonorcount	1	2D
Kappashapeindices	3	2D
Largestchain	1	2D
LargestPisystem	1	2D
Longestaliphaticchain	1	2D
MannholdLogP	1	2D
McGowanvolume	1	2D
Moleculardistanceedge	19	2D
Molecularlinearfreeenergyrelation	6	2D
Petitjeannumber	1	2D
Ringcount	34	2D
Rotatablebondscount	1	2D
Ruleoffive	1	2D
Topologicalpolarsurfacearea	1	2D
Vertexadjacencyinformation(magnitude)	1	2D
Weight	1	2D
Weightedpath	5	2D
Wienernumbers	2	2D
XlogP	1	2D
Zagrebindex	1	2D

IV. Feature Selection

In this work, two feature selection methods are evaluated:

A. Wrapper Method

Wrapper method, as its name suggests, wraps (actual) feature selection, evaluation, and learning algorithms as a black box[15]. In the black box part, it evaluates sets of features generated by the selection algorithm and keeps the best set as a final result. Its architecture is described by Figure 1. In this research, Genetic Algorithm (GA) implementation of DEAP [16] is used as the feature selection and learning algorithm. Linear SVM is used to evaluate every set of features produced by GA based on accuracy score.

B. SVM Recursive Feature Elimination

SVM recursive feature elimination (SVM-RFE) is another feature selection method which initially introduced to choose relevant genes in cancer classification task [17]. SVM-RFE

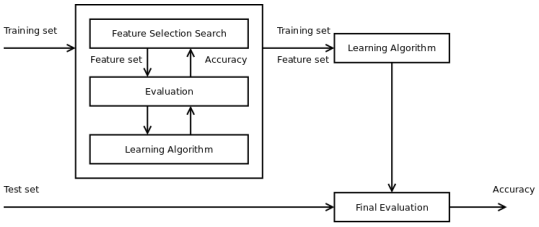

Fig. 1. Wrapper method architecture [15]

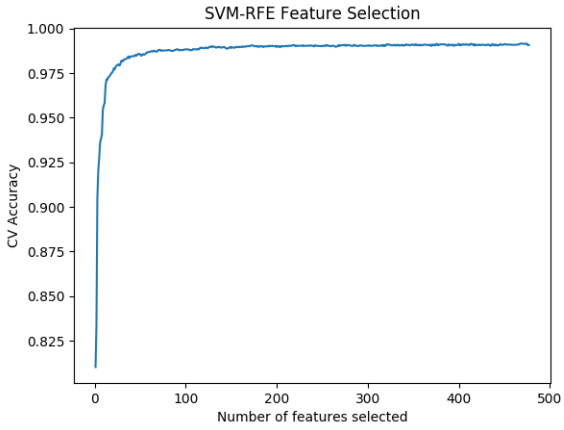

Fig. 2. SVM recursive feature elimination accuracy per feature sets

ranks feature based on their weights in hyperplane and progressively reduces the number of features from the lowest rank. In a nutshell, it does these four steps:

1) Trains linear SVM classifiers using training data
2) Sort features based on their ranks
3) Drop the lowest-ranked feature
4) Repeat steps using remaining features until none left

In this research, Scikit-Learn implementation of SVM-RFE is used [18]. Internally, it uses Linear SVM to rank features for incremental/recursive removal.

V. Experiments

The first dataset, which is a combination of PubChem BioAssay + DUD-E decoys, is used for feature selection using both methods, wrapper method (WM) and SVM recursive feature elimination. For both methods, the accuracy score is used as a metric to determine the performance of features set.

The SVM-RFE method selects a set of 471 features which achieves 0.9915 accuracy score on PubChem BioAssay + DUD-E decoys dataset. Figure 2 shows a chart visualizing relations between the number of features selected by SVM-RFE and their accuracy. It can be observed that the score starts decreasing when the number of selected features is below 50. However, even with only a single highest-ranked feature, the linear SVM achieves accuracy > 0.8.

The wrapper method, which uses a Genetic Algorithm with 100 generations and 20 population per generation as a selection algorithm, chooses only 249 features achieving

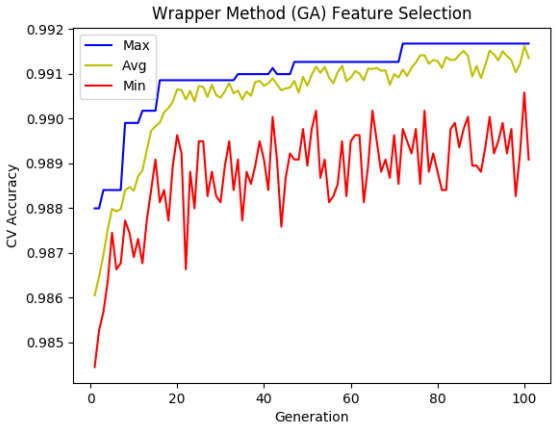

Fig. 3. Wrapper method with Genetic Algorithm accuracy scores

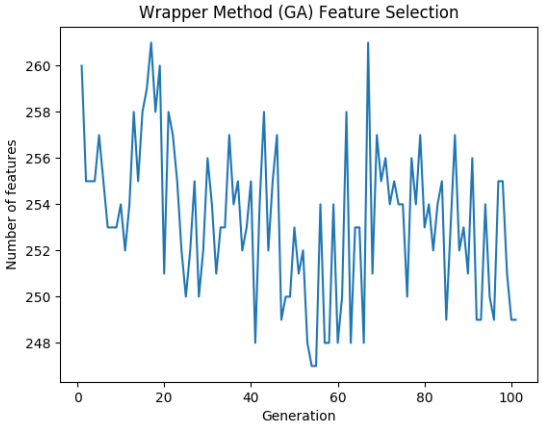

Fig. 4. Wrapper method with Genetic Algorithm number of features

maximum Linear SVM accuracy of 0.9916. There are 244 features in common with the result from SVM-RFE. The chart in Figure 3 shows that even from the first generation, the average accuracy has achieved > 0.9. While Figure 4 shows that every best candidate in each generation use between 240-265 features.

Additionally, the execution time comparison for SVM-RFE and WM is shown by Figure 5. It shows that WM requires a much longer time than SVM-RFE to get its final result. This experiment measures total time to execute each method' script on an Ubuntu 16.04 LTS machine with Core i7 5500U, 8 GB RAM, and 256 SSD storage.

Having obtained two sets of features selected by SVM-RFE and WM, the following experiments aim to compare performance between them and without feature selection. Metrics used in this experiment are based on true/false positive/negative scores: area under the curve (AUC), accuracy, sensitivity, specificity, and precision. The additional metrics are important because unlike the first dataset, the second one has an unbalanced ratio of positive and negative samples.

