

DOI : [http://doi.org/10.22438/jeb/40/3\(SI\)/Sp-21](http://doi.org/10.22438/jeb/40/3(SI)/Sp-21)

# Optimized bio-inspired kernels with twin support vector machine using low identity sequences to solve imbalance multiclass classification

Paper received: 17. 04. 2018

Revised received: 14. 11.2018

Accepted: 03.01. 2019

## Authors Info

S.K. Guramand<sup>1\*</sup>,  
R.D.R. Saedudin<sup>2</sup>, R. Hassan<sup>1</sup>,  
S. Kasim<sup>3</sup>, R. Ramlan<sup>4</sup> and  
B. W. Salim<sup>5</sup>

<sup>1</sup>School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia

<sup>2</sup>School of Industrial Engineering, Telkom University, 40257 Bandung, West Java, Indonesia

<sup>3</sup>Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400, Johor, Malaysia

<sup>4</sup>Faculty of Technology Management and Business, Universiti Tun Hussein Onn Malaysia, 86400, Johor, Malaysia

<sup>5</sup>College of Engineering, Nawroz University, Duhok, Iraq

\*Corresponding Author Email : [shahreen@uthm.edu.my](mailto:shahreen@uthm.edu.my)

## Edited by

Prof. Muhammad Aqeel Ashraf

## Reviewed by

Dr. Hongbo Zhang  
Dr. Xiaopeng Jia

## Abstract

The function of enzymes is performed differently depending on their bio-chemical mechanisms and important to the prediction of protein structure and function. In order to overcome the weaknesses of imbalance data distribution in subclasses prediction we proposed Bio-Twin Support Vector Machine (Bio-TWSVM). The TWSVM approach also allows for kernel optimization where in this study we have introduced the bio-inspired kernels such as the Fisher, spectrum and mismatch kernels which at the same time incorporate the biological information regarding the protein evolution in the classification process.

**Key words :** Bio-inspired kernels, Bio-Twin Support Vector Machine, Enzymes, Fisher, Kernel optimization

**Citation:** Guramand, S.K., R.D.R. Saedudin, R. Hassan, S. Kasim, R. Ramlan and Baraa Wasfi Salim: optimized bio-inspired kernels with twin support vector machine using low identity sequences to solve imbalance multiclass classification. *J. Environ. Biol.*, **40**, 563-576 (2019).

## Introduction

The enzymes function differently depending on their biochemical mechanisms. Eventually, this has led to the genesis of an interesting problem in Bioinformatics, which how can we classify a given protein sequence as an enzyme and accurately predict its function. In light of the key biological role of enzyme proteins, Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) has created a hierarchical classification scheme based on the functional mechanism of enzymes (Yehia, 2017; Hanedar *et al.*, 2017). Through protein enzyme functional and sub-functional class prediction, biologists are able to determine the biological function of each protein which can be attained from the biologically inspired functional kernel. On the other hand, as a quick and simple measure, sequence identity is also widely used as an indication of functional similarity in dividing enzymes family into subfamilies. However, due to lack of a rigorously established sequence identity threshold, the division of an enzyme family into subfamilies may require human intervention (Ong *et al.*, 2017; Daya *et al.*, 2017). Besides, the enzyme sub-functional class prediction is an imbalance prophecy. Therefore, it is of great significance to establish the threshold of sequence identity above which functional similarity using Twin SVM as classifier and bio-inspired kernel function can be affirmed to overcome the imbalance classification problem by introducing a novel technique.

Various machine learning algorithms have been proposed for enzymes functional and sub-functional class prediction such as Support Vector Machine (SVM), Decision trees, Naïve Bayesian (NB), K-Nearest Neighbor (KNN), C4.5, Random Forest and Artificial Neural Network (ANN). All of these studies have used amino acid compositions derived from sequence and employs the technique that uses protein sequences to compute their input feature vectors and predictor models to predict enzyme main and sub-functional class. Most of these previous studies are limited to the scope of enzyme sequences lesser than the current available number of sequences for every class and utilizes  $\leq 40\%$  rate of sequence similarities. To date, SVM has been used extensively and provides more accurate results in most of the studies done previously. Unfortunately, when faced with imbalanced problem, the performance of SVM drops significantly and it should be pointed out that the prediction of enzyme sub-functional class is an imbalance multi-class classification problem due to the fact that the number of proteins in each subclass makes a great difference. In this study, we have introduced the modified version of SVM, Twin SVM (TWSVM) incorporated into Bio-TWSVM (TWSVM with Bio-inspired kernels) in enzyme sub-functional classification. TWSVM can construct two nonparallel hyperplanes; positive and negative hyperplane without restriction of parallel, has lots of advantages such as high efficiency, high

detection rate and low false positive rate and achieves an overall accuracy above 90%. Although the results are promising, these studies use the basic kernel function for optimization (H'ng *et al.*, 2018; Hasan, 2018).

