A Machine Learning-Based Virtual Screening for Natural Compounds Potential on Inhibiting Acetylcholinesterase in the Treatment of Alzheimer's Disease

Ulfah Nur Azizah[1], *Eri* Dwi Suyanti[1], *Muhammad* Rezki Rasyak[2], *Yekti* Asih Purwestri[1], and *Lisna* Hidayati[1]
[1]Faculty of Biology, Universitas Gadjah Mada, Jl. Teknika Selatan, Sekip Utara, Yogyakarta, 55281
[2]Eijkman Center for Molecular Biology, National Research and Innovation Agency, Jakarta, Indonesia

Abstract. Alzheimer's disease (AD) is a progressive neurodegenerative disease caused by neural cell death, characterized by the overexpression of acetylcholinesterase (AChE) and extracellular deposition of amyloid plaques. Currently, most of the FDA-approved AChE-targeting drugs can only relieve AD symptoms. There is no proven treatment capable to stop AD progression. Many natural products are isolated from several sources and analyzed through preclinical and clinical trials for their neuroprotective effects in preventing and treating AD. Therefore, this study aims to explore and determine potential candidates from natural bioactive compounds and their derivatives for AD treatment targeting AChE. In this study, feature extraction was carried out on 1730 compounds from six plants resulting from literature studies with limitations on international journals with a minimum publication year of 2018 and database searches, then classified using machine learning algorithms: Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM). Hit compounds predicted to be active and inactive in the selected model were then processed through ensemble modelling. From 1730 compounds, there are 986 predicted active compounds and 370 predicted inactive compounds in the LR and RF ensemble modelling. Quercetin, Kaempferol, Luteolin, Limonene, γ-Terpinene, Nerolidol, and Linalool predicted active found overlapping in two to three plants in both LR and RF models.

Keywords: AChE inhibitor, Alzheimer's disease, machine learning, natural bioactive compounds

1 Introduction

Alzheimer's Disease (AD) is a neurodegenerative disorder associated with increasing age and is included in progressive dementia. AD is clinically one of the main causes of dementia which can trigger memory, cognitive, executive dysfunction, and behavioural changes that lead to mental disorders in patients [1]. According to WHO (2022), AD is a global disease contributing to the number of dementia cases by 60-70% compared to other neurodegenerative diseases. The incidence of AD until now continues to increase every year, where every 3 seconds there is 1 person in the world experiencing AD. The prevalence of dementia caused by AD globally reaches 55 million cases and it is estimated that there are 10 million new cases each year. In Indonesia, cases of death due to AD in 2020 reached 27,054 people [2]. The high number of AD cases is influenced by the accumulation and aggregation of β-amyloid in the brain [3].

β-Amyloid is a peptide that accumulates abnormally in brain tissue and forms extracellular plaques that can induce neurodegeneration [4]. AD is formed from the accumulation of Aβ40 and Aβ42 peptides which are the result of an abnormal process of amyloid protein precursors between β-secretase and $\boldsymbol{\gamma}$-secretase and an imbalance in production and synthesis pathways [5]. Several enzymes are known to be involved in increasing neurodegenerative disorders including cholinesterase (Acetylcholinesterase (AChE) and Butyrylcholinesterase (BuChE)), Prolyl endopeptidase (PEP) or oligopeptidase (POP), and the cleavage enzyme APP β (BACE1) [6]. AChE is found mainly in blood and nerve synapses [7]. Therefore, this enzyme is a suitable target for the treatment of AD. AChE is strongly suspected of interacting directly with Aβ plaque formation. This confirms that AChE inhibitors play an essential role in curing AD rather than as a palliative measure [6].

Commercial drugs currently used as AChE inhibitors are tacrine, rivastigmine, and donepezil. However, these drugs have several side effects causing nausea, vomiting, diarrhea, bleeding, and shrinking of brain tissue [8]. Seeing these problems, it is necessary to find alternative natural ingredients that are safer with minimal side effects. As a tropical country, Indonesia has many potential natural ingredients that have the potential to become candidates for anti-Alzheimer's drugs.

Currently, the discovery and development of new drugs for AD treatment required a long time and are quite expensive. It takes between 10 and 15 years of research and testing. Therefore, the approved drug therapies for AD that temporarily relieve the symptoms and slow down the disease progression could only be countered by hand [9]. It is indicated that drug discovery efforts for Alzheimer's treatment still need enhancement [10]. In the traditional discovery of natural chemical compounds, the compounds were isolated randomly, then their biological activity was identified by a simple test. In addition, in the wet lab experimental tests, not all isolated compounds were tested for their therapeutic activity. With the development of information technology, the drug discovery process can be simulated in silico more quickly and accurately through virtual screening. Machine learning-based virtual screening can be an alternative way to select natural product compounds more effectively than compounds that contain the desired activity [11].

Previous research by Periwal *et al.* (2022) used a machine-learning approach in the form of a trained classification model to explore natural compounds and their derivatives more quickly on a much larger scale [10]. In this study, the similarity was predicted between the approved drug and its natural compounds.

Therefore, this study aims to explore and predict Indonesian natural product compounds that can potentially become acetylcholinesterase inhibitors as AD treatment.

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

2 Material and Methods

2.1 Data set and preparation

Indonesian herbal plants that have the potential to have anti-Alzheimer's activity were searched through literature studies with limitations on international journals with a minimum publication year of 2018 (Table 1). The compounds contained in each plant are then searched through the ChEMBL Database (https://www.ebi.ac.uk/chembl/) and KNaPSAcK (http://www.knapsackfamily.com/). Each SMILES ID is taken and saved with Notepad++ in .smi format.
Via DUD-E Docking (https://dude.docking.org/targets) the active and decoy compounds of Acetylcholinesterase (AChE) (Code: ACES) were downloaded. At this stage, the SMILES ID of the active and decoy compounds was stored in Notepad++ in .smi format.

2.2 Environment used

This study needs Java programming language to obtain optimal results through the software that is used. The environment variables are set by adding Java SE Development Kit (JDK) and Java SE Runtime Environment (JRE).