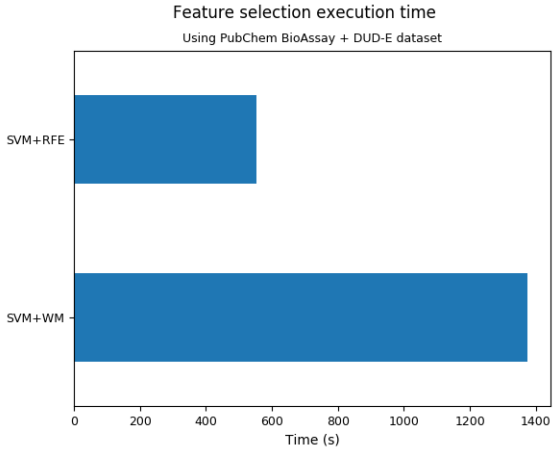

Fig. 5. Feature selection time comparison

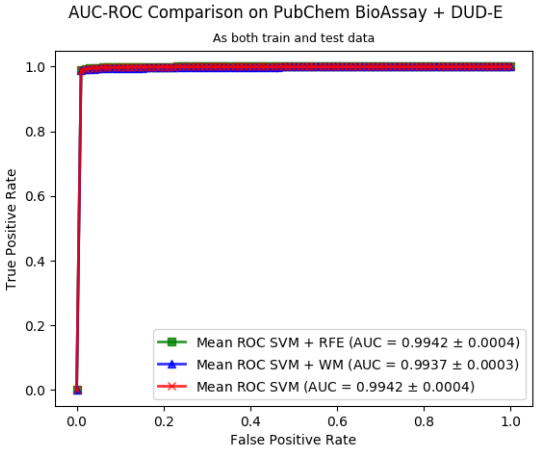

Fig. 6. AUC-ROC curve comparison on PubChem BioAssay + DUD-E dataset

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

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

$$specificity = \frac{TN}{TN + FP} \quad (3)$$

$$precision = \frac{TP}{TP + FP} \quad (4)$$

$$(5)$$

Figure 6 shows the ROC (Receiver Operating Characteristics) curve and AUC score of three models when trained and tested with single PubChem BioAssay + DUD-E dataset (cross-validation): Linear SVM with SVM-RFE, WM, and without feature selection. It shows that there is almost no improvement made by feature selections, as the AUC score of the model without feature selection is the same as the SVM-RFE model, and even slightly higher than the WM model. A similar result is also indicated by the accuracy/sensitivity/precision/specificity chart in Figure 7 where Linear SVM without feature selection is slightly better than both with WM and SVM-RFE.

When the same models trained using the first dataset but tested on second the dataset, which is the manual docking result between Indonesian Herbal DB and HIV-1 protein, the AUC-ROC curve in Figure 9 shows significantly lower performance compared to the previous experiment but also shows no improvement due to feature selection. Although most metrics in Figure 9 reflect the lower performance also shown by the AUC-ROC curve, precision remains relatively high. These results indicate that the problem lies in detecting true negative samples.

To gain a better understanding of the performance, another experiment is done by training and testing the models using only Indonesian Herbal DB dataset. Figure 10 and 11 show better results than experiment that uses PubChem BioAssay +

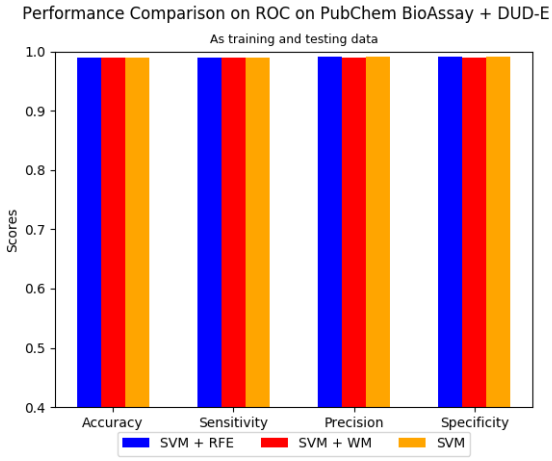

Fig. 7. Classification performance comparison on PubChem BioAssay + DUD-E dataset