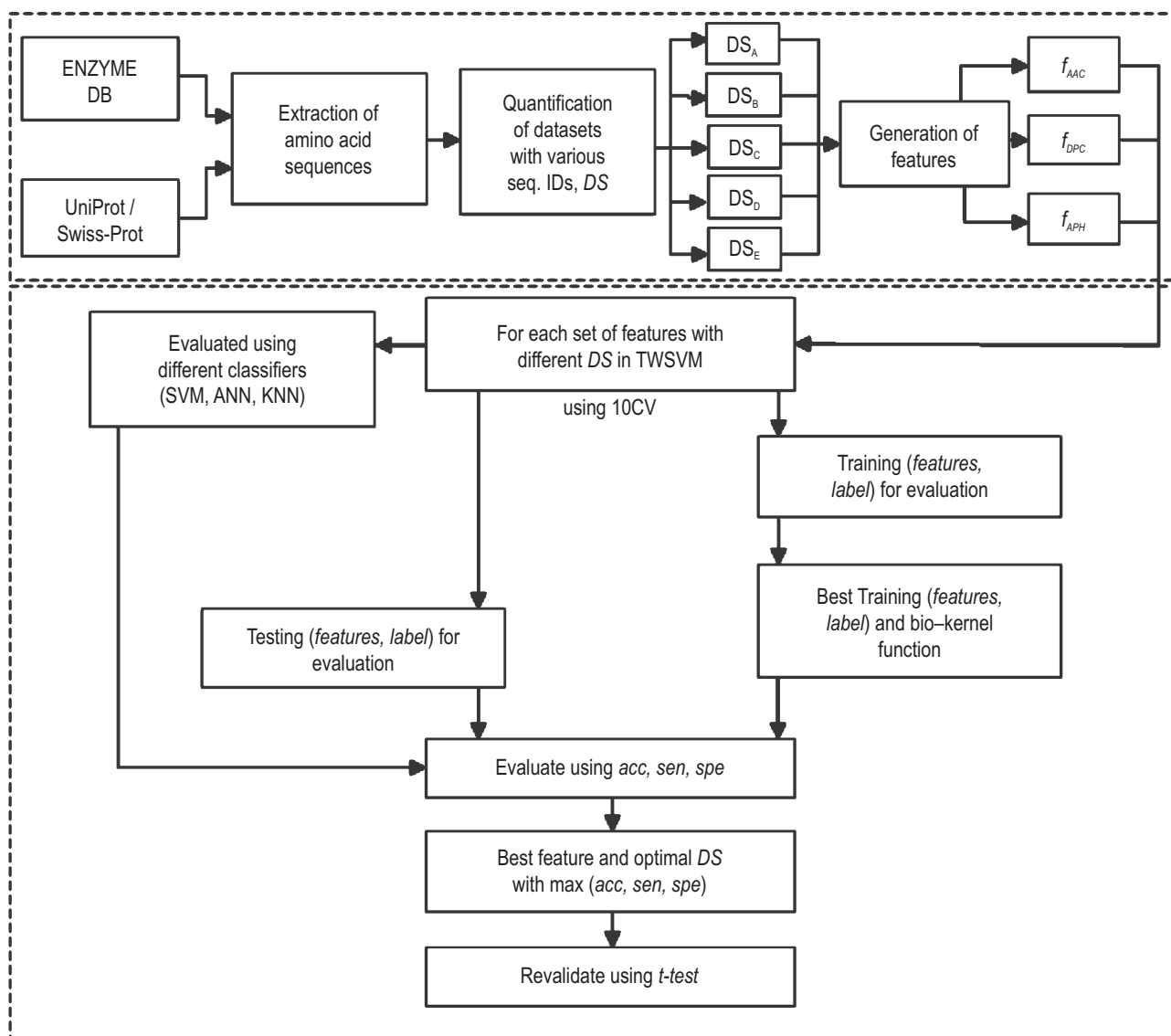
Kernel function as the core in SVM can be categorized into two categories: standard and bio-inspired. However, different kernel functions of SVM will give different result in protein enzymes sub-functional class prediction. Currently, there are four standard kernel functions in SVM, namely linear, polynomial, Radial Basis Function (RBF) and sigmