Performance of multivariate mutual information and autocorrelation encoding methods for the prediction of protein-protein interactions

Alhadi Bustamam, Mohamad Irlin Sunggawa, Titin Siswantining
Department of Mathematics, Universitas Indonesia, Depok, Indonesia

Article Info

Article history:

Received Jun 28, 2021
Revised Feb 11, 2022
Accepted Feb 27, 2022

Keywords:

Autocorrelation
Machine learning
Multivariate mutual information
Protein-protein interactions
Support vector machine

ABSTRACT

Protein interactions play an essential role in the study of how an organism can be infected with a disease and also its effects. One of the challenges in computational methods in the prediction of protein-protein interactions is how to represent a sequence of amino acids in a vector so that it can be used in machine learning to create a model that can predict whether or not an interaction occurs in a protein pair. This paper examined the qualitative feature encoding methods of amino acid sequence, namely, multivariate mutual information (MMI), and the quantitative feature encoding methods, namely, autocorrelation. We develop the new design for MMI and autocorrelation feature encoding methods which give better results than the previous research. There are four ways to build the MMI method and six ways to build the autocorrelation method that we tested. We also built four types of MMI-autocorrelation (mixed) method and look for the best form of each type of MMI, autocorrelation, and mixed-method. We combine these feature encoding methods with support vector machine (SVM) as machine learning methods. We also test the encoding methods we propose to several machine learning classifier methods, such as random forest (RF), k-nearest neighbor (KNN), and gradient boosting.

Corresponding Author:

Alhadi Bustamam
Department of Mathematics, Universitas Indonesia
St. Margonda Raya, Pondok Cina, Beji, Depok, Jawa Barat, Indonesia
Email: alhadi@sci.ui.ac.id

1. INTRODUCTION

Proteins are linear polymers of amino acids linked together by peptide bonds [1]. Amino acids are organic molecules that have a distinctive structure that consists of a central alpha carbon atom (C_a) bound to four different chemical groups: amino groups ($-NH_2$), carboxyl groups or carboxylic acids ($-COOH$), a hydrogen atom (H), and a variable group called a side chain group, or R. Protein-protein interactions are known as processes of physical contact between proteins contained in cells or an organism. Moreover, the physical contact between two proteins does not necessarily produce an interaction; thus, the physical contact referred to in protein interactions needs to be elucidated and specified. According to Rivas and Fontanillo [2], the definition of protein-protein interactions should consider that interactions must occur intentionally; the interactions that occur result from certain selected biomolecular events. The occurring interactions must be non-generic, evolved for specific purposes that differ from the generic functions, such as protein production and degradation of functions. In general, protein interactions are divided into two types: the first one is the interaction between proteins in an organism, and the second is the interaction between proteins from different organisms, which is also known as pathogen-host interaction (PHI) [3].

Identification of protein interactions is very important to explain the function of proteins and identify the biological processes that occur in cells. Protein interactions induce a variety of changes, including changes in signals in cells and enzymatic reactions. Knowledge of protein interactions can help in the study of the mechanism of a disease caused by a virus infection [4], [5], in which its results are also used for developing a drug design for the corresponding disease. Although there has been a rapid development in technology for studying protein interactions, the cost of experimentation in studying protein interactions is still high. Also, the research process requires a lot of time and effort. For example, suppose there are 100 proteins in organism A and 200 proteins in organism B; studying which proteins interact requires testing as much as 20, 000 times. In reality, the amount of protein in an organism can reach thousands or tens of thousands, so that the testing process becomes longer; thus, it is impossible to study all of the possible pairs in a short time. Due to limited costs, time and effort, it is important to choose protein pairs that will be prioritised in the study and ignore the other protein pairs.

The bioinformatics researchers are vying to develop a computational approach for accurately and efficiently studying and predicting protein-protein interactions. This method was developed to help researchers in the field of biological sciences select which protein pairs are unlikely to have interactions, so that researchers can eliminate other protein pairs. Until now, various types of coding techniques and machine learning-based computing methods for determining whether or not proteins interact have been widely developed. According to the previous research there are some methods that using information of evolution [6], using natural language processing [7], and using clustering methods [8], [9] to learn protein-protein interaction. The most popular method is predicting the interaction between two proteins based on the sequences of amino acids. In the process, the researchers manipulated amino acid sequences into vectors to be processed using machine learning methods to create a model that could predict the interaction between two proteins.

In manipulating proteins into a vector, which is called the feature encoding method, researchers have developed several types of methods. In general, we classify the encoding methods of those features into three types. The first one includes encoding methods that calculate values based on the composition, transition and distribution of amino acid sequences. In this method, the researchers categorised 20 types of amino acids into several classes. The amino acid sequence was converted into a sequence of numbers representing the class of each amino acid. A mathematical calculation was then performed to produce values that will be used as features to be studied via machine learning. We called this type of encoding method as qualitative encoding method. The encoding methods incorporated into this qualitative type are conjoint triad [10], multivariate mutual information (MMI) [11] and local descriptor [12]. The second type of feature encoding method uses the characteristic values of each amino acid's physicochemical properties. In this method, the sequences of amino acids, which are sequences of letters, will be converted into sequences of numbers representing the physicochemical property value of each corresponding amino acid. The sequences of numbers will be processed using mathematical or statistical methods, and this second method is called the quantitative method. The encoding methods incorporated into this quantitative type are Moran autocorrelation (MAC) [12] and normalised Moreau-Broto autocorrelation (NMBAC) [11]. The third method is the feature encoding method that uses matrices to manipulate amino acid sequences into a vector. Aside from these three types of methods, some studies used a combination of two or more methods. Ding *et al.* [11] compared the effectiveness of the MMI encoding method, NMBAC and combinations of those two types of encoding methods. These methods are combined with a random forest (RF) machine learning classifier.

In this study, we developed new ways to build a qualitative and quantitative feature encoding methods. The qualitative encoding method that we used was MMI, while the quantitative encoding method that we used was the method that adopts the calculation of autocorrelation coefficients using physicochemical property values. We also examined the results of the combination of the MMI and autocorrelation feature encoding method and compared them with those of each feature encoding method. We employed this encoding method using a support vector machine (SVM) classifier to predict the possibility of protein interaction based on the data on protein interactions in HIV-1 and humans. Subsequently, we identified methods that can optimise the prediction results of protein interactions by calculating the values of accuracy, sensitivity, specificity and F1 score. We used the protein interaction data between HIV-1 and humans from the National Centre for Biotechnology Information (NCBI) website.

2. RESEARCH METHOD

We collected data on the protein interactions between HIV-1 and humans. Then, the data was divided into training data and testing data at a ratio of 4:1. Furthermore, sequences of amino acids were converted into specific dimensional vectors using the feature encoding method. Subsequently, the vectors formed from the training data were used and studied using an SVM to predict the possibility of protein interaction of protein pairs. In the final stage, the model was tested using data testing, and the values of accuracy, sensitivity, specificity and F1 score were calculated to evaluate the model created. We conducted five trials to avoid assumption that the model obtained will yield good results for each type of feature encoding method as the training and testing data were coincidentally divided into accurate proportions. After conducting the five trials, an average was taken as the value representing each of the models studied. To identify the best feature encoding method, we tested each method in the prediction of protein interactions using the SVM classifier. We took each of the two best qualitative and qualitative types to develop a combined feature encoding method.

2.1. Datasets

According to Kösesoy *et al.* [3], one of the common problems in the development of computational methods for predicting protein interactions is scarce protein interaction data. We used data on the protein interactions between HIV-1 and humans, which was downloaded from the NCBI website. In the list of interacting proteins that were downloaded, there were 16,215 pairs of protein interaction data. We reduced the duplicated data so that the existing data will also be reduced to 7,919 pairs. We downloaded the amino acid sequences found in both HIV-1 and humans. We compiled them into datasets of interacting protein pairs, which are called positive datasets. Aside from the scarcity of data, the unavailability of data pairs of proteins that do not interact was also a problem [13]. To solve this problem, we randomly selected 7,919 pairs of proteins that did not exist in the downloaded protein interaction database to be used as assumed pairs of non-interacting proteins. The pairs were chosen to be used as negative datasets. Thus, there were 15,838 pairs of HIV-1 and human proteins in the dataset used, which consisted of 7919 pairs of positive datasets and 7,919 pairs of negative datasets.

2.2. Feature encoding methods

In this study, we used MMI and three different formula of autocorrelation coefficient as feature encoding methods. There are four types of MMI feature encoding method and six types autocorrelation feature encoding methods we test in this paper. We develop the MMI methods by using different combination of classification of amino acids and formula. We also test the different formula of autocorrelation coefficient, which are moran, normalised Moreau-Broto and geary autocorrelation.

2.2.1. Multivariate mutual information (MMI)

In 1948, Shannon introduced the concept of representing the measuring value of non-linear relationships between two variables, which is known as the concept of mutual information (MI), in the future [14]. The concept was developed to measure the non-linear relationships of three or more variables, further known as MMI. Ding *et al.* [11] adopted the concept of MI and MMI to be used as features to describe amino acid sequences. They manipulated the form of amino acid sequences into sequences of numbers by classifying amino acid sequences based on the specific properties of each amino acid, after which they calculated the MI and MMI values of each protein. In the final stages, they created a vector containing the calculated MI and MMI values, where they are used as vectors representing the corresponding amino acid pairs.

a. Conversion of amino acid sequences into sequences of numbers

According to previous studies [10]–[12], protein interactions can be explained by the dipole scale values and volume of the side chain of each amino acid. Based on these two things, 20 types of amino acids are divided into 7 classes, as presented in Table 1. In the early stages, a sequence of amino acids is converted into a sequence of numbers indicating each amino acid class. For example, suppose that there is a sequence of amino acids ACFHLP. Inspired by Ding *et al.* [11], we convert the amino acid sequence into 013433 because based on Table 1, A is in class 0, C is in class 1, H is in class 4, and F, L and P are in class 3.

b. Calculation of the values of MI and MMI

After converting a sequence of amino acids into a sequence of numbers, we calculate the frequency of element a (n_a) that appears in that sequence $(a = 0, 1 \dots, 6\ )$. Furthermore, we calculate the values of the 2-g feature $(n_{00}, n_{01}, ., n_{66})$ and the 3-g feature $(n_{000}, n_{001}, \dots, n_{666})$ as shown in Figure 1. The next step is to calculate the values of non-linear relationships between two variables (MI) and three variables (MMI) using the values of n_a, n_{ab} and n_{abc}, according to the formula used by Ding *et al.* [11]:

$$I(a,b) = f(a,b) ln\left(\frac{f(a,b)}{f(a)f(b)}\right) \tag{1}$$

$$I(a,b,c) = I(a,b) - I(a,b\,|c) \tag{2}$$

where

$$I(a,b|c) = H(a|c) - H(a|b,c) \tag{3}$$

$$H(a|c) = -\frac{f(a,c)}{f(c)} l\,n\left(\frac{f(a,c)}{f(c)}\right) \tag{4}$$

$$H(a|b,c) = -\frac{f(a,b,c)}{f(b,c)} l\,n\left(\frac{f(a,b,c)}{f(b,c)}\right) \tag{5}$$

Table 1. Classification of 20 amino acids based on dipole scales and volume of side chains [10]–[12]

Class	Group	Dipole Scale	Volume Scale
0	*A,G,V*	$Dipole < 1.0$	$Volume < 50$
1	*C*	$1.0 < Dipole < 2.0$ *(form disulphide bonds)*	$Volume > 50$
2	*D,E*	$Dipole > 3.0$ *(opposite orientation)*	$Volume > 50$
3	*F,I,L,P*	$Dipole < 1.0$	$Volume > 50$
4	*H,N,Q,W*	$2.0 < Dipole < 3.0$	$Volume > 50$
5	*K,R*	$Dipole > 3.0$	$Volume > 50$
6	*M,S,T,Y*	$1.0 < Dipole < 2.0$	$Volume < 50$

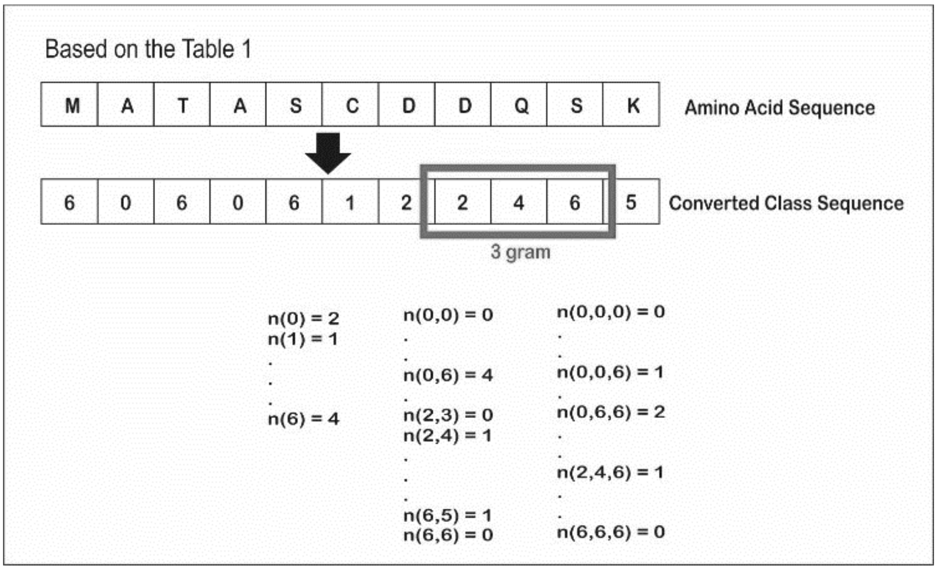


Figure 1. Converting amino acid sequence into a class sequence

Considering the possibility of the absence of element a in a converted sequence ($n_a = 0$), if we define $f(a) = n_a$, the MI and MMI values will become undefined, as $f(a)$ is the denominator in each formula above. Thus, the value $f(a)$ needs to be defined to avoid the MI and MMI values from becoming undefined according to Ding *et al*. [11]:

$$f(a) = \frac{n_a+1}{L+1} \tag{6}$$

where L denotes the length of the sequence. As the multivariate value of information between a and b is equal to the mutual value of information between b and a, which means $I(a,b) = I(b,a)$, we only use one value, that is, $I(a,b)$, and ignore $I(b,a)$, as well as the values of MMI.

In this study, we also attempted to build an MMI feature encoding method by applying the same concept but using a combination of different ways amino acid classification and defining the $f(a)$ function. In the feature encoding method of global encoding [15], 20 amino acids are grouped into 6 classes based on the carbon chain structure of each amino acid, as shown in Table 2.

Furthermore, considering the condition when each value n_a added by one, the length of the sequence will increase according to the number of possible values of a. For example, if each value $n_0.n_1,\dots,n_6$ is added by one, then the length of the sequence will increase to seven because there are seven possibilities of a: 0, 1, 2, 3, 4, 5 and 6. Therefore, we define variable c so that:

$$f(a)=\frac{n_a+1}{L+c} \quad (7)$$

where c denotes the number of possible values of a. Since there are seven classes, when calculating $f(a)$, we used $c=7$. To calculate each value of $f(a,b)$, we used $c=28$, as there are 28 pairs of a and b (00, 01, …, 66). To calculate $f(a,b,c)$, we used $c=84$, as there are 84 values of a,b,c that can be obtained from combinations of those seven classes. To represent a sequence of amino acids, we created a 119-dimensional vector $(7+28+84)$ that contains $f(a), I(a,b)$ and $I(a,b,c)$. Because there are two proteins (host and pathogen), the vector formed has a dimension of $119\times 2=238$. Similarly, when classifying 20 amino acids based on Table 2, there are 6 values of $f(a)$, 21 values of $I(a,b)$ and 56 values of $I(a,b,c)$; thus, the vector formed has a dimension of $83\times 3=166$. By permutating two ways of classifying amino acids and two ways of defining the function $f(a)$, there are four types of MMI that we will test:

i) MMI with Table 1 and $f(a)=\frac{n_a+1}{L+1}$
ii) MMI with Table 1 and $f(a)=\frac{n_a+1}{L+c}$
iii) MMI with Table 2 and $f(a)=\frac{n_a+1}{L+1}$
iv) MMI with Table 2 and $f(a)=\frac{n_a+1}{L+c}$

Table 2. Classification of 20 amino acids based on each structure of carbon chains [15]

Group	Amino Acid classification	Amino Acids
0	*Aliphatic*	*A,V,L,I,M,C*
1	*Aromatic*	*F,W,Y,H*
2	*Polar*	*S,T,N,Q*
3	*Positive*	*K,R*
4	*Negative*	*D,E*
5	*Special*	*G,P*

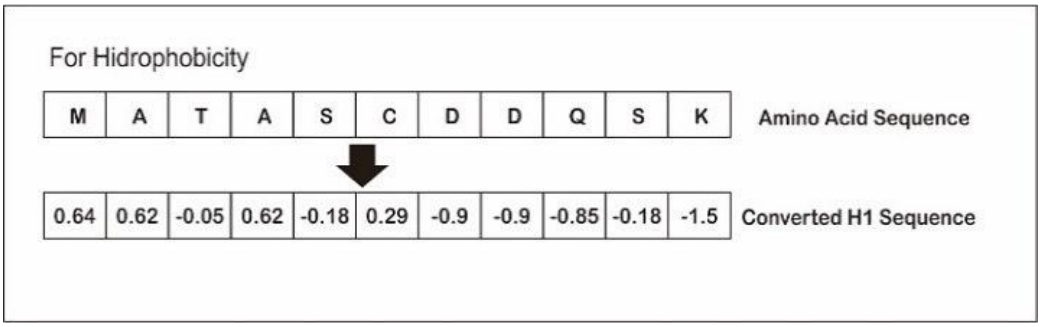
2.2.2. Autocorrelation

Inspired by previous studies using physicochemical properties, Ding *et al.* [11] adopted the concept of NMBAC value calculation using six physicochemical property values to convert amino acid sequences. The six values are hydrophobicity (H_1), volumes of side chains of amino acids (VSC), polarity (P_1), polarizability (P_2), solvent-accessible surface area (SASA) and net charge index of side chains (NCISC). Wang *et al.* [16] applied the concept of autocovariance by using seven physicochemical properties, of which six were the same as those in the research by Ding *et al.* [11] and the seventh one was hydrophilicity (H_2). In this study, we applied autocorrelation concepts and utilised the physicochemical properties of each amino acid to convert amino acid sequences. We tested the use of both six and seven physicochemical properties to compare the results.

a. Conversion of amino acid sequences into sequences of physicochemical properties

Using the values of physicochemical properties as shown in Table 3, we convert sequences of amino acids into sequences of numbers. Assuming a sequence of numbers formed as a time series, we calculate the autocorrelation coefficient values using a variety of different lag values. We use lag values from 1 to 30 similar to those used by Ding *et al.* [11]. Furthermore, we tested three autocorrelation concepts, namely, NMBAC, MAC and GAC, each divided into two types, using six physicochemical properties (without H_2) and seven physicochemical properties (with H_2).

Each sequence of amino acids will be converted into six or seven sequences containing numbers, where each sequence represents the values of a particular type of physicochemical property as shown in Figure 2. Since the lag values we used are 1 to 30, if we build a vector using seven physicochemical properties, we obtain a 210-dimensional vector (30×7) to represent an amino acid sequence. Also, as there are two proteins (host and pathogen), we obtain a 420-dimensional vector. Similarly, we created a 360-dimensional vector $(30\times 6\times 2)$ to represent a pair of proteins using the concept of autocorrelation with six values of physicochemical properties.



For Hidrophobicity

M	A	T	A	S	C	D	D	Q	S	K	Amino Acid Sequence
0.64	0.62	-0.05	0.62	-0.18	0.29	-0.9	-0.9	-0.85	-0.18	-1.5	Converted H1 Sequence

Figure 2. Conversion of amino acid sequences into sequences of physicochemical properties

Table 3. The values of physicochemical properties of amino acids [17]

No	H_1	H_2	V	P_1	P_2	SASA	NCI
A	0.62	−0.5	27.5	8.1	0.046	1.181	0.007187
C	0.29	−1	44.6	5.5	0.128	1.461	−0.03661
D	−0.9	3	40	13	0.105	1.587	−0.02382
	−0.74	3	62	12.3	0.151	1.862	0.006802
F	1.19	−2.5	115.5	5.2	0.29	2.228	0.037552
G	0.48	0	0	9	0	0.881	0.179052
H	−0.4	−0.5	79	10.4	0.23	2.025	−0.01069
I	1.38	−1.8	93.5	5.2	0.186	1.81	0.021631
K	−1.5	3	100	11.3	0.219	2.258	0.017708
L	1.06	−1.8	93.5	4.9	0.186	1.931	0.051672
M	0.64	−1.3	94.1	5.7	0.221	2.034	0.002683
N	−0.78	2	58.7	11.6	0.134	1.655	0.005392
P	0.12	0	41.9	8	0.131	1.468	0.239531
Q	−0.85	0.2	80.7	10.5	0.18	1.932	0.049211
R	−2.53	3	105	10.5	0.291	2.56	0.043587
S	−0.18	0.3	29.3	9.2	0.061	1.298	0.004627
T	−0.05	−0.4	51.3	8.6	0.108	1.525	0.003352
V	1.08	−1.5	71.5	5.9	0.14	1.645	0.057004
W	0.81	−3.4	145.5	5.4	0.409	2.663	0.037977
Y	0.26	−2.3	117.3	6.2	0.298	2.368	0.023599

b. Calculation of the values of NMBAC

To calculate the NMBAC values, we used the same equation used by Ding *et al*. [11]:

$$NMBAC_{lag,j} = \frac{1}{n-lag}\sum_{i=1}^{n-lag}\left(X_{i,j} \times X_{i+lag,j}\right) \tag{8}$$

c. Calculation of the values of MAC

To calculate the MAC values, we used the same equation used by You *et al*. [12]:

$$MAC_{lag,j} = \frac{\frac{1}{n-lag}\sum_{i=1}^{n-lag}\left(X_{i,j}-\overline{X_{\iota,j}}\right)\times\left(X_{i+lag,j}-\overline{X_{\iota,j}}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(X_{i,j}-\overline{X_{\iota,j}}\right)^2} \tag{9}$$

d. Calculation of the values of GAC

To calculate the GAC values, we used the same equation used by Du *et al*. [18]:

$$GAC_{lag,j} = \frac{\frac{1}{2\times(n-lag)}\sum_{i=1}^{n-lag}\left(x_{i,j}-x_{i+lag,j}\right)^2}{\frac{1}{N}\sum_{j=1}^{N}\left(x_{i,j}-\overline{x_{\iota,j}}\right)^2} \tag{10}$$

2.3. Classification methods

There are several machine learning classifier techniques, one of which is SVM. In this research, we used SVM as a machine learning classifier for predicting protein-protein interaction. SVM is a machine learning classifier that used to solve classification problem. We compare our result with the previous result that also used SVM as machine learning classifier.

2.3.1. Support vector machine (SVM)

SVM is a machine learning method developed based on the Vapnik-Chervonenkis theory [19]. SVM is generally used for some problems, such as classification, regression and density estimation [20]. SVM utilises data features as coordinates of points in a particular dimension that will be transformed into higher dimensions using a kernel function so that the data can be more easily classified [21]. The schematic of the main idea of SVM can be seen from Figure 3. In Figure 3, we can see that the dots in the cross and dot shaped classes in the two-dimensional plane are separated by a line. With the same concept, if the dots are in a three-dimensional plane, the separator is called a plane, whereas for dots that are at a higher dimension, the separator boundary is called a hyperplane.

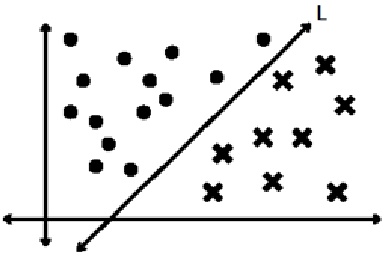


Figure 3. Two classes separated by a line

As can be seen from Figures 2 and 3, the main problem of SVM is finding the best hyperplane that can separate the data into two classes; mathematically, the hyperplane should have a maximum distance to the nearest point. Furthermore, the problem is solved by using several methods, one of which is the Lagrange multiplier [22]. With the complexity of the dataset, it can sometimes be difficult to create hyperplanes that can separate data into two classes. Therefore, a method known as a kernel function was developed. It transformed the data into a higher dimension, resulting in the creation of a hyperplane to make it possible to separate the two classes. The simulation of the role of kernel functions in SVM is presented in Figure 4, where it can be seen that when the data is mapped onto two-dimensional fields, the cross and dot shaped data are randomly separated. On the left side of the image, the data with different classes are mixed into one. It can be said that it is impossible to create a hyperplane that can accurately divide data into two classes. Furthermore, each point is mapped onto a higher dimension (three dimensions) using the kernel function, thus creating a shape on the right sides, which can be seen from Figure 4, so that the hyperplane that separates the two classes becomes easier to create. In identifying the best hyperplane, a parameter must be optimised, namely, C. C is used to control the trade-off between margin and error classification. There are several kernel functions that can be used in SVM, such as linear, polynomial, radial basis function (RBF), sigmoid, multi-quadratic inverse and additive [23]. In this study, we employed RBF as a kernel function using equation $k(x, x_i) = \exp(-\gamma \|x - x_i\|^2)$ [24]. Based on the equation of the RBF function, one more parameter must be optimised, which was the gamma (γ) parameter.

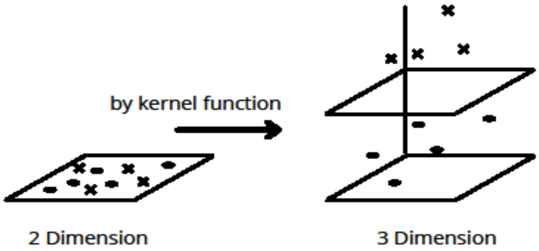


Figure 4. Simulation of the kernel function's work

2.4. Evaluation measurements

This study has four parameters: accuracy, sensitivity, specificity and F1 score. These parameters are defined:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (12)$$

$$Specificity = \frac{TN}{TN+FP} \quad (13)$$

$$F1\ Score = \frac{2 \times SN \times PPV}{SN+PPV} \quad (14)$$

where TP, TN, FP and FN denote true positive (both predicted and actual classes are positive), true negative (both predicted and actual are negative), false positive (predicted is positive but actual is negative) and false negative (predicted is negative but actual is positive), respectively.

3. RESULTS AND DISCUSSION

We train each encoding method feature using SVM. On the process, we apply the k-fold cross-validation method using the values k=5, 6, 7, 8, 9 and 10. The optimum model is obtained from the results when k=7; thus, we use the value k=7 every time we train the data. There are differences in C and gamma values that optimise the accuracy values of each type of feature encoding method. We used C and the gamma values are 10^n $(n = -3, -2, -1, 0, 1, 2, 3)$. We used the grid search method to find the combination of C and gamma that offers the best accuracy.

3.1. The results of MMI methods

In the early stages of the research, we tested the combination of MMI and SVM. We tested each type of combination five times and averaged the values of each evaluation measurement. In the two initial experiments in each MMI type, C and gamma that provide the optimum model had the same value, which was 0.1; thus, we conducted advanced training using the same C and gamma values in the remaining three experiments.

The results for each type of MMI can be seen from Table 4. From the table, we can see that the encoding method of MMI type 1 (MMI method used in the research by Ding *et al.* [11]) has the smallest average values of accuracy, sensitivity and F1 score when compared with other types. Meanwhile, the average value of specificity yields the best results. In this case, the application of MMI encoding method type 1 can predict the pair of proteins that do not interact well compared with other methods. This is following the purpose of the computational method in predicting protein interactions that is to reduce the number of protein pairs that are unlikely to interact. MMI type 2 and MMI type 4 provide the first- and second-best average accuracy values, respectively, where MMI type 2 is superior in average sensitivity values, whereas MMI type 4 excels negligibly in the average specificity value and F1 score. The advantages of MMI type 4 are few; thus, we concluded that MMI type 2 and type 4 have the same performance in predicting protein pairs with the target class that does not interact. MMI types 2 and 4 have a reasonably good specificity value. However, the value is still below that of MMI type 1, but the average specificity difference between type 2 and type 4 with type 3 is also insignificant (about 0.2%); thus, we concluded that the three types do not exhibit significant differences in the ability to predict protein pairs that do not interact. MMI type 3 has an advantage in predicting the pair of interacting proteins, but in terms of accuracy, specificity and F1 score, MMI type 3 is still not suitable compared with MMI types 2 and 4. Thus, we obtained two types of MMI that can optimise the model, namely, MMI type 2 and type 4.

3.2. The results of autocorrelation methods

We continue our research by testing a combination of autocorrelation methods and SVM. There are six types of encoding methods used in the autocorrelation feature:

i) NMBAC with six physicochemical properties
ii) NMBAC with seven physicochemical properties
iii) MAC with six physicochemical properties
iv) MAC with seven physicochemical properties
v) GAC with six physicochemical properties
vi) GAC with seven physicochemical properties

As in the MMI method tests, we identified the C and gamma values that could optimise the accuracy of each model formed. In each encoding method of the autocorrelation feature, the C value that provides the best accuracy is 1, whereas the gamma value remains 0.1. The results for each type of autocorrelation method are presented in Table 4. When combined with SVM, the table demonstrates that the feature encoding methods with the autocorrelation concept provide an average prediction accuracy of above 72.9%. The value is superior to the average values of accuracy obtained when using the MMI feature encoding method. Similarly, in most other averages above 72.3%, only GAC methods with seven physicochemical properties produced an average sensitivity below 72.3%.

The MAC method with seven physicochemical properties and NMBAC also with seven physicochemical properties are two of the best feature encoding methods, as indicated by their average accuracy, which reached 73%. Contrary to the MAC method with 7 physicochemical properties which consistently has average evaluation values above 73%, the NMBAC method with 7 physicochemical properties has a poor average value in terms of sensitivity. However, the average accuracy value and F1 score of this method (72.94%) are quite larger compared with those of other methods. Thus, this method can also be evaluated as a pretty good method in terms of precision and recall.

Table 4. The results of MMI, autocorrelation and mixed methods

Feature Encoding Method	Type	Average of			
		Accuracy	Sensitivity	Specificity	F1 Score
MMI	Type 1	0.7090	0.6946	0.7232	0.7062
	Type 2	0.7241	0.7275	0.7210	0.7221
	Type 3	0.7191	0.7312	0.7072	0.7207
	Type 4	0.7220	0.7229	0.7212	0.7224
Autocorrelation	NMBAC 6 PP	0.7290	0.7331	0.7252	0.7280
	NMBAC 7 PP	0.7301	0.7268	0.7333	0.7294
	MAC 6 PP	0.7282	0.7324	0.7238	0.7317
	MAC 7 PP	0.7313	0.7326	0.7300	0.7342
	GAC 6 PP	0.7278	0.7313	0.7243	0.7277
	GAC 7 PP	0.7293	0.7150	0.7435	0.7239
Mixed	Mixed 1	0.7236	0.7080	0.7390	0.7187
	Mixed 2	0.7184	0.7122	0.7248	0.7180
	Mixed 3	0.7188	0.7269	0.7109	0.7188
	Mixed 4	0.7291	0.7229	0.7356	0.7299

3.3. The results of mixed MMI+autocorrelation methods

The above results indicate that the two types of MMI methods that yield the best results when combined with SVM are MMI type 2 and type 4, whereas for the autocorrelation type, the two methods of feature encodings that yield the best results are NMBAC and MAC with seven physicochemical properties. Thus, the four types of combined method feature encodings are as:

i) Mixed 1: MMI Type 2 and MAC 7 physicochemical properties
ii) Mixed 2: MMI Type 4 and MAC 7 physicochemical properties
iii) Mixed 3: MMI Type 2 and NMBAC 7 physicochemical properties
iv) Mixed 4: MMI Type 4 and NMBAC 7 physicochemical properties

Each of these mixed methods is retested using SVM, and the results are presented in Table 4. Each type uses the parameters C=1 and gamma=0.1. The values of C and gamma were selected using the same methods as those in the selection of parameter values in previous tests.

From Table 4, it can be seen that if the MAC (7 physicochemical properties) feature encoding method yields better results when combined with MMI type 2 than MMI type 4, as the average values of accuracy, specificity and F1 score obtained by mixed 1 bigger than the values obtained by mixed 2 feature encoding method. With regard to the average sensitivity value, the combination of MAC and MMI type 4 is better than that of MAC and MMI type 2, but the values of sensitivity do not matter when compared with the accuracy and specificity values considering that the purpose of the computing method is to filter and reduce pairs that are unlikely to interact. Furthermore, the encoding method of the NMBAC feature is more suitable when combined with MMI type 2 than MMI type 4, which is also seen from the average values of accuracy, specificity and F1 score. Overall, the combination of NMBAC and MMI type 4 provides the best average accuracy values and F1 score compared with other methods, whereas the combination of MAC and MMI type 2 is superior in terms of specificity to other combinations of MMI and autocorrelation methods.