2.3 Feature extraction

Feature extraction is performed on each active and decoy compound with PaDEL Descriptor software using PubChem Fingerprint. At this stage, remove salt, aromaticity detector, and standardize nitro groups are selected.

PubChem has 881 binary structural keys that indicate the presence or absence of a certain group of chemical features in a compound. Compared to other fingerprints that use a floating point number and require 32 bits for one feature, PubChem fingerprint only requires one bit of storage for each feature in the compound. The small bit of fingerprint can speed up the machine learning process. PubChem Fingerprint uses the 2D structure of the compound which is used as a measure of the similarity of the compound to the compounds that have been found on the website http://pubchem.ncbi.nlm.nih.gov.

The fingerprint results of the active and decoy compounds are then combined into one big data in .csv format. Each natural product compound is given a class label. The active compound is labeled 1, while the decoy compound is labeled 0.

2.4 Machine Learning

Machine learning (ML) is a subfield of Artificial Intelligence (AI) that focuses on developing algorithms and statistical models that enable computers to learn and improve their performance on specific problems without being explicitly programmed [11]. Machine learning has become an increasingly valuable tool in research due to its ability to analyze large and complex datasets, identify patterns, and make predictions.

This article utilized a machine-learning approach to extract compounds from the data obtained from ChEMBL and KNaPSaCK databases. This approach has developed a set of algorithms that have been optimized for performance and accuracy using advanced computational techniques. These algorithms can process large volumes of data efficiently and extract valuable information from it.

In this study, a machine learning approach was used with supervised learning. The selected algorithms include Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM). All three have different ways of classifying objects (compounds) into their classes. RF is a classifier that consists of a collection of tree-structured classifiers. RF uses multiple trees to average (regression) or calculates the most votes (classification) in the terminal leaf nodes to make a prediction. In decision trees, each node is separated by the best separation among all variables. While in RF, each node is divided by the best among the predictor subsets chosen randomly at that node. On decision trees, at each decision node, the features are split into two branches and are repeated until the leaf nodes are reached to make the final prediction. compared to RF, decision trees can't generalize well the unseen data and are notorious for overfitting. RF is able to overcome overfitting by using a decision tress ensemble where the values are random and independent. RF is suitable for medium to large datasets. In RF, not all predictor variables are used at once so that when the number of independent variables is greater than the number of observations, this algorithm can run, unlike the LR algorithm which will not run because the predicted parameters exceed the number of observations [12].

Logistic Regression (LR) is one of the most commonly utilized linear statistical models in which the response variable is quantitative. The response variable in LR is in the form of a log of the possibilities that are classified in group I of binary responses or multi-class responses. LR makes several assumptions such as independence, the responses (logits) at each subpopulation level of the variables are normally distributed, and the constant variance between the responses and all explanatory value variables. a transformation to the variable is applied over the output classes between 0 and 1, called "logistic" or "sigmoid". LR has a vulnerability to underfitting and has low accuracy [13]. In addition, the LR algorithm also has problems with class imbalance in datasets with high dimensions [14].

Support Vector Machine (SVM) is a technique for finding hyperplanes that can separate two data sets from two different classes. SVM has the basic principle of a linear classifier. Even so, SVM can also work on non-linear problems through the kernel concept in high-dimensional space. In a high-dimensional space, we will look for a hyperplane that can maximize the margin between data classes. In addition, SVM uses structural risk minimization (SRM), which is the inductive principle of nature having a learning model from a limited set of training data. The advantage of this model is that it is able to determine the distance to the support vector so that the computational process runs faster.

The machine-learning approach was conducted on Lenovo Ideapad Slim 3i, Intel Core i3-14IGL05, RAM 5 GB, SSD 256 GB. Analysis of the machine learning approach was carried out using the Orange Data Mining application with 75% fixed proportion data and 5- cross-validation. The results obtained are then evaluated based on ROC Analysis and Confusion Matrix to select the algorithm to be used for the next process (Figure 1).

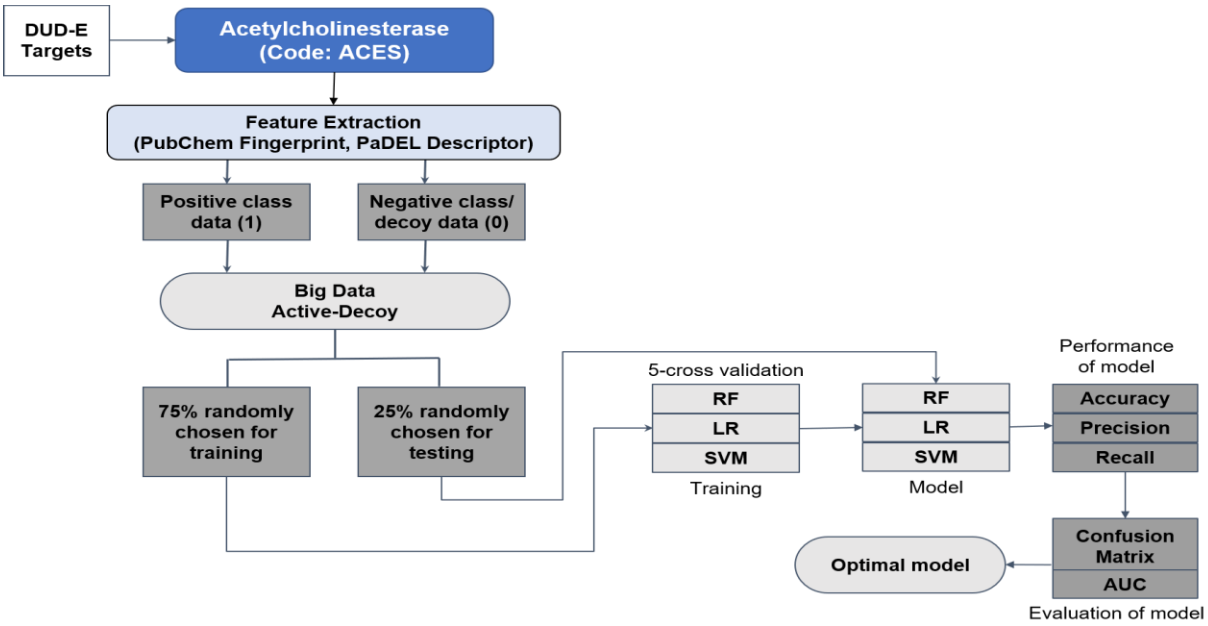


Fig. 1. A schematic representation of the machine learning approach to choose the optimal and validated model for predicting natural compounds

2.5 Predicting Indonesian Herbal Compounds

Prediction of Indonesian medicinal compounds is investigated with selected models. The natural ingredient compounds from the 6 selected plants do not have labels. Therefore, each plant compound is predicted to be active and decoy. From the classification results obtained, an ensemble was performed on compounds that were predicted to be active and inactive in both models (Figure 2).

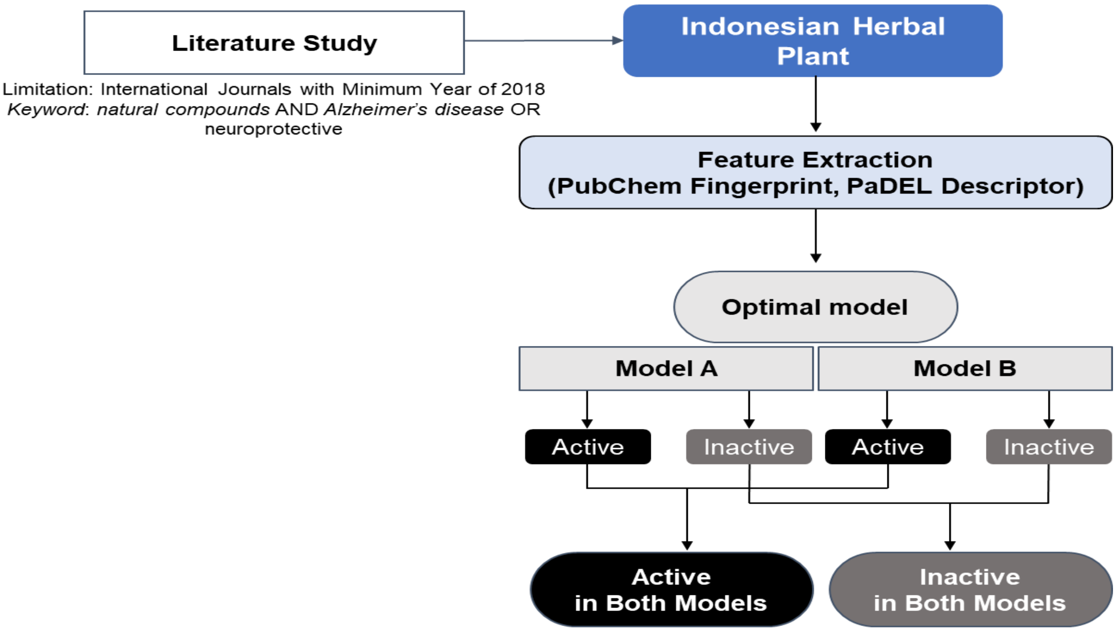

Literature Study
Limitation: International Journals with Minimum Year of 2018
Keyword: *natural compounds* AND *Alzheimer's disease* OR neuroprotective
Indonesian Herbal Plant
Feature Extraction (PubChem Fingerprint, PaDEL Descriptor)
Optimal model
Model A
Model B
Active
Inactive
Active
Inactive
Active in Both Models
Inactive in Both Models

Fig. 2. The scheme of predicting natural compounds using the optimal and validated model

2.6 Pharmacokinetic Analysis

The pharmacokinetic tests of the compounds included the HIA and toxicity (BBB) test through PreADMET (https://preadmet.webservice.bmdrc.org/adme/) and the Lipinski Rule of Five (Ro5) test through SwissADME (http://www.swissadme.ch/).
In addition, Predictions of Activity Spectra for Substances (PASS) were carried out via the http://www.pharmaexpert.ru/passonline/predict.php page. This test was carried out to determine the activity potential of a compound based on the relationship between the structure of the compound and its biological activity (Structure Activity Relationship/SAR).

3 Results and Discussion

3.1 Data source

Based on the results of a literature review search, there are 6 candidate plants that had the potential as anti-Alzheimer's (Tabel 1). Based on ChEMBL and KNaPSAcK, a total of 1730 compounds were obtained from 6 selected plants. There are 14 compounds in *Moringa oleifera*, 31 compounds in *Zingiber officinale*, 43 compounds in *Allium sativum*, 1292 compounds in *Annona crassiflora*, 142 compounds in *Citrus aurantium*, and 208 compounds in *Annona muricata*.

Table 1. List of Indonesian herbal compound related to anti-Alzheimer.

Plants	Bioactivity	Reference
Moringa oleifera	Inhibitor of butyrylcholinesterase (BChE); inhibitor of acetylcholinesterase (AChE); lower the glycemic index, total cholesterol, triglycerides, and low-density lipoprotein cholesterol (LDL-C) level; increase high-density lipoprotein cholesterol (HDL-C) in plasma	[15]
Zingiber officinale	Antioxidant, anti-inflammatory, increase expression of nerve growth factor (NGF)	[16]
Allium sativum	Antioxidant, neuroprotective agent, inhibitor of AChE and BChE enzymes, anti-neuroinflammatory	[17]
Annona crassiflora	Antibacterial, antimutagenic, anti-inflammatory, antinociceptive, hepatoprotective, and antitumoral, inhibitor of AChE	[18]
Citrus aurantium	Anti-oxidative, antihypertensive, anti-hyperlipidemia, anti-diabetic, anti-inflammatory, and hepato-protective potentials, neuroprotective	[19]
Annona muricata	Anticancer, antioxidant, antiviral, anti-haemolytic, sedative and neuroprotective	[20]

3.2 Machine learning

Through the machine learning approach, we conducted each model with 75% fixed proportion and 5-cross validation to obtain the optimal prediction model. The performance prediction model calculated using the validation dataset is shown in Figure 3.

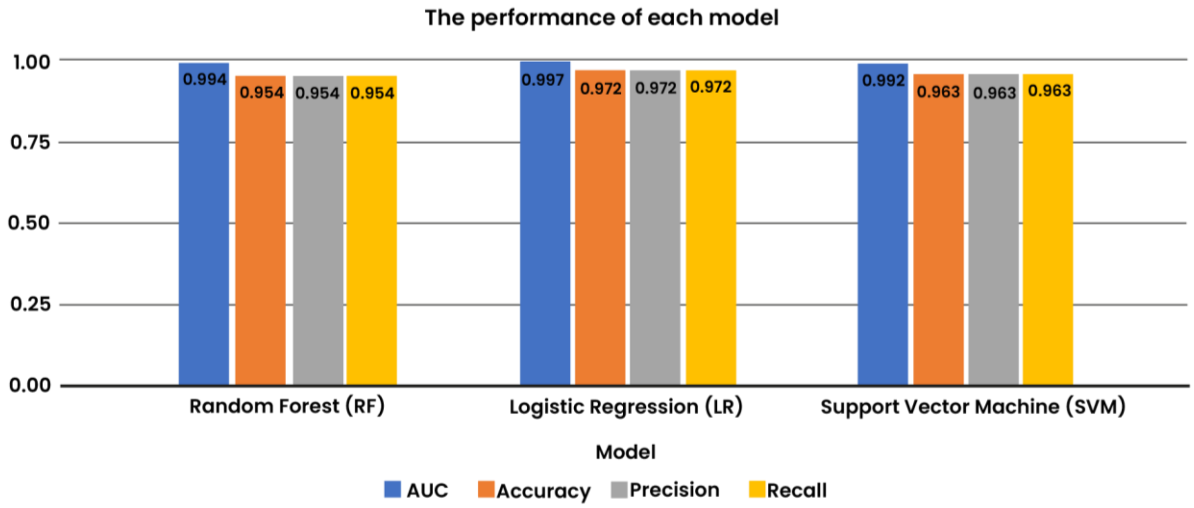


Fig. 3. The performance prediction of each model

The performance prediction model contains AUC, accuracy, precision, and recall. Area Under Curve (AUC) is the area under the ROC curve. If the value is close to 1, it means that the model obtained is more accurate. Accuracy is true positive and true negative divided by a total number of positive and negative. Precision is the number of true positive divided by the total number of positive predictions. Recall is the number of true positive divided by the total number of true positive and false negative [21]. Both precision and recall are focusing on positive examples and prediction.

Based on Figure 3, the LR model has the highest accuracy value (0.972), followed by SVM (0.963) and RF (0.954), likewise with the value of precision and recall. The LR model also has the highest AUC value. The higher the AUC value, the better the model performance in distinguishing positive and negative classes.

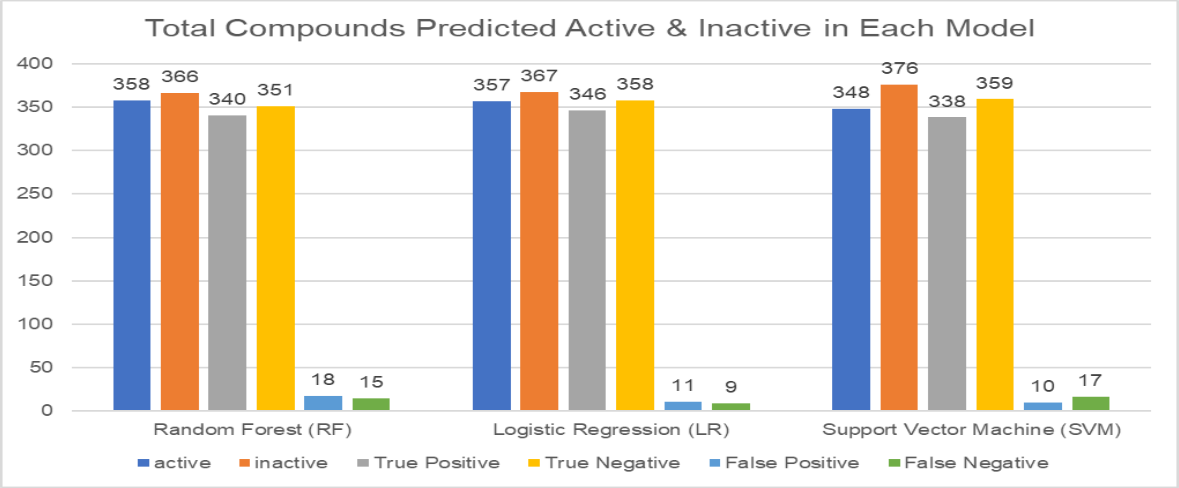


Fig 4. Total compounds predicted active and inactive in each model

The confusion matrix has four categories. They are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positive (TP) means that the actual data is predicted correctly as positive. True Negative (TN) means that the actual data is predicted correctly as negative. Sometimes, the system can make a mistake by predicting the actual negative value as a positive value (False Positive/FP) or the actual positive value is predicted as a negative value (False Negative) [21].

As mentioned in Figure 4, the LR model has the highest TP value (346) and the SVM model has the lowest TP value (338).

