Classification of IGF1R ligand compounds for identification of herbal extracts using extreme gradient boosting

Mohammad Hamim Zajuli Al Faroby[a,1,*], Siti Amiroch[b,2], Bernadus Anggo Seno Aji[c,3], Avriono Aritonang[a,4]

[a,] Department of Data Science, Faculty of Information Technology and Business, Institut Teknologi Telkom Surabaya, Surabaya, Indonesia.
[b] Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Islam Darul 'Ulum, Lamongan, Indonesia.
[c] Department of Information Technology, Faculty of Information Technology and Business, Institut Teknologi Telkom Surabaya, Surabaya, Indonesia.
[1] alfaroby@ittelkom-sby.ac.id*; [2] siti.amiroch@unisda.ac.id; [3] bernadus.seno@ittelkom-sby.ac.id; [4]avriono.aritonang@student.ittelkom-sby.ac.id
* Corresponding Author

Received 9 February 2022; accepted 6 July 2022; published 6 September 2022

ABSTRACT

Diabetes Mellitus is a serious disease that requires serious treatment. The cause of this disease is due to malfunctions in insulin and insulin-producing organs. One of the proteins that become insulin signaling receptors is IGF1R, which has an important role in activating and maximizing insulin performance. In this study, we aimed to obtain herbal compounds that can activate the function of the IGF1R protein by utilizing compound data in an open database and modeling it using the ensemble method, namely extreme gradient boosting. We found that this method produces the best classification model than with other algorithms. We predicted 844 data for herbal compounds, but only 15 data met the threshold of 0.6. We got one plant from the fifteen herbal compounds, namely Zostera Marine, which was confirmed to have compounds that bind to IGF1R. These compounds have the highest probability value in the classification model that we formed compared to others.

KEYWORDS
Molecular Fingerprint
Extreme Gradient Boosting
Machine Learning
Herbal Compound
IGF1R

1. Introduction

Diabetes Mellitus is a serious problem in the human body's metabolic system. Diabetes Mellitus can be inherited genetically or due to unhealthy lifestyle conditions [1]. Based on the origin of the cause of Diabetes Mellitus, this disease is divided into three types, namely Type-1 Diabetes, which is a genetic disorder that is inherited [2]. Type 2 diabetes is caused by impaired insulin function, so it cannot maximally convert glucose in the blood into energy [3]. While the last type is gestational diabetes, the cause is hormonal fluctuations during pregnancy to childbirth [4].

In biological systems, insulin action is triggered by the activation of proteins in a network of glycolysis systems. These proteins are known as receptors. Insulin receptor proteins, namely Insulin Receptor (INSR) and Insulin-like Growth Factor-1 Receptor (IGF1R), play an important role in activating and maximizing insulin performance in carrying out its biological role [5]. In previous studies, the role of IGF1R on insulin productivity was very significant [6]. The body needs an active compound (ligand) that matches the 3D construction of the receptor protein in order to trigger the receptor of the protein. So it requires a detailed analysis of the appropriate characteristics between the target protein and the ligand. Ligand compounds can be synthetic compounds or natural compounds (herbs). Synthetic compounds come from experiments and research on bonds in chemical structures [7], while herbal compounds come from substances produced by plants.

The chemical bonding characteristics of synthetic and herbal compounds have a complex structure. So, if it is analyzed manually, it will require high costs, both in terms of time and budget. The important role of artificial computer intelligence makes it easier for humans to simplify complex problems. Research that utilizes machine learning technology on biological objects is often carried out to facilitate analysis and efficiency. Machine learning methods such as Neural Network [8], Support Vector Machine [9], Random

Forest [10] have been used to screen compounds. In addition, an analysis of the interactions of interconnected proteins has also been carried out [6]. This condition is also supported by the database of compounds related to the IGF1R target protein. Open-source databases such as ChEMBL [11], DUD-E database [12], PubChem [13], and Super Target [14] are very helpful in collecting data on synthetic compounds that have been analyzed in previous studies. These synthetic compounds become a source of analysis to detect herbal compounds with similar characteristics and can be candidates for diabetes mellitus drugs.

Herbal compounds are potential sources of environmentally friendly drugs rather than synthetic compounds. Indonesia has abundant biodiversity, so potential sources of herbs are easy to obtain [15]. For example, herbal medicine in Indonesia is often only information that is not basic related to its efficacy. Thus, with this research, it is hoped that the potential of these herbs has a scientific basis for publication, especially as an alternative medicine for diabetes mellitus. This research aims to build a machine learning model using the Extreme Gradient Boosting (XGB) method as a classifier based on the data of ligands related to the IGF1R protein. The classification model formed becomes a predictor to determine herbal compounds. It is hoped that the results of this study will find herbal compounds that play an active role in triggering the activation of IGF1R protein and increasing insulin production. However, the results of this study require validation under both in vitro and in vivo conditions. In addition, molecular dynamics can also help increase the possibility of significant compounds that can be alternative drugs from the IGF1R target protein.

2. Methods

This research consists of several stages in the analysis process. An overview of the current research process is shown in Fig. 1. The activity was preceded by collecting data on ligand compounds from the open-source database DUD-E database [16]. After getting the data, the next process is data cleaning and adjustment of data construction. This process removes unnecessary data noise in the text and adjusts the data frame data. Data adjustment is by using a smile code on compound data.

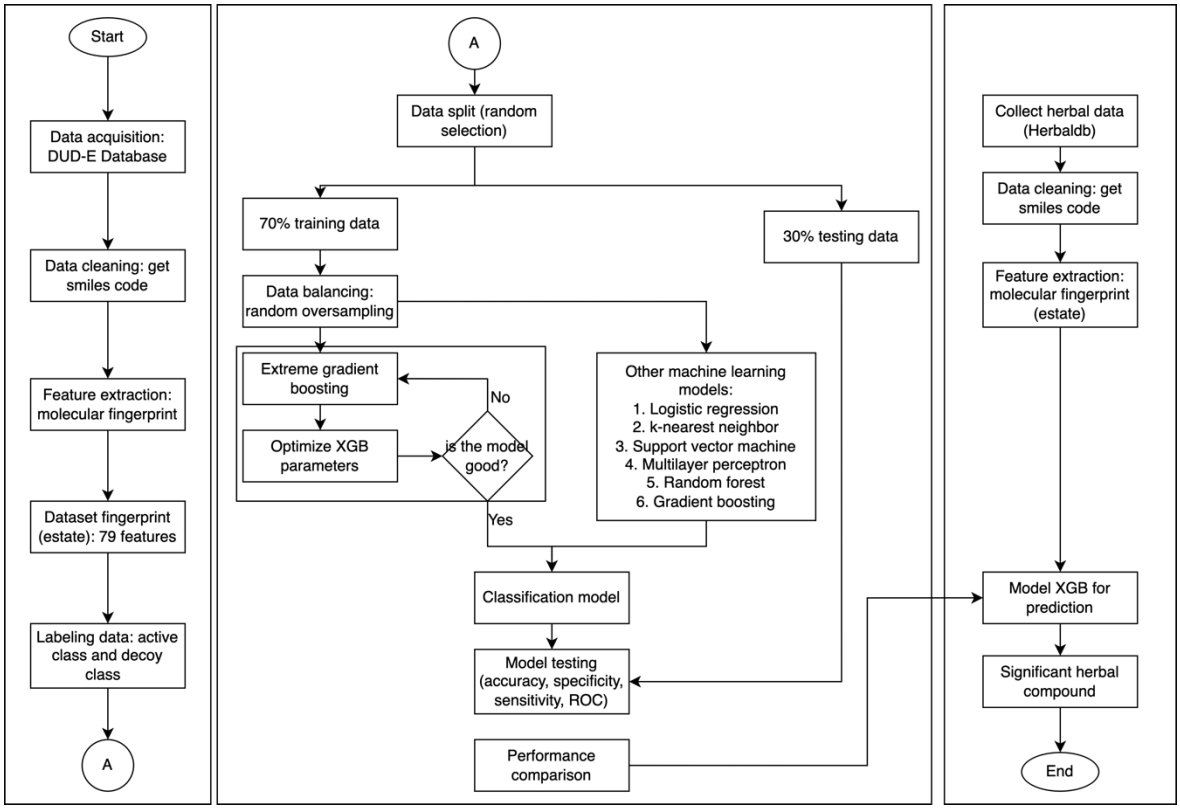

Fig. 1. A research flow chart that shows the initial process of data mining to the data processing to produce a predictive model of herbal compounds.

The next process is extracting data features using the Molecular Fingerprint method. This method extracts the characteristics of the compounds in the form of smiles into special characteristics that are coded Binary. The extraction results with molecular fingerprints yielded 79 data features [17]. Before modeling the data using machine learning algorithms, the data is first labeled as active data and decoy data. Thus, the available dataset has two classes. After being labeled, the data were separated into training data and test data. In the training data, we balance the proportion of the number of each data class. balancing data using random oversampling method. The next process is to model the data with the Extreme Gradient Boosting (XGB) algorithm. The modeling process separates the dataset into two parts, namely as training data and data as test data. This modeling process is done many times to get the optimal algorithm parameters. The resulting optimal model will predict whether the herbal compound is compatible with insulin function. This study also compares the XGB classification model with other classification algorithms. The XGB model generated by the algorithm was used to predict herbal compounds. Prediction results become a reference for each compound that is significant to the IGF1R target protein.

2.2. Random Oversampling

The problem of unbalanced data backfires from the research results. This condition is sometimes not realized in the formation of the model because the algorithm parameters cannot detect the sample size condition in each class. The random oversampling method provides a solution to balance the data size in each dataset class. In its application, this method duplicates the sample with the minor class [18]. That is, in classes with a smaller size than other classes, the sample data in that class will be duplicated randomly until the size of each data class is balanced.

2.3. Molecular Fingerprint

Each type of data requires a feature extraction process so that the modeled variables are uniform. Extraction of features related to data smiles can use a molecular fingerprint. This method binary confirms the bonds and elements present in the compound. If it is confirmed that there is a bond in the compound, the fingerprint index will be coded with 1; if it is not confirmed a bond in the compound, the fingerprint index will be coded with 0 [19]. An illustration of feature extraction with a molecular fingerprint is shown in Fig. 2.

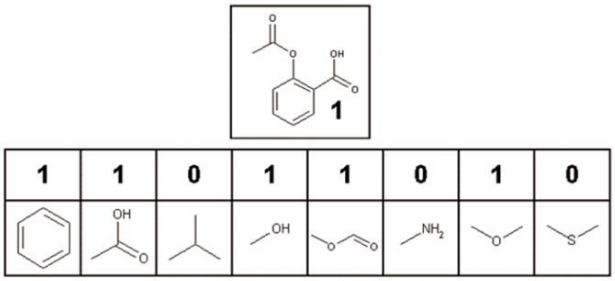

Fig. 2. The feature extraction process in the bottom image confirms the bonds and elements contained in the compound in the top image.

There are several types of molecular fingerprints, such as standard fingerprint, PubChem fingerprint, Klekota-Roth fingerprint, MACCS fingerprint, estate fingerprint, and circular fingerprint. Each type of fingerprint has different bits, such as the PubChem fingerprint, which has 881 bits [20], MACCS with 166 bits [21], and the fingerprint estate of 79 bits [22]. In this study, the limitation of feature extraction uses the fingerprint estate method with the RDkit library .

2.4. Extreme Gradient Boosting (XGB)

Tianqi Chen became the first initiator of a method belonging to this type of ensemble. He explained that the XGB method has the same rules as a decision tree. XGB is a machine learning method in terms of tree enhancement in the form of classification and regression models. The advantage of this method is its scalability in all scenarios of the dataset form [23]. XGBoost's scalability due to algorithmic optimization has innovations, including an algorithm for handling sparse data and a weighted quantitative sketch procedure that allows handling of sample weights.

The XGB model uses an additive function to determine its predictive value. Suppose we have a dataset of size n and has m features, $\phi = \{(x_i, y_i)\}$ $(|\phi| = n,\ x_i \in \mathbb{R}^m)$, the output function is [23],

$$\hat{y}_i = \sum_{i=1}^{K} f_k(x_i), \qquad f_k \in \psi \tag{1}$$

value $\psi = \{f(x) = w_q(x)\}$ with $q: \mathbb{R}^m \to T,\ w \in \mathbb{R}^T$ represents the regression tree space (1). The representation of the structure of each tree is expressed in q; this value maps the dataset to the appropriate leaf index. f_k is a function that declares the tree structure independent of q and corresponds to the leaf weight w. The objective function of the model learning system is to minimize the loss function and its complexity function [24],

$$\mathcal{L}(x) = \sum_{i=1} l(\hat{y}_i, y_i) + \sum_{k=1} \Omega(f_k) \tag{2}$$

where,

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

The loss function is expressed in terms of $l(\hat{y}_i, y_i)$ and can be distinguished between the predicted $\hat{y}_i$ and the target y_i. This implies that the loss function must be differentiable. The model complexity function is expressed in the form $\Omega(f)$.

Learning that occurs in equation (2) is trained additively. So it takes tth iteration to complete the objective function. If $\hat{y}_i^{(t)}$ is the prediction of the ith event in the tth iteration, it requires f_t to minimize the following objective

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \left(l\left(y_i, \hat{y}_i^{(t-1)}\right) + f_t(x_i) \right) + \Omega(f_t)$$

The addition of f_t to improve the model in equation (2). We can use the second-order Taylor expansion to solve the optimization quickly [25],

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} \left(l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right) + \Omega(f_t)$$

where, $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ are first order and second order gradient statistics of the loss function. The constant condition l can be omitted to obtain a simple objective function at step t,

$$\tilde{\mathcal{L}}^{(t)} \simeq \sum_{i=1}^{n} \left(g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right) + \Omega(f_t) \tag{3}$$

Let the set of leaves j be denoted by $I_j = \{i |\ q(x_i) = j\}$. Expanding the function in equation (3) can be written as,

$$\tilde{\mathcal{L}}^{(t)} \simeq \sum_{i=1}^{n} \left(g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

$$= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \tag{4}$$

We can find the optimum value of a function of equation (4) with $\mathcal{L}' = 0$,

$$\frac{d}{d_{w_j}}\sum_{j=1}^{T}\left[\left(\sum_{i\epsilon I_j} g_i\right)w_j + \frac{1}{2}\left(\sum_{i\epsilon I_j} h_i + \lambda\right)w_j^2\right] + \gamma T = 0$$

$$\sum_{i\epsilon I_j} g_i + \frac{1}{2}\left(\sum_{i\epsilon I_j} h_i + \lambda\right)w_j = 0$$

From the results of the above operation, we can determine the tree structure by calculating the optimal weight w_j^* of the j by leaves,

$$w_j^* = -\frac{\sum_{i\epsilon I_j} g_i}{\sum_{i\epsilon I_j} h_i + \lambda} \tag{5}$$

and calculate the optimal correspondence value by,

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\epsilon I_j} g_i\right)^2}{\sum_{i\epsilon I_j} h_i + \lambda} + \gamma T \tag{6}$$

Equations (5) and (6) can be used to construct an ensemble tree. The number of ensemble trees generated depends on the number of iterations. To determine the nodes in the tree is similar to the structure of a decision tree. However, the decision is to calculate the right and left functions of the set of leaves. Let I_L and I_R be the set of left and right nodes after splitting. The set of nodes is expressed in $I = I_L \cup I_R$, then the loss reduction equation after splitting is [26],

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\epsilon I_L} g_i\right)^2}{\sum_{i\epsilon I_L} h_i + \lambda} + \frac{\left(\sum_{i\epsilon I_R} g_i\right)^2}{\sum_{i\epsilon I_R} h_i + \lambda} - \frac{(\sum_{i\epsilon I} g_i)^2}{\sum_{i\epsilon I} h_i + \lambda}\right] \tag{6}$$

2.5. Herbal Compounds

Herbal compounds are compounds derived from plants. Indonesia is one of the locations with the potential for abundant flora diversity. Variants of flora allow the potential of the herbal compounds produced to vary. The content of herbal compounds is claimed to have great results and relatively low side effects [15]. For this reason, we investigated using data on herbal compounds to prove their effect on the IGF1R protein. The herbal compound data we collected from HerbalDB 2.0 [27]. This database was developed by the faculty of pharmacy, University of Indonesia. We do a scraping technique to get the data on the website database. In the HerbalDB 2.0 database, we got 6756 herbal compound data, but only 844 compound data were confirmed with PubChem id. So in this study, we screened 844 herbal data whether they were compatible with the IGF1R protein.

3. Results and Discussion

3.1. Data Construction

IGF1R active compounds were collected from the DUD-E database. In the database, we found 148 active compounds of IGF1R. As for the decoy data, the database provides data for 9,300 compounds. The difference between active and decoy data is high, so we randomly selected 296 decoy data. After selection, the proportion of each class is compared between 1:2, with a total of 444 data collected. The compounds we collected were in the form of the Simplified molecular-input Line-entry System (SMILES). Extraction of features from these compounds using fingerprint estate [18]. This method extracts the characters from the compound into a 79-bit fingerprint. Each fingerprint shows certain characteristics contained in these compounds. Each operational and decoy data is extracted to its character with the help of Python's RDkit package [28]. The active data group produces a data matrix of 148 x 79 and in the feed data, the extraction process forms a data matrix with a matrix size of 296 x 79. Fingerprint characteristics become variables in the classification model. The dataset size of the process is a 444 x 79 matrix. This dataset matrix becomes the main dataset.

Before the classification algorithm training process, we divided it into 2 types: training data and test data. The proportion of training data is 70% of the total data, while the test data is 30%. The amount of data that becomes training data is 311 data and 133 test data. The training data will be used to train the algorithm used, while the test data will measure the quality of the model formed. on the training data, the number of active classes we have is 209 data while the number of decoy classes is 102 data. To form a classification model, we balance the two data classes before being trained against a machine learning algorithm.

The data contained in the training data section is balanced by the random oversampling method [29]. This method is relatively simple because the selection of data duplication is made randomly. Duplicated data comes from minor data classes, namely those with a smaller size than other classes. The duplication process repeatedly occurs until the proportions between classes are balanced. Data duplication is carried out on minor data classes, namely active data classes. The amount of active data is balanced randomly as much as the number of data from the decoy class. The results of this oversampling process produce a total of 418 training data with balanced class conditions.

3.2. Model Construction and Optimization

The classification model of Extreme Gradient Boosting (XGB) develops the previous boosting tree method. Before the model is ready to predict herbal compounds, we look for optimal parameters in the algorithm. This parameter search uses repeated experiments on the formed model (trial and error). The optimization looks at the graph of the loss function in Fig. 3a. and the graph of misclassification in Fig. 3b. We set some XGB model parameters like n_estimator, learning_rate, n_jobs, max_depth, gamma, subsample, and colsample_bytree. The goal is to optimize the model to produce better accuracy and ROC scores.

Optimizing the model is by testing the number of booster trees generated (n_estimator). In the initial experiment, we generated as many as more than 500, which resulted in an overfitting model. The graph condition is also constant when the generated tree exceeds 120. So we limit the booster tree generated to 80 trees. This condition makes sense for us to prune more trees. In addition number of trees, the learning_rate parameter is also an important key to how the model learns about the data. After several trials, we did the best learning ratio at 0.15.

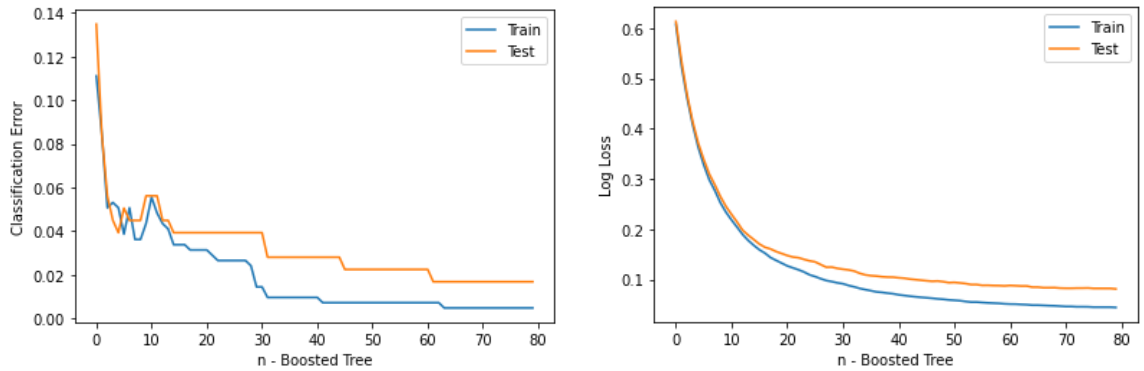

Fig. 3. The optimization of the model is based on observations of losses that occur in the objective function and on errors in the data classification model for each booster tree formed. (a) a graph showing the classification errors that occur in the model for each booster tree formed; (b) a graph showing the resulting loss on the objective function for each booster tree formed.

The optimal parameters of XGB against the IGF1R fingerprint data are summarized in Table 1. There are seven parameters that we set to get the optimal model. As in the previous explanation, the number of trees raised greatly affects the quality of the model. In the 80th tree generated, the loss on the objective function is 0.03973. If the next booster tree is raised again, the loss function in the training model experiences a persistent condition at a value of 0.03973. Other parameters also show a significant effect. A learning ratio of fewer than 0.15 results in a larger misclassification. If greater than 0.15, it will experience similar conditions, depending on the tree depth and the specified gamma. So, with repeated experiments to determine the optimal parameters, we use the parameter values in Table 1 as a determination to build a classification model with the Extreme Gradient Boosting algorithm.