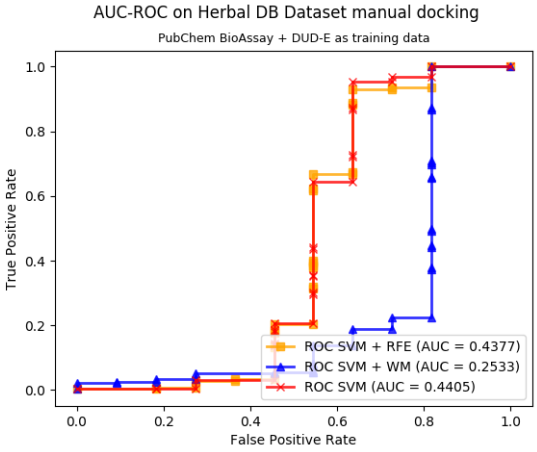

Fig. 8. AUC-ROC curve comparison on Herbal DB dataset

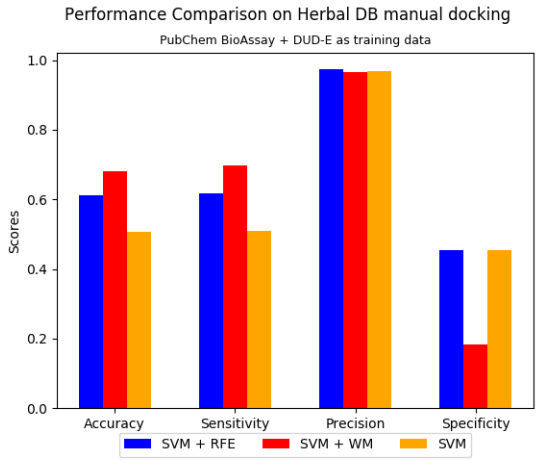

Fig. 9. Classification performance comparison on Herbal DB dataset

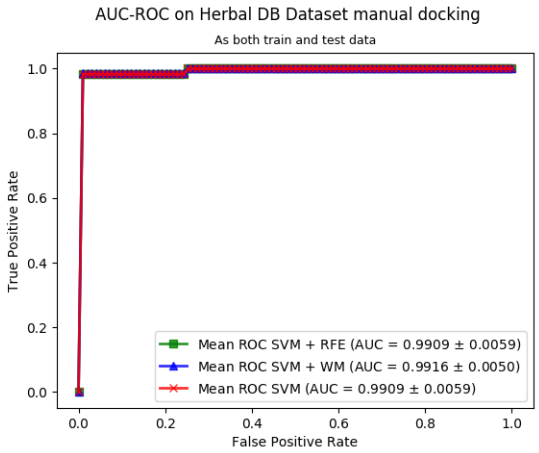

Fig. 10. AUC-ROC curve comparison of the model trained and tested with Herbal DB dataset

DUD-E dataset as training data. However, specificity remains very low achieving only 0.5 for all models.

In the last experiment, the classification time of each model for both datasets is also calculated. As expected, Figure 12 shows that the time required to classify samples is proportional to the number of features used.

VI. Analysis

The experiments show that SVM-RFE and WM feature selections do not improve the performance of Linear SVM classifier for both datasets. However, since there is no significant decrease in performance, they are still useful to improve efficiency by reducing the number of features processed and ultimately the whole classification time by 25%-50%.

Compared to SVM-RFE, WM with Genetic Algorithm requires twice longer time to select features. Despite that, WM manages to choose half of the SVM-RFE features. Since the feature selection process only needs to be done once, and the classification process is done multiple times, WM achieves

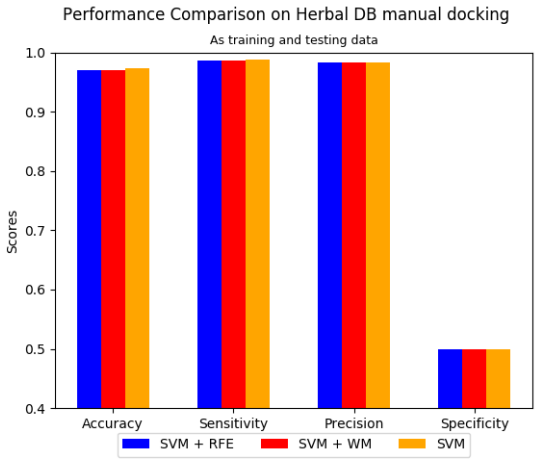

Fig. 11. Classification performance comparison of the model trained and tested with Herbal DB dataset

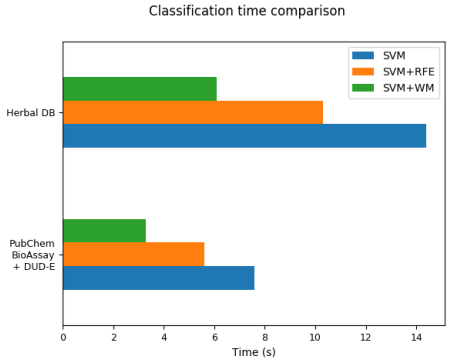

Fig. 12. Classification time comparison

better efficiency than SVM-RFE without sacrificing significant classification performance.

After visualizing both datasets using t-Distributed Stochastic Neighbor Embedding (t-SNE) [19] with 100 perplexities, it becomes clear why Linear SVM classification performance for Indonesian Herbal DB dataset is lower than performance

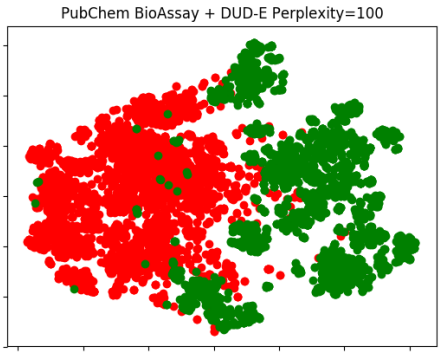

Fig. 13. TSNE visualization of PubChem BioAssay + DUD-E dataset

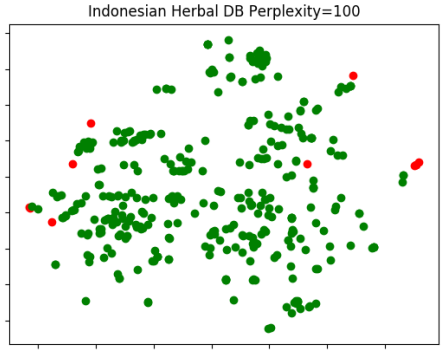

Fig. 14. TSNE visualization of Indonesian Herbal DB dataset

on PubChem BioAssay + DUD-E dataset. Figure 14 shows that positive (green dots) and negative samples (red dots) are visually separable. While in Figure 14, the negative samples are scattered among the positives. Comparison of these two figures indicates that classifying samples in Indonesian Herbal DB is more complicated than in PubChem BioAssay + DUD-E dataset. Most likely, this is caused by the labeling based on manual docking.