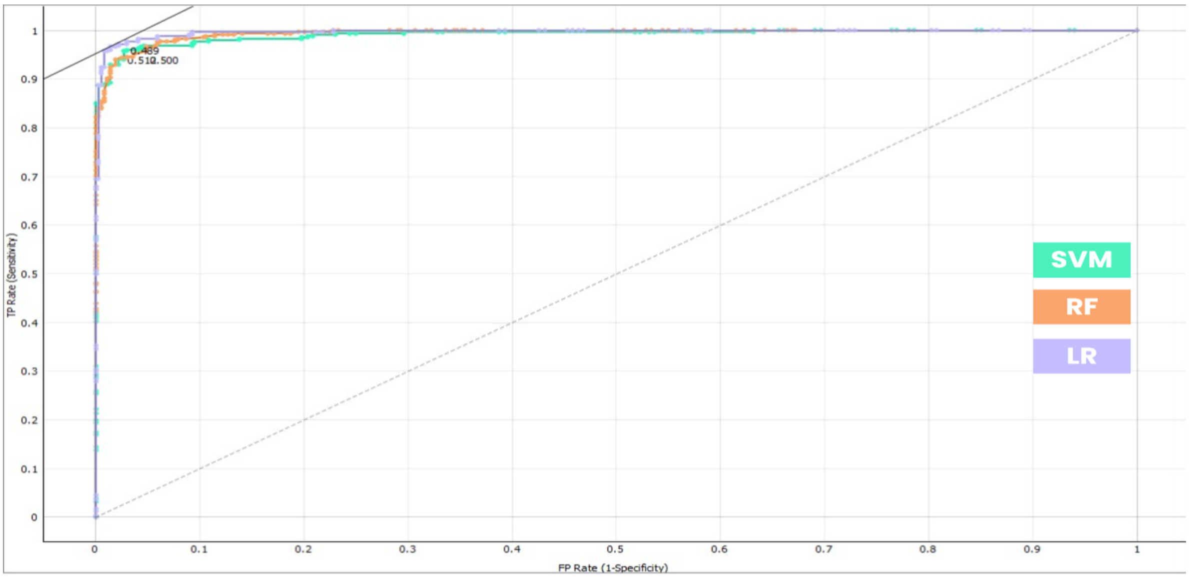


Fig. 5. ROC curve of each model

Provos *et al.* (1998) have argued that the results can be misleading if we look only at the accuracy results. When evaluating binary decision problems, it is recommended to use Receiver Operating Characteristic (ROC) [22]. ROC is the cross-validated performance measurement for the classification model [23]. ROC curve is fixed. An ROC curve begins at the (0,0) coordinate which is the decision threshold at which all test results are negative, and forms diagonal line ending at the (1,1) coordinate which is the decision threshold at which all test results are positive. This diagonal line called the “chance diagonal” and represents the ROC curve of a prediction test with no ability to distinguish positive and negative results [22].

ROC curve has False Positive Rate (FPR/specificity/recall) on the *x*-axis and True Positive Rate (TPR/sensitivity/precision) on the *y*-axis. The FPR measures wrongly the actual negative value as a positive value. The TPR measures the actual positive value correctly as a positive value. A significant change in the number of false positives just makes a small change in the FPR in ROC analysis [22].

Figure 5 shows that at a specificity of 0–0.05, the LR model has the highest sensitivity, followed by SVM, then RF. At a specificity of 0.05–0.1, the LR model has the highest sensitivity, followed by RF, then SVM. The higher the sensitivity, the better the positive class is classified correctly. The discrimination has a better particular model if the curve is more convex and approaches the upper left corner [24]. While looking at Figure 5, the LR model appears to be fairly close to optimal.

Compared to the confusion matrix, ROC has several advantages. ROC can be extended to multi-class through one vs. one (OVO) or one vs. all (OVA) methodologies. ROC depends on TPR and FPR which are calculated against true positive and true negative independently, so the change of class distribution will not impact the ROC curve. Obuchowski (2004) states that for classifying the results, the basic measures of accuracy require a decision rule or positivity threshold. ROC curves do not depend on the decision threshold although constructed from sensitivity and specificity [25].

To obtain the optimal results, here ensemble model is used. The ensemble model refers to combining multiple models into one. Based on the ROC curve, the LR and RF models are chosen here to predict the active and inactive compounds of the six plants above. This is intended to minimize the bias that can occur when only referring to one model [26].

3.3 Predicting Indonesian Herbal Compounds

Compounds of each plant are predicted using LR and RF models by its fingerprint. The results can be shown in Figure 6.