Table 1. The parameters setting of the XGB classification method after optimizing the model training process.

Parameters	Description	Parameter Value
n_estimators	A number of trees generated by the gradient. This number is equivalent to the iteration of the upgrade.	80
learning_rate	Algorithm learning speed ratio.	0.15
n_jobs	Number of parallel threads to run the algorithm.	4
max_depth	Maximum depth of each booster tree.	4
gamma	The minimum pruning ratio for partitioning the tree nodes.	0.6
subsample	Subsample ratio of training data.	0.8
colsample_bytree	Column subsample ratio when constructing tree.	0.8

The variance of the data variables also influences the optimal model. Fig.4. shows the magnitude of the f-score value of each feature contained in the data. Feature number eighteen became the most influential feature informing the classification model because the f-score was the highest of other features with more than 0.14. The second most influential feature is feature number twenty-nine, with an f-score of more than 0.1. The score for other features is still below 0.1. The value of this feature indicates that the feature will often be encountered as the root of the generated booster tree. If the feature value is zero, the feature is not used in the booster tree as part of the classification.

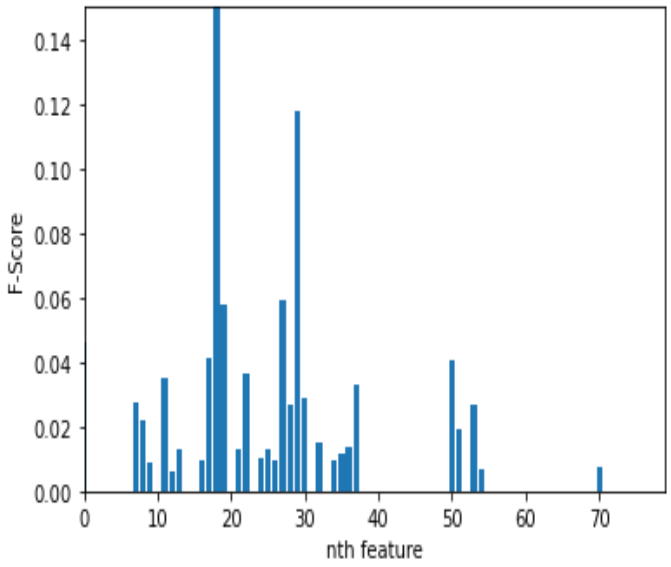

Fig. 4. The features that significantly influence the formation of the model, the greater the f-score value on the features, the more influential these features will be in the classification model.

3.3. Comparison of Model with Other Methods

The XGB classification model that has been optimized becomes a model for predicting herbal compounds. However, we wanted to prove whether the model was the best compared to other classification methods. We compare the other six classification models to the XGB model that has been formed. Other classification models that we use are Logistic Regression (LR), K-nearest Neighbor (K-nn), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forest (RF), and Gradient boosting (GB). The test results use a confusion matrix, and we use four measures to determine the quality of the resulting model. The results of this measurement are evidence of whether the XGB model has better quality than other machine learning models.

The test matrices for better model quality are accuracy, specificity, sensitivity, and Receiver Operating Characteristics (ROC curve). The measurement results of each test matrix are presented in Table 2. In the table, it can be seen that the accuracy of the XGB model is greater than the accuracy of other models.

The accuracy matrix describes how accurately the model is classifying the data correctly [30]. The higher the accuracy of the model, the less likely the model has errors in classification. This result is also supported by the specificity and ROC score value, which has a greater value than other classification models. These results prove that the XGB model has better quality than other machine learning models in terms of classifying compound fingerprint data. In addition, in Fig. 5, we can conclude that the ROC graph of the XGB model has a larger area under the graph than the other classification models. The SVM model is known on the graph as the worst model in this classification. Thus, SMV is not suitable for modeling compound fingerprint datasets.

Table 2. Measurement results on the test matrix in each classification model

Classification Models	Accuracy	Specificity	Sensitivity	ROC score
LR	0.9551	0.9451	0.9655	0.9553
K-nn	0.7416	0.5055	0.9885	0.7470
SVM	0.7135	0.4396	1.0000	0.7197
MLP	0.9607	0.9670	0.9540	0.9605
RF	0.9607	0.9451	0.9770	0.9610
GB	0.9775	0.9780	0.9770	0.9775
XGB	**0.9831**	**0.9890**	0.9770	0.**9830**

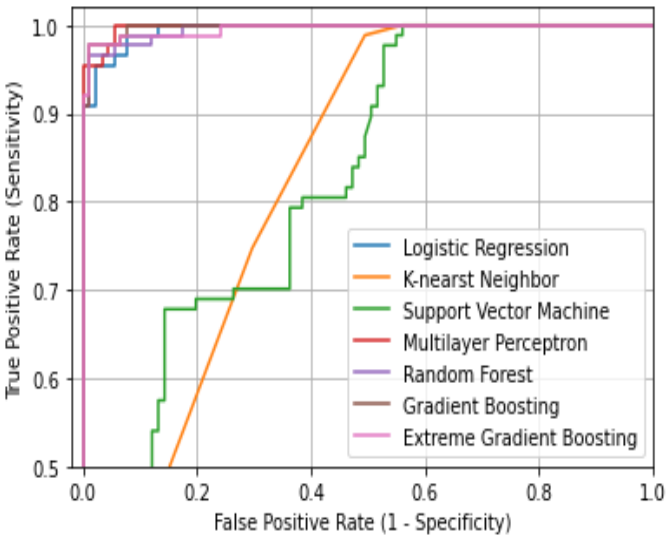

Fig. 5. The ROC/AUC graph for each classification method, the graph from SVM shows the condition of the model having the worst quality compared to others.