VII. Conclusion

Based on experimental results and analysis above, some conclusions are made:

1) Feature selection using SVM Recursive Feature Elimination (SVM-RFE) and Wrapper Method (WM) able to improve the Linear SVM drug target classification efficiency, but not effectiveness. Therefore, they are useful to increase the efficiency of Ligand-Based Screening (LBS).
2) WM using Genetic Algorithm is more suitable than SVM-RFE for molecular descriptors feature selection because it selects almost half the number of features.
3) Indonesian Herbal DB dataset possess different characteristics than PubChem BioAssay + DUD-E dataset. Further collaboration with experts is required to improve the quality of the dataset in the future.

References

[1] A. Yanuar, H. Suhartanto, A. Mun *et al.*, "Virtual screening of indonesian herbal database as hiv-1 protease inhibitor," *Bioinformation*, vol. 10, no. 2, p. 52, 2014.

[2] J. J. F. Chen and D. P. Visco Jr, "Developing an in silico pipeline for faster drug candidate discovery: Virtual high throughput screening with the signature molecular descriptor using support vector machine models," *Chemical Engineering Science*, vol. 159, pp. 31–42, 2017.

[3] S. Korkmaz, G. Zararsiz, and D. Goksuluk, "Drug/nondrug classification using support vector machines with various feature selection strategies," *computer methods and programs in biomedicine*, vol. 117, no. 2, pp. 51–60, 2014.

[4] P. BioAssay, "update wang," 2014.

[5] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey *et al.*, "The chembl bioactivity database: an update," *Nucleic acids research*, vol. 42, no. D1, pp. D1083–D1090, 2014.

[6] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker *et al.*, "Pubchem substance and compound databases," *Nucleic acids research*, vol. 44, no. D1, pp. D1202–D1213, 2015.

[7] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman, "Zinc: a free tool to discover chemistry for biology," *Journal of chemical information and modeling*, vol. 52, no. 7, pp. 1757–1768, 2012.

[8] A. Janecek, W. Gansterer, M. Demel, and G. Ecker, "On the relationship between feature selection and classification accuracy," in *New Challenges for Feature Selection in Data Mining and Knowledge Discovery*, 2008, pp. 90–105.

[9] M. Hilman, "Analisis teknik data mining dan kinerja infrastruktur komputasi cloud sebagai bagian dari sistem perancangan obat terintegrasi," *Graduate Thesis*, 2012.

[10] C. W. Yap, "Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints," *Journal of computational chemistry*, vol. 32, no. 7, pp. 1466–1474, 2011.

[11] M. M. Mysinger, M. Carchia, J. J. Irwin, and B. K. Shoichet, "Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking," *Journal of medicinal chemistry*, vol. 55, no. 14, pp. 6582–6594, 2012.

[12] A. Yanuar, A. Mun'im, A. B. A. Lagho, R. R. Syahdi, M. Rahmat, and H. Suhartanto, "Medicinal plants database and three dimensional structure of the chemical compounds from medicinal plants in indonesia," *arXiv preprint arXiv:1111.7183*, 2011.

[13] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson, "Autodock4 and autodocktools4: Automated docking with selective receptor flexibility," *Journal of computational chemistry*, vol. 30, no. 16, pp. 2785–2791, 2009.

[14] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open babel: An open chemical toolbox," *Journal of cheminformatics*, vol. 3, no. 1, p. 33, 2011.

[15] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data classification: Algorithms and applications*, p. 37, 2014.

[16] F.-A. Fortin, F.-M. De Rainville, M.-A. Gardner, M. Parizeau, and C. Gagné, "DEAP: Evolutionary algorithms made easy," *Journal of Machine Learning Research*, vol. 13, pp. 2171–2175, jul 2012.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[19] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.