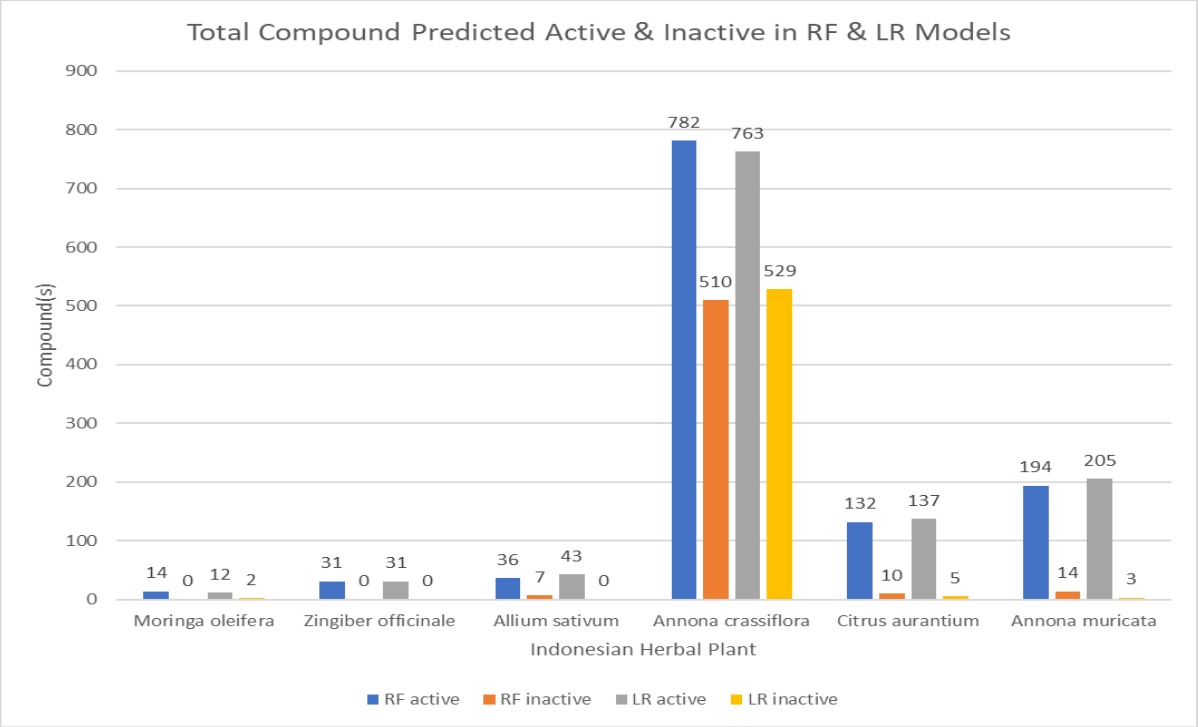


Fig 6. Total compounds predicted active and inactive in LR and RF models

Based on Figure 6, all compounds of *Moringa oleifera* are predicted as active compounds in the RF model, but there are only 12 compounds predicted as active compounds in the LR model. All compounds of *Zingiber officinale* are predicted as active compounds in both RF and LR models. All compounds of *Allium sativum* are predicted as active compounds in LR model, but there are only 36 compounds predicted as active compounds in the RF model. In *Annona crassiflora,* there are 782 predicted active compounds in the RF model and 763 compounds predicted active compounds in the LR model. In *Citrus aurantium*, there are 132 active compounds in the RF model and 137 active compounds in the LR model. In *Annona muricata,* there are 194 active compounds in the RF model and 205 active compounds in the LR model.

Compounds that are predicted active in both LR and RF models then are collected. From 1730 compounds, there are 986 predicted active compounds and 370 predicted inactive compounds in the LR and RF ensemble modeling. From these compounds, there are lots of predicted active compounds found overlapping in two to three plants in both LR and RF models (Table 2).

Table 2. List of Indonesian herbal compound related to anti-Alzheimer.

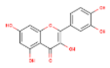
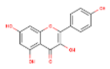
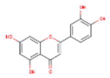
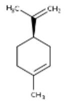
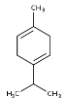
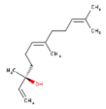
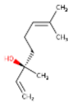
Compounds	Molecular Formula	SMILES ID	Plants
Quercetin 	C15H10O7	c1(cc(c2c(c1)oc(c(c2=O)O)c1ccc(c(c1)O)O)O)O	*Allium sativum, Annona muricata,* and *Citrus aurantium.*
Kaempferol 	C15H10O6	c1(cc(c2c(c1)oc(c(c2=O)O)c1ccc(cc1)O)O)O	*Allium sativum* and *Annona muricata*
Luteolin 	C15H10O6	c1(cc(c2c(c1)oc(cc2=O)c1cc(c(cc1)O)O)O)O	*Annona muricata* and *Citrus aurantium*
Limonene 	C10H16	C1=C(CC[C@H](C1)C(=C)C)C	*Annona muricata* and *Citrus aurantium*
γ-Terpinene 	C10H16	C1C=C(CC=C1C)C(C)C	*Allium sativum* and *Citrus aurantium*
Nerolidol 	C15H26O	CC(=CCC/C(=C/CC[C@](C=C)(C)O)/C)C	*Allium sativum* and *Citrus aurantium*
Linalool 	C10H18O	CC(=CCC[C@](C=C)(O)C)C	*Annona muricata* and *Citrus aurantium*

Table 3. Toxicity and HIA Analysis

Compounds	HIA (%)	BBB
Quercetin	77.21	-
Kaempferol	79.44	0.286076
Luteolin	79.43	0.367582
Limonene	100.00	8.27823
γ-Terpinene	100.00	8.03745
Nerolidol	100.00	13.9838
Linalool	100.00	6.12506
Donepezil	97.95	0.187923

3.4 Pharmacokinetics Analysis

The seven overlapping compounds were then further analyzed to determine the HIA value and their toxicity through PreADMET (Table 3).

Based on the predicted results of Human Intestinal Absorption (HIA) in Table 3, it shows that the Quercetin, Kaempferol, and Luteolin compounds have a value below 70%. While Limonene, γ-Terpinene, Nerolidol, and Linalool have HIA values of 100%. HIA in the range of 80% -100% indicates that the compound has good absorption in the intestinal wall [27].

The Blood Brain Barrier (BBB) values for the compounds shown in Table 3 obtained the highest to lowest BBB values respectively Nerolidol (13.9838), Limonene (8.27823), γ-Terpinene (8.27823), Linalool (6.12506), Luteolin (0.367582), and Kaempferol (0.286076). All compounds have a BBB value higher than Donepezil as a positive control. The BBB value indicates the absorption of the compound into the blood-brain barrier [27].

Compounds in Table 2 are then analyzed further with the Lipinski Rule of Five (Ro5) test to determine the similarity of drugs or chemical compounds with their pharmacological properties. Several parameters considered in the Lipinski test include molecular weight (< 500 Da), lipophilicity (iLOGP, XLOGP3, WLOGP, MLOGP, SILICOS-IT) (< 5), number of donor hydrogen atoms (< 5, number of OH and NH), and the number of acceptor hydrogen atoms (< 10, the number of N & O) [28].

Table 4. Lipinski Rule of Five of Selected Compounds.

Compounds	Molecular Weight (g/mol)	H-bond acceptors	H-bond donors	LogP
Quercetin	302.24	7	5	1.63
Kaempferol	286.24	6	4	1.70
Luteolin	286.24	6	4	1.86
Limonene	136.23	0	0	2.72
γ-Terpinene	136.23	0	0	2.73
Nerolidol	222.37	1	1	3.64
Linalool	154.25	1	1	2.71
Donepezil	379.49	4	0	4.00

Based on Lipinski test in Tabel 4, it shows that all compounds have no violation and have fulfilled the Lipinski Rule of Five. The molecular weight of all compounds is lighter than donepezil as a positive control. Compounds that have a molecular weight of less than 500 Da, indicate that the compounds are orally active. If the molecular weight of a compound is high, the permeability of the compound in the intestine and central nervous system is lower [29].