3.4. Herbal Compound Prediction

Before predicting herbal compounds, there are several processes to adjust the data to become model inputs. The data provided on HerbalDB 2.0 is similar to the compound PubChem code. So we need to get the Smiles code of each of these herbal compounds. Therefore, we did technical scraping to the Pubchem database and collected Smiles codes on each data. After getting it, we extracted the herbal data features with the same technique during algorithm training feature extraction. The reason is that these herbal compounds can be used as input in the classification model. After that, the data will predict its association with the IGF1R protein, whether it becomes an active compound or not. To state that the compound is significant for IGF1R, we set a threshold of 0.6. If the probability value of the prediction results in the model is less than 0.6, then the compound is declared insignificant to IGF1R; if it is greater than 0.6, then the compound is declared significant to IGF1R protein.

The predictions of herbal compounds can be seen in Table 3. Compounds with Pubchem ID 135596554 get the highest probability value compared to others. The predicted value of 0.97 means that the herbal compound is very close to the synthetic compound that has been shown to bind to the IGF1R

target protein. From the prediction results of 844 herbal compounds collected, only 15 compounds met the probability value threshold. So we concluded that these compounds were significant for IGF1R protein based on the classification model with IGF1R synthesis compounds.

Table 3. The prediction result of herbal compounds, the data are shown in the table, are compounds with a probability value of more than 0.6 or the top 15.

Compound names	Pubchem ID	Smiles	Probabilitas Values
2-amino-9[(2R,3R,4S,5R)-3,4-dihidroxy-5-(hydroxymethyl)oxolan-2-yl]-3H-purin-6-one	135596554	C1=NC2=C(N1[C@H]3[C@@H]([C@@H]([C@H](O3)CO)O)O)NC(=NC2=O)N	0.9671942
6,7-dimethoxy-1-methylisoquinoline	20725	C1=NC=CC2=CC(=C(C=C12)OC)OC	0.9474775
2-(1H-indol-3-yl)ethanol	10685	C1=CC=C2C(=C1)C(=CN2)CCO	0.9447276
3,5-dioxa-11-azapentacyclo[10.7.1.02,6.08,20.014,19]icosa-1(20),2(6),7,9,11,14,16,18-octaen-13-one	10144	C1OC2=C(O1)C3=C4C(=C2)C=CN=C4C(=O)C5=CC=CC=C53	0.9341754
(R)-[(2S,5R)-5-ethenyl-1-azabicyclo[2.2.2]octan-2-yl]-(6-methoxyquinolin-4-yl)methanol	8549	COC1=CC2=C(C=CN=C2C=C1)[C@H]([C@@H]3CC4CCN3C[C@@H]4C=C)O	0.8888987
3-(prop-2-enyldisulfanyl)prop-1-ene	16590	C=CCSSCC=C	0.8616078
(2R,3S,4R,5S)-hexane-1,2,3,4,5,6-hexol	11850	C([C@H]([C@@H]([C@@H]([C@H](CO)O)O)O)O)O	0.8564644
1,3-dimethyl-7H-purine-2,6-dione	2153	CN1C2=C(C(=O)N(C1=O)C)NC=N2	0.8285811
3,7-dimethylpurine-2,6-dione	5429	CN1C=NC2=C1C(=O)NC(=O)N2C	0.8285811
(2S)-2-amino-3-(1H-indol-3-yl)propanoic acid	6305	C1=CC=C2C(=C1)C(=CN2)C[C@@H](C(=O)O)N	0.8131784
1,3,7-trimethylpurine-2,6-dione	2519	CN1C=NC2=C1C(=O)N(C(=O)N2C)C	0.7910511
2-(1H-indol-3-yl)ethanamine	1150	C1=CC=C2C(=C1)C(=CN2)CCN	0.7677137
(5-ethenyl-1-azabicyclo[2.2.2]octan-2-yl)-(6-methoxyquinolin-4-yl)methanol	1065	COC1=CC2=C(C=CN=C2C=C1)C(C3CC4CCN3CC4C=C)O	0.7548318
1H-indole-3-carbaldehyde	10256	C1=CC=C2C(=C1)C(=CN2)C=O	0.617963
1,8-dihydroxy-3-(hydroxymethyl)anthracene-9,10-dione	10207	C1=CC2=C(C(=C1)O)C(=O)C3=C(C2=O)C=C(C=C3O)CO	0.6046094
octanedioic acid	10457	C(CCCC(=O)O)CCC(=O)O	0.6046094

3.5. Discussion

This section only discusses the potential of data mining for herbal drug discovery. All the data we get comes from an open database that has been validated. The classification model is the right choice because we have collected the compound IGF1R synthesis in the DUD-E database. These synthetic compounds are the source of training for the classification model. The average herbal data have not been classified as whether they are significant to IGF1R. So we aimed to use this model to label whether the herbal compound is significant for IGF1R.