3.4. Discussions

We compare the results of the research to identify the best way to develop MMI and autocorrelation methods for predicting protein-protein interaction. There are four ways to build MMI methods by using combination of two type of classification of 20 amino acids and two formula define to calculate MMI. To build autocorrelation methods, there are six ways that we try that produce by combination of number of physicochemical properties we used and the formula of autocorrelation we used.

3.4.1. Division of 20 amino acids

To determine the best way to classify 20 types of amino acids to be used in developing the MMI feature encoding method, we compared the results of the MMI type 1 method with those of the MMI type 3 method. Then, we compared the MMI type 2 method with the MMI type 4 method. The type 1 and 2 methods use the classification of amino acids based on the dipole scale and volume of the side chain of each amino acid, whereas the type 3 and 4 methods use amino acid classification based on the carbon chain structure of each amino acid. From the results of the comparison of each of the corresponding types, we concluded that the selection of how twenty amino acids being classified in building the MMI feature encoding method has little effect on the model being built, can be seen from the inconsistency of the accuracy obtained, where the MMI type 3 is superior to the MMI type 1 while the MMI type 2 provides greater accuracy than MMI type 4.

3.4.2. Defining of the $f(a)$ function in the MMI

We compared MMI type 1 to MMI type 2 and also MMI type 3 to MMI type 4to determine the best equation to represent a particular element in a sequence of amino acids. By comparing the results, we found that (7) yields better results than (6). The average accuracy and F1 score of MMI types 2 and 4 were higher than those of MMI types 1 and 3, respectively. Thus, we recommended the use of $f(a) = \frac{n_a+1}{L+c}$ to represent a ratio of element a in the amino acid sequences to obtain the values of MI and MMI.

3.4.3. Use of physicochemical properties

We compared the result obtained from NMBAC, MAC and GAC method with 6 physicochemical properties and those using 7 physicochemical properties. From the results of the comparison, it can be concluded that to convert amino acid sequences into a vector form using the autocorrelation method, the use of seven physicochemical properties is recommended. This is because based on Table 4, all autocorrelation methods that apply the seven physicochemical properties yield better average values of accuracy, specificity and F1 score than those using only six physicochemical properties. However, suppose that you want to predict which pair of proteins interact. In that case, the use of six physicochemical properties yields good results. The application of machine learning in the screening process is recommended to reduce the non-interacting pairs. The values of specificity and accuracy are superior to that of sensitivity.

3.4.4. Qualitative vs. quantitative descriptors of the SVM classifier

By comparing the values in Table 4, it can be observed that almost in various aspects of quantitative encoding method (autocorrelation), superior to qualitative type method (MMI). We assume that quantitative methods are better as they uniquely describe each amino acid. Thus, these methods are better for studying amino acid sequences than the qualitative methods, which consider some acids to be the same element as they have similarities in certain respects.

3.4.5. Separated vs. combined methods with the SVM classifier

Based on the values in Table 4, the combination of qualitative and quantitative encoding methods does not always yield better results than these methods separately. The combination of MMI type 4 and NMBAC 7 physicochemical properties has an average accuracy of 72.91%. This value of average accuracy is smaller than that obtained through the application of the NMBAC type 7 physicochemical properties method alone, which has an average accuracy of 73.13%. Meanwhile, the average values of accuracy of the combination method are higher than those of the MMI method, which was 72.41%. This can be explained by the fact that the SVM machine learning model, MMI and autocorrelation values are used as coordinates in the hyperplane. Combining two different encoding methods will cause the mapping of points on the SVM operations to vary and become more complex. When two feature encodings are different from each other in the mapping coordinates, the hyperplane formed becomes less optimal and leads to a more significant error in the division of data into two classes. Thus, in some cases, the combination of the two feature encoding methods does not produce better results compared with the separate implementation of the methods. The combination of MMI and autocorrelation methods yields better results in the prediction of protein pairs that do not interact due to the sensitivity value of the combination of MMI type 2 and MAC 7 physicochemical properties (73.90%). The combination of MMI type 4 and NMBAC 7 PP (73.56%) is superior to MMI type 2 alone (72.10%), MMI type 4 alone (72.12%), NMBAC 7 physicochemical properties alone (73.33%) and MAC 7 physicochemical properties alone (73%). Thus, the combination of MMI and autocorrelation methods is quite good if used for reducing protein pairs that are unlikely to interact. This will make research on protein interactions efficient.

3.4.6. Proposed method vs. previous research using the SVM classifier

Based on the dataset (HIV-1 and human) and types of machine learning techniques used, we compared the results obtained using the MMI and autocorrelation method with those obtained by Bustamam *et al*. [25], which are presented in Table 5. In their paper, a combination of pseudoSMR and SVM and global encoding and SVM were tested. Both methods yielded the best accuracy value below 70%, whereas our methods obtained an average accuracy of above 70%. The specificity obtained by both of global encoding and pseudoSMR encoding methods are very small, where this is not in line with the purpose of the method developed that is to eliminate protein pairs that have the possibility of the target class does not interact. Although not good enough in the average of sensitivity value when compared by the sensitivity value obtained by global encoding method, our method gives a more balanced result, where the average values of sensitivity and specificity are above 70%. This result also indicates that our methods can classify well positive and negative data and even better predict protein pairs that do not interact. For the F1 score, we obtained a value of 72.21%-73.42%, which was higher than that obtained in previous research wherein the pseudoSMR method and global encoding method got the best F1 scores of 67.76% and 71.50%, respectively. These values indicate that the MMI encoding methods, autocorrelation method and mixed methods provide a better percentage value in terms of precision and recall

Table 5. Comparison between previous research and proposed methods

Feature encoding method	Average of			
	Accuracy	Sensitivity	Specificity	F1 Score
PseudoSMR (Bustamam *et al*, 2019)	0.6555	0.7145	0.5950	0.6776
GE (Bustamam *et al*, 2019)	0.6201	0.9433	0.2902	0.7150
MMI	0.7241	0.7275	0.7210	0.7221
MAC 7	0.7313	0.7326	0.7300	0.7342
MMI+NMBAC	0.7291	0.7229	0.7356	0.7299

3.4.7. Performance MMI and autocorrelation method in other classifiers

We also tested MMI, autocorrelation and mixed methods using other machine learning classifiers, such as RF, gradient boosting and k-nearest neighbours (KNN). We compared the results of these experiments with those obtained by Bustamam *et al*, 2019 [25], which are presented Table 6. The results in Table 6 indicate that when combined with RF, the separate application of autocorrelation and MMI methods yields better results than when they are combined. From the results obtained, when combined with RF the MMI and autocorrelation methods were not found to be better than the global encoding methods and pseudoSMR methods, but the result of MMI and autocorrelation methods are more balanced than global encoding and PseudoSMR methods.