The hydrogen-bond acceptor of all compounds has compiled the Lipinski Rule of Five (H-bond acceptor $\leq$ 10), which means that all compounds bind well with solvents such as water. The hydrogen-bond donor of all compounds has compiled the Lipinski Rule of Five (H-bond donor $\leq$ 5), which means that all compounds have the ability to penetrate the bilayer membrane. The higher the hydrogen bonding capacity of a compound/molecule, the higher the energy required for the absorption process [30].

The lipophilicity values of all compounds showed a value of no more than 5 and complied with the Lipinski rule. All compounds have lower lipophilicity than donepezil. The lipophilicity value indicates the degree of absorption of the compound and is the algorithm of the ratio of the drug partitioning to the organic phase which is in the aqueous phase. The positive lipophilicity values indicate that the compounds easily penetrate the lipid bilayer membrane [29].

3.5 PASS (Prediction of Activity Spectra for Substances) Analysis

The PASS test is performed to predict pharmacological effects, types of biological activity, mechanism of action, and specific toxicity of different chemical compounds. The predicted activity spectrum in PASS is presented by a list of activities with the probability of active (Pa) and being active inactive (Pi). Pa value > 0.7 indicates that the compound is very likely to show activity in the experiment. A Pa value of 0.5 < Pa < 0.7 indicates that the compound is likely to show activity in experiments, but is less likely and unlike any known pharmaceutical agent. Then, if the Pa value <0.5 then the compound may not show activity in the experiment [31].

The PASS web server predicts various biological activities of the compounds, but the focus of the research here is on the prediction of the Acetylcholine neuromuscular blocking agent.

Table 5. Biological Activity of Each Compound Related to Alzheimer's Disease.

Compounds	Acetylcholine neuromuscular blocking agent	CYP3A4 substrate	CYP2D6 substrate	CYP3A4 inhibitor
Quercetin	0.512	0.617	0.41	0.294
Kaempferol	0.545	0.623	0.425	0.275
Luteolin	0.57	0.567	0.429	0.223
Limonene	0.743	0.309	0.254	-
γ-Terpinene	0.556	0.383	0.258	-
Nerolidol	0.261	-	0.353	-
Linalool	0.34	0.198	0.314	-
Donepezil	0.561	0.231	0.188	-

Several cholinesterase inhibitors currently used in the treatment of Alzheimer's disease are metabolized via CYP-related enzymes [32]. This drug can interact with many other drugs that are substrates, inhibitors or inducers of the CYP system. Some cholinesterase inhibitors (tacrine, donepezil, galantamine) are metabolized via CYP-related enzymes, especially CYP2D6, CYP3A4, and CYP1A2 [33].

Based on Table 5, Quercetin, Kaempferol, Luteolin, and γ-Terpinene are likely to show activity in experiments, but are less likely and unlike any known pharmaceutical agent ($0.5 < Pa < 0.7$) of Acetylcholine neuromuscular blocking agent, same as Donepezil. Only Limonene is very likely to show activity in the experiment to this biological activity. When looking at the CYP system, Quercetin, Kaempferol, and Luteolin are also likely to show activity in experiments, but are less likely and unlike any known pharmaceutical agent ($0.5 < Pa < 0.7$) for CYP3A4 substrate. For CYP2D6 substrate, all compounds do not show activity in the experiment.

From a drug metabolism standpoint, when a compound acts as a substrate the bioavailability will be reduced due to these compounds will be metabolized into relatively more polar compounds to make it easier excreted. However, if compounds play a role as an inhibitor that inhibits the action of cytochromes P450, then the compound causes bioavailability of other compounds will increase that can cause toxicity [34].

4 Conclusions

In this study, a machine learning approach was used to look for candidate compounds that have the potential to become acetylcholinesterase inhibitors for the treatment of Alzheimer's disease. Analysis through machine learning and evaluation through the ROC curve shows that the RF and LR models are optimal models in differentiating active and inactive classes. From 1730 compounds, there are 986 predicted active compounds and 370 predicted inactive compounds in the LR and RF ensemble modeling. Quercetin, Kaempferol, Luteolin, Limonene, γ-Terpinene, Nerolidol, and Linalool predicted active found overlapping in two to three plants in both LR and RF models. Based on the results of pharmacokinetic analysis, Limonene, γ-Terpinene, Nerolidol, and Linalool showed optimal results. Nevertheless, further research both in silico and in vitro study still needs to be developed.

Acknowledgement

The author thanks to MBKM research grant from Faculty of Biology Universitas Gadjah Mada.

References

1. L. Xue., F. Xiaojin, S. Xiaodong, H. Ningning, H. Fang, L. Yongping. Global, regional, and national burden of Alzheimer's disease and other dementias, 1990-2019, Frontiers in Aging Neuroscience (2022)
2. World Health Organzation. Dementia. Accessed 2 June 2023. URL: www.who.int/news-room/fact-sheets/detail/dementia (2022)
3. S. Xiaojuan, C. Wei-Dong, W. Yan-Dong. β-Amyloid: the key peptide in the pathogenesis of Alzheimer's disease, Frontiers in Pharmacology **6,** 221 (2015)
4. T. Elena., G. Michela., V. Vasciaveo, T. Massimo. Oxidative stress and beta amyloid in Alzhemimer's disease, which comes first: the chicken or the egg?, Antioxidants **10**, 9 (2021)
5. D. Ture, A. Michael, D. Dennis. The neuropathological diagnosis of Alzheimer's disease, Molecular Neurodegeneration **14,** 32 (2015)