This study did not involve in vitro analysis of biomolecules, so to verify whether these compounds are present in herbal plants, we searched the relevant literature to verify their validity. The compound with Pubchem ID 135596554 belongs to the guanosine family of compounds. This compound was found in

the plant Zostera Marina [31]. The Zostera Marine plant is a sea clump plant that belongs to the Zosteraceae family. However, some compounds have not been verified whether they are present in certain plants, such as compounds with Pubchem ID 20725 and 10685. So, we think that Zoster Marine plants can cure Diabetes Mellitus. These plants have compounds that bind to the target protein IGF1R, which is significant in the formation of insulin in the body. So, it could be that the Zoster Marine plant can be a cure for Diabetes Mellitus. This claim is temporary because it requires in vitro proof of the benefits of Zoster Marine plants. Future studies need to prove this bond by docking to the 3D structure of the protein. If the docking process meets, it can proceed to the next stage, namely testing on living things.

4. Conclusion

Bioinformatics data processing research has benefits, one of which is the discovery of new drugs. We used the data available in open databases, collecting synthetic compounds that affect the IGF1R protein. We extracted the compound features in code smiles into the fingerprint estate, which has 79 features. We classify the data using the ensemble method, namely Extreme Gradient Boosting. In addition, we also model it with other algorithms as a comparison.

The XGB model we got is the best classification model compared to other algorithms. We get an XGB model accuracy of 0.9831 and a ROC score of 0.9830. These values are the highest from other classification models. We used the XGB model to predict herbal compounds that we got from the HerbalDB 2.0 database. The prediction results state that the compound with Pubchem ID 135596554 has the highest probability value compared to other compound data. The resulting prediction value is 0.967, and this compound is identified as being contained in Zoster Marine plants. We set a threshold of 0.6 to determine whether or not the herbal compound binds to IGF1R. The threshold results indicate that fifteen herbal compounds have the potential to bind (significantly) to the IGF1R target protein. The results of this study show that Zoster Marine plants can be used as a drug for Diabetes Mellitus.

Acknowledgment

This research is a collaborative research institution between Institut Teknologi Telkom Surabaya, especially department of data science and Universitas Islam Darul 'ulum Lamongan. Funding from this research was obtained from both parties who mutually support the implementation of this research.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.
Funding statement. None of the authors have received any funding or grants from any institution or funding body for the research.
Conflict of interest. The authors declare no conflict of interest.
Additional information. No additional information is available for this paper.

References

[1] C. C. Regina, A. Mu'ti, and E. Fitriany, "Diabetes Mellitus Type 2," *Verdure Heal. Sci. J.*, vol. 3, no. 1, pp. 8–17, Jun. 2021, Accessed: Dec. 18, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK513253/.

[2] J. B. Cole and J. C. Florez, "Genetics of diabetes mellitus and diabetes complications," *Nat. Rev. Nephrol.*, vol. 16, no. 7, pp. 377–390, Jul. 2020, doi: 10.1038/S41581-020-0278-5.

[3] "Type 2 diabetes mellitus, oxidative stress and inflammation: examining the links - PubMed." https://pubmed.ncbi.nlm.nih.gov/31333808/ (accessed Dec. 18, 2020).

[4] H. D. McIntyre, P. Catalano, C. Zhang, G. Desoye, E. R. Mathiesen, and P. Damm, "Gestational diabetes mellitus," *Nat. Rev. Dis. Prim.*, vol. 5, no. 1, Dec. 2019, doi: 10.1038/S41572-019-0098-8.

[5] E. N. Gonc *et al.*, "Genetic IGF1R defects: new cases expand the spectrum of clinical features," *J. Endocrinol. Invest.*, vol. 43, no. 12, pp. 1739–1748, Dec. 2020, doi: 10.1007/S40618-020-01264-Y.

[6] M. H. Z. Al Faroby, M. I. Irawan, and N. N. T. Puspaningsih, "XGBoost and Network Analysis for Prediction of Proteins Affecting Insulin based on Protein Protein Interactions," *Kinet. Game Technol. Inf. Syst. Comput.*

Network, Comput. Electron. Control, vol. 4, no. Cc, pp. 253–262, 2020, doi: 10.22219/kinetik.v5i4.1076.

[7] Y. Khajebishak, L. Payahoo, M. Alivand, and B. Alipour, "Punicic acid: A potential compound of pomegranate seed oil in Type 2 diabetes mellitus management," *J. Cell. Physiol.*, vol. 234, no. 3, pp. 2112–2120, Mar. 2019, doi: 10.1002/JCP.27556.

[8] K. A. Carpenter and X. Huang, "Machine Learning-based Virtual Screening and Its Applications to Alzheimer's Drug Discovery: A Review," *Curr. Pharm. Des.*, vol. 24, no. 28, pp. 3347–3358, 2018, doi: 10.2174/1381612824666180607124038.

[9] Y. Peng and M. H. Nagata, "An empirical overview of nonlinearity and overfitting in machine learning using COVID-19 data," *Chaos, Solitons and Fractals*, vol. 139, 2020, doi: 10.1016/j.chaos.2020.110055.

[10] Y. Zhou *et al.*, "Quantitative Structure-Activity Relationship (QSAR) Model for the Severity Prediction of Drug-Induced Rhabdomyolysis by Using Random Forest," *Chem. Res. Toxicol.*, vol. 34, no. 2, pp. 514–521, Feb. 2021, doi: 10.1021/ACS.CHEMRESTOX.0C00347/SUPPL_FILE/TX0C00347_SI_001.ZIP.