Contrary to the results obtained using RFs, when combined with the machine learning gradient boosting or KNN methods, the MMI and autocorrelation methods yielded better results than the global encoding and pseudoSMR methods. As can be seen from Table 6, most of the average evaluation values that we obtained when using the MMI and autocorrelation methods were above 70% in terms of accuracy, sensitivity, specificity and F1 score. While the use of global encoding and pseudoSMR provide results where the majority of the evaluation measurement values are below 70%. In the use of the machine learning gradient boosting method, only the sensitivity value and F1 score were recorded to be above 70%; however, this sensitivity and F1 score value is still below sensitivity values and F1 score from MMI and autocorrelation. Contrary to the results of other machine learning methods that are pretty balanced in terms of sensitivity and specificity values, when combined with KNN, the specificity value in the MMI, autocorrelation and mixed feature encoding methods was much higher than the specificity values obtained. The specificity values obtained by autocorrelation method and the mixed methods are above 74.2%, where these values are the highest score we obtain in our study. Thus, it can be said that the model formed from the combination of the MMI feature encoding method and autocorrelation with machine learning method KNN can predict protein pairs that do not interact well. From Table 6, it can be seen that the combination of the MMI and autocorrelation feature encoding methods works better in predicting protein interactions than using these encoding methods separately, as demonstrated by the average values of the evaluation measurement using a combination of methods higher than the use of these methods separately. In the machine learning method kNN, the results demonstrated no significant difference in terms of performance between the use of methods combined and each.

Table 6. Results of gradient boosting classifier

Machine learning classifier	Method		Average			
			Accuracy	Sensitivity	Specificity	F1 Score
Random Forest (RF)	Our	MMI 2	0.7361	0.7410	0.7311	0.7379
		MMI 4	0.7316	0.7366	0.7278	0.7342
		MAC 7 PP	0.7305	0.7340	0.7271	0.7305
		NMBAC 7 PP	0.7341	0.7277	0.7407	0.7343
		Mixed 2	0.7280	0.7355	0.7207	0.7286
		Mixed 3	0.7304	0.7350	0.7258	0.7316
	Bustamam [25]	GE	0.7684	0.7272	0.8104	0.7603
		PseudoSMR	0.7689	0.7261	0.8125	0.7605
Gradient Boosting	Our	MMI 2	0.7102	0.7059	0.7144	0.7085
		MMI 4	0.7086	0.7163	0.7007	0.7120
		MAC 7 PP	0.7069	0.7057	0.7081	0.7079
		NMBAC 7 PP	0.7095	0.7181	0.7010	0.7116
		Mixed 2	0.7305	0.7367	0.7243	0.7311
		Mixed 3	0.7304	0.7336	0.7273	0.7320
	Bustamam [25]	GE	0.6800	0.6916	0.6682	0.6859
		PseudoSMR	0.6975	0.7246	0.6698	0.7076
k-Nearest Neighbours' (KNN)	Our	MMI 2	0.6973	0.6655	0.7294	0.6878
		MMI 4	0.7124	0.7027	0.7221	0.7108
		MAC 7 PP	0.7154	0.6647	0.7684	0.7008
		NMBAC 7 PP	0.7113	0.6715	0.7512	0.6994
		Mixed 2	0.7057	0.6492	0.7627	0.6889
		Mixed 3	0.7176	0.6924	0.7420	0.7085
	Bustamam [25]	GE	0.6188	0.6488	0.5882	0.6323
		PseudoSMR	0.6686	0.7068	0.6294	0.6836

4. CONCLUSION

The use of encoding methods MMI and autocorrelation features yields good results in the prediction of protein-protein interactions. The MMI method gives optimum results when the ratio of an element in a sequence of amino acids is defined using formula that we developed. The conclusion we obtained as the result of MMI type 2 and MMI type 4 which are better than the other types of MMI. Therefore, the methods we propose gives better result than the previous research that uses MMI type 1 as feature encoding method. Then, for building a feature encoding method using autocorrelation concept, the use of seven physicochemical properties gives better result than those using only six physicochemical properties. Combination of two or more encoding feature methods not always gives better result than those methods separately. Furthermore, the new design of MMI and autocorrelation feature encoding methods provides the better results rather than the previous design, and also are better than the other feature encoding methods when combined with several machine learning methods. Overall, the improved MMI and autocorrelation methods give good and balanced performance in predicting protein-protein interactions.

SUPPLEMENTARY FILES

We share all the results and all the dataset of pairs interaction protein between HIV-1 and human at the link http://ipmuonline.com/supp_files/IJAI/IJAI_21472_SP.zip.

ACKNOWLEDGEMENTS

This research is supported by the Q2 International Publication Assistance Grant 2021 from Directorate of Research and Community Service, Universitas Indonesia with contract number NKB-589/UN2.RST/HKP.05.00/2021.

REFERENCES

[1] H. Lodish *et al.*, *Molecular cell biology*, 8th editio. New York : W.H. Freeman-Macmillan Learning, 2016.

[2] J. De Las Rivas and C. Fontanillo, "Protein-protein interactions essentials: key concepts to building and analyzing interactome networks," *PLoS Computational Biology*, vol. 6, no. 6, Art. no. e1000807, Jun. 2010, doi: 10.1371/journal.pcbi.1000807.

[3] İ. Kösesoy, M. Gök, and C. Öz, "A new sequence based encoding for prediction of host-pathogen protein interactions," *Computational Biology and Chemistry*, vol. 78, pp. 170–177, Feb. 2019, doi: 10.1016/j.compbiolchem.2018.12.001.

[4] X. Tang, Q. Xiao, and K. Yu, "Breast cancer candidate gene detection through integration of subcellular localization data with protein-protein interaction networks," *IEEE Transactions on NanoBioscience*, vol. 19, no. 3, pp. 556–561, Jul. 2020, doi: 10.1109/TNB.2020.2990178.

[5] R. Ginanjar, A. Bustamam, and H. Tasman, "Implementation of regularized Markov clustering algorithm on protein interaction networks of schizophrenia's risk factor candidate genes," in *2016 International Conference on Advanced Computer Science and*

Information Systems (ICACSIS), Oct. 2016, pp. 297–302, doi: 10.1109/ICACSIS.2016.7872726.

[6] J. Li *et al.*, "Using weighted extreme learning machine combined with scale-invariant feature transform to predict protein-protein interactions from protein evolutionary information," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–1, 2020, doi: 10.1109/TCBB.2020.2965919.

[7] K. Yu, T. Zhao, P. Zhao, and J. Zhang, "Extraction of protein-protein interactions using natural language processing based pattern matching," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov. 2017, pp. 1292–1295, doi: 10.1109/BIBM.2017.8217847.

[8] A. Bustamam, T. Siswantining, T. P. Kaloka, and O. Swasti, "Application of BiMax, POLS, and LCM-MBC to find bicluster on interactions protein between HIV-1 and human," *Austrian Journal of Statistics*, vol. 49, no. 3, pp. 1–18, Feb. 2020, doi: 10.17713/ajs.v49i3.1011.