6. S. Tomas, A. Marketa, O. Lubomir, C. Lucie., J. Daniel, H. Martina, K. Jiri, C. Jakub. Cholinesterase and prolyl oligopeptidase inhibitory activities of alkaloids from Argemone platyceras (Papaveraceae), Molecules **22**, 1181 (2017)
7. B., Buket, U. Duygu, N. Nurlu, K. Gulen, U. Nehir. Chemical profile, acetylcholinesterase, butyrylcholinesterase, and prolyl oligopeptidase inhibitory activity of Gaucium corniculatum subsp. Refractum, Brazilian Journal of Pharmaceutical Sciences **58,** (2022)
8. G. Marucci, M. Buccioni, D. Ben, C. Lambertucci, R. Volpini, F. Amenta. Efficacy of acetylcholinesterase inhibitors in Alzheimer's disease, Nuropharmacology, **190,** (2021)
9. P. Long, P. Quan. Virtual screening strategies in drug discovery: A brief overview, Vietnam Journal of Science and Technology **59**, 4 (2021)
10. V. Periwal, S. Bassler, S. Andrejev, N. Gabrielli, K. Patil, A. Typas, K. Patil. Bioactivity assessment of natural compounds using machine learning models trained on target similarity between drugs, PLOS Computational Biology **18**, 4 (2022)
11. I. Fernandez, J. Peters. Machine learning and deep learning in medicine and neuroimaging, Annals of the Child Neurology Society **1,** 2 (2023)
12. A. Liaw, M. Wiener. Classification and regression by random forest, R News **2**, 3 (2002)
13. Harrington, P. Machine Learning in Action. Manning Publications Co. (2012)
14. C. Lin, R. Weng, S. Keerthi. Trust Region Newton Method for Large-Scale Logistic Regression, Journal of Machine Learning Research. (2008)
15. A. Ademosun, G. Oboh, O. Ajeigbe. Influence of Moringa (Moringa oleifera) enriched ice creams on rats' brain: Exploring the redox and cholinergic systems, Current Research in Food Science **5**, (2002)
16. R. Arcusa, D. Villano, J. Marhuenda, M. Cano, B. Cerda, P. Zafrilla. Potential Role of Ginger (Zingiber officinale Roscoe) in the Prevention of Neurodegenerative Diseases, Front. Nutr. **9**, (2022)
17. P. Tedeschi, M. Nigro, A. Travagli, M. Catani, A. Cavazzini, S. Merighi, S. Gessi. Therapeutic Potential of Allicin and Aged Garlic Extract in Alzheimer's Disease, Int. J. Mol. Sci. **23**, 6950 (2022)
18. M. Barbosa, A. Justino, M. Martins, K. Belaz, F. Ferreira, R. de Oliveira, A. Danuello, F. Espindola, M. Pivatto. Cholinesterase inhibitors assessment of aporphine alkaloids from Annona crassiflora and molecular docking studies, Bioorganic Chemistry **120**, (2022)
19. P. Shayan, A. Amir. Evaluation of antioxidant and inhibitory properties of Citrus aurantium L. on the acetylcholinesterase activity and the production of amyloid nano–bio fibrils, International Journal of Biological Macromolecules **182,** (2021)
20. W. Kim, Y. Kim, E. Cho, E. Byun, W. Park, H. Song, K. Kim, S. Park, E. Byun. Neuroprotective effect of Annona muricata-derived polysaccharides in neuronal HT22 cell damage induced by hydrogen peroxide, Bioscience, Biotechnology, and Biochemistry **84**, (2020)
21. D. Chicco, G, Jurman. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics. (2020)
22. J. Davis, M. Goadrich. The relationship between precision-recall and ROC curves. (2006)
23. P. Ananat, P. Gupta. Application of machine learning in understanding bioactivity of beta-lactamase AmpC, Journal of Physics: Conference Series **2273**, (2022)
24. K. Gajowniczek, T. Zabkowski. Estimating the ROC curve and its significance for classification models' assessment, Quantitative Methods in Economics **15**, 2 (2014)
25. N. Obuchowski. Fundamental of Clinical Research for Radiologists, American Journal of Roentgenology **184**, (2005)
26. T. Mahesh, D. Kumar, V. Kumar, J. Asghar, B. Bazezew, R. Natarajan, V. Vivek. Blended ensemble learning prediction model for strengthening diagnosis and treatment of chronic diabetes disease, Computational Intelligence and Neuroscience (2022)
27. T. Hou. J. Wang, Y. LiADME evaluation in drug discovery: the prediction of human intestinal absorption by a support vector machine, J. Chem. Inf. Model **47**, 6 (2007)
28. D. Sen, K. Nandi, D. Saha. Rule of five: The five men army to cross the blood brain barrier for therapeutically potent, World Journal of Advance Healthcare Research **5**, 3 (2021)
29. M. PollastriOverview on the rule of five, Current Protocols in Pharmacology **49,** (2010)
30. T. Altamash, A. Amhamed, S. Aparicio, M. Atilhan. Effect of hydrogen bond donors and acceptors on CO2 absorption by deep eutectic solvents, Processes **8,** (2020)
31. M. Basanagouda, J. Jadhav, M. Kulkarni, R. Rao. Computer Aided Prediction of Biological Activity Spectra: Study of Correlation between Predicted and Observed Activities for Coumarin-4-Acetic Acids, Indian J Pharm Sci. **73**, 1 (2011)
32. R. Cacabelos, R. Llovo, C. Frail, L. Fernández-Novoa L. Pharmacogenetic aspects of therapy with cholinesterase inhibitors: the role of CYP2D6 in Alzheimer's disease pharmacogenetics, Curr Alzheimer Res. **4,** 4 (2007)
33. S. Ruangritchankul, P. Chantharit, S. Srisuma, L. Gray. Adverse Drug Reactions of Acetylcholinesterase Inhibitors in Older People Living with Dementia: A Comprehensive Literature Review, Ther Clin Risk Manag **17,** (2021)

34. J. Hakkola, J. Hukkanen, M. Turpeinen, O. Pelkonen. Inhibition and induction of CYP enzymes in humans: an update, Springer Science and Business Media Deutschland GmbH **94,** (2020)