[11] A. Capecchi, D. Probst, and J. L. Reymond, "One molecular fingerprint to rule them all: Drugs, biomolecules, and the metabolome," *J. Cheminform.*, vol. 12, no. 1, pp. 1–15, Jun. 2020, doi: 10.1186/S13321-020-00445-4/FIGURES/8.

[12] 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," *J. Med. Chem.*, vol. 55, no. 14, pp. 6582–6594, 2012, doi: 10.1021/jm300687e.

[13] S. Kim *et al.*, "PubChem in 2021: New data content and improved web interfaces," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1388–D1395, 2021, doi: 10.1093/nar/gkaa971.

[14] M. Bagherian, E. Sabeti, K. Wang, M. A. Sartor, Z. Nikolovska-Coleska, and K. Najarian, "Machine learning approaches and databases for prediction of drug-target interaction: A survey paper," *Brief. Bioinform.*, vol. 22, no. 1, pp. 247–269, 2021, doi: 10.1093/bib/bbz157.

[15] Y. Y. S. Rahayu, T. Araki, and D. Rosleine, "Factors affecting the use of herbal medicines in the universal health coverage system in Indonesia," *J. Ethnopharmacol.*, vol. 260, p. 112974, Oct. 2020, doi: 10.1016/J.JEP.2020.112974.

[16] P. I. Koukos, M. Réau, and A. M. J. J. Bonvin, "Shape-Restrained Modeling of Protein-Small-Molecule Complexes with High Ambiguity Driven DOCKing," *J. Chem. Inf. Model.*, vol. 61, no. 9, pp. 4807–4818, 2021, doi: 10.1021/acs.jcim.1c00796.

[17] N. R. Das, S. P. Mishra, and P. G. R. Achary, "Evaluation of molecular structure based descriptors for the prediction of pEC50(M) for the selective adenosine A2A Receptor," *J. Mol. Struct.*, vol. 1232, p. 130080, May 2021, doi: 10.1016/J.MOLSTRUC.2021.130080.

[18] A. Salazar, L. Vergara, and G. Safont, "Generative Adversarial Networks and Markov Random Fields for oversampling very small training sets," *Expert Syst. Appl.*, vol. 163, p. 113819, Jan. 2021, doi: 10.1016/J.ESWA.2020.113819.

[19] A. Fitriawan, I. Wasito, A. F. Syafiandini, M. Amien, and A. Yanuar, "Deep belief networks using hybrid fingerprint feature for virtual screening of drug design," *2016 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2016*, pp. 417–420, Mar. 2017, doi: 10.1109/ICACSIS.2016.7872737.

[20] A. Capecchi, M. Awale, D. Probst, and J. L. Reymond, "PubChem and ChEMBL beyond Lipinski," *Mol. Inform.*, vol. 38, no. 5, May 2019, doi: 10.1002/MINF.201900016.

[21] K. Dührkop *et al.*, "SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information," *Nat. Methods*, vol. 16, no. 4, pp. 299–302, Apr. 2019, doi: 10.1038/S41592-019-0344-8.

[22] S. Kim, P. A. Thiessen, E. E. Bolton, and S. H. Bryant, "PUG-SOAP and PUG-REST: Web services for programmatic access to chemical information in PubChem," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W605–W611, 2015, doi: 10.1093/nar/gkv396.

[23] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, doi: 10.1145/2939672.

[24] M. Rahman, Y. Cao, X. Sun, B. Li, and Y. Hao, "Deep pre-trained networks as a feature extractor with XGBoost to detect tuberculosis from chest X-ray," *Comput. Electr. Eng.*, vol. 93, p. 107252, Jul. 2021, doi: 10.1016/J.COMPELECENG.2021.107252.

[25] M. R. Mohammadi *et al.*, "Modeling hydrogen solubility in hydrocarbons using extreme gradient boosting and

equations of state," *Sci. Rep.*, vol. 11, no. 1, pp. 1–20, 2021, doi: 10.1038/s41598-021-97131-8.

[26] T. Avian *et al.*, "SS symmetry Machine Learning for the Prediction of Antiviral Compounds," 2022.

[27] R. R. Syahdi, J. T. Iqbal, A. Munim, and A. Yanuar, "HerbalDB 2.0: Optimization of construction of three-dimensional chemical compound structures to update Indonesian medicinal plant database," *Pharmacogn. J.*, vol. 11, no. 6, pp. 1189–1194, 2019, doi: 10.5530/pj.2019.11.184.

[28] R. Singh *et al.*, "Classification of beta-site amyloid precursor protein cleaving enzyme 1 inhibitors by using machine learning methods," *Chem. Biol. Drug Des.*, vol. 98, no. 6, pp. 1079–1097, Dec. 2021, doi: 10.1111/CBDD.13965.

[29] S. Bagui and K. Li, "Resampling imbalanced data for network intrusion detection datasets," *J. Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-020-00390-x.

[30] R. Couronné, P. Probst, and A. L. Boulesteix, "Random forest versus logistic regression: A large-scale benchmark experiment," *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–14, 2018, doi: 10.1186/s12859-018-2264-5.

[31] N. K. Hepler, A. Bowman, R. E. Carey, and D. J. Cosgrove, "Expansin gene loss is a common occurrence during adaptation to an aquatic environment," *Plant J.*, vol. 101, no. 3, pp. 666–680, Feb. 2020, doi: 10.1111/TPJ.14572.