[9] P. P. Tampubolon, A. Bustamam, D. Lestari, and W. Mangunwardoyo, "A biclustering procedure using BicBin algorithm for HIV-1 human protein interaction database in NCBI," in *AIP Conference Proceedings*, 2019, Art. no. 020008, doi: 10.1063/1.5094272.

[10] Y. E. Göktepe and H. Kodaz, "Prediction of protein-protein interactions using an effective sequence based combined method," *Neurocomputing*, vol. 303, pp. 68–74, Aug. 2018, doi: 10.1016/j.neucom.2018.03.062.

[11] Y. Ding, J. Tang, and F. Guo, "Predicting protein-protein interactions via multivariate mutual information of protein sequences," *BMC Bioinformatics*, vol. 17, no. 1, Art. no. 398, Dec. 2016, doi: 10.1186/s12859-016-1253-9.

[12] Z.-H. You, Y.-K. Lei, L. Zhu, J. Xia, and B. Wang, "Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis," *BMC Bioinformatics*, vol. 14, no. 8, Art. no. 10, May 2013, doi: 10.1186/1471-2105-14-S8-S10.

[13] C. N. Kopoin, N. T. Tchimou, B. K. Saha, and M. Babri, "A feature extraction method in large scale prediction of human protein-protein interactions using physicochemical properties into Bi-gram," in *2020 IEEE International Conf on Natural and Engineering Sciences for Sahel's Sustainable Development-Impact of Big Data Application on Society and Environment (IBASE-BF)*, Feb. 2020, pp. 1–7, doi: 10.1109/IBASE-BF48578.2020.9069594.

[14] S. Mohammadi, V. Desai, and H. Karimipour, "Multivariate mutual information-based feature selection for cyber intrusion detection," in *2018 IEEE Electrical Power and Energy Conference (EPEC)*, Oct. 2018, pp. 1–6, doi: 10.1109/EPEC.2018.8598326.

[15] D. Lestari, S. Aprilia, and A. Bustamam, "Performance analysis of support vector machine combined with global encoding on detection of protein-protein interaction network of HIV virus," in *AIP Conference Proceedings*, 2018, Art. no. 020228, doi: 10.1063/1.5064225.

[16] X. Wang, R. Wang, Y. Wei, and Y. Gui, "A novel conjoint triad auto covariance (CTAC) coding method for predicting protein-protein interaction based on amino acid sequence," *Mathematical Biosciences*, vol. 313, pp. 41–47, Jul. 2019, doi: 10.1016/j.mbs.2019.04.002.

[17] Z. Li, J. Tang, and F. Guo, "Identification of 14-3-3 proteins phosphopeptide-binding specificity using an affinity-based computational approach," *Plos One*, vol. 11, no. 2, p. e0147467, Feb. 2016, doi: 10.1371/journal.pone.0147467.

[18] X. Du, J. Cheng, T. Zheng, Z. Duan, and F. Qian, "A novel feature extraction scheme with ensemble coding for protein-protein interaction prediction," *International Journal of Molecular Sciences*, vol. 15, no. 7, pp. 12731–12749, Jul. 2014, doi: 10.3390/ijms150712731.

[19] M. Moukhafi, K. El Yassini, and B. Seddik, "Intrusions detection using optimized support vector machine," *International Journal of Advances in Applied Sciences*, vol. 9, no. 1, p. 62, Mar. 2020, doi: 10.11591/ijaas.v9.i1.pp62-66.

[20] O. W. Chuan, N. F. Ab Aziz, Z. M. Yasin, N. A. Salim, and N. A. Wahab, "Fault classification in smart distribution network using support vector machine," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 3, pp. 1148–1155, Jun. 2020, doi: 10.11591/ijeecs.v18.i3.pp1148-1155.

[21] C. M. Bishop, "Sparse kernel machines," in *Pattern Recognition and Machine Learning*, 2006.

[22] A. Maslan, K. M. Bin Mohamad, and F. Binti Mohd Foozy, "Feature selection for DDoS detection using classification machine learning techniques," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 9, no. 1, pp. 137–145, Mar. 2020, doi: 10.11591/ijai.v9.i1.pp137-145.

[23] N. M. G. Dwi Purnamasari, M. A. Fauzi, I. Indriati, and L. S. Dewi, "Cyberbullying identification in twitter using support vector machine and information gain based feature selection," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 3, pp. 1494–1500, Jun. 2020, doi: 10.11591/ijeecs.v18.i3.pp1494-1500.

[24] E. A. Gheni and Z. M. Algelal, "Human face recognition methods based on principle component analysis (PCA), wavelet and support vector machine (SVM) : a comparative study," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 20, no. 2, pp. 991–999, Nov. 2020, doi: 10.11591/ijeecs.v20.i2.pp991-999.

[25] A. Bustamam, M. I. S. Musti, S. Hartomo, S. Aprilia, P. P. Tampubolon, and D. Lestari, "Performance of rotation forest ensemble classifier and feature extractor in predicting protein interactions using amino acid sequences," *BMC Genomics*, vol. 20, no. S9, Art. no. 950, Dec. 2019, doi: 10.1186/s12864-019-6304-y.

BIOGRAPHIES OF AUTHORS

Assoc. Prof. Alhadi Bustamam Ph.D. Head of post graduate program in department of mathematics in Universitas Indonesia (UI). Chairman of Data Science Center (DSC) Lembaga Sains Terapan FMIPA UI, also co-founder of Bioinformatics and Advanced Computing Laboratory, Department of Mathematics FMIPA UI. Graduated from Universitas Indonesia (UI) in 1996 (Bachelor of Computational Mathematics) and 2002 (Master of Computer Science in Parallel Computing). Received PhD in 2012 from Institute for Molecular Biosciences (IMB), The University of Queensland (UQ) in the field of Bioinformatics. He can be contacted at email: alhadi@sci.ui.ac.id.

Mohamad Irlin Sunggawa, S. Pd, M. Si Graduated of post graduate program in department of mathematics Universitas Indonesia, Depok, Indonesia in 2021. Received his bachelor degree in Mathematics Education, Universitas Islam Negeri Sunan Gunung Djati, Bandung, Indonesia, 2012-2016. He can be contacted at email: mohamad.irlin@sci.ui.ac.id.

Dr. Titin Siswantining, D.E.A Received her Bachelor degree in Statistics from Sepuluh Nopember Institute of Technology, Indonesia in 1984, Master degree in Applied Mathematics from EHESS Paris in 1990 and Doctoral degree in Statistics from Bogor Agricultural Institute in 2013. Her research focuses on Applied Statistics and Bioinformatics. Currently, she is a Senior Research Fellow of BACL and working as an Associate Professor in the Department of Mathematics at Universitas Indonesia. She can be contacted at email: titin@sci.ui.ac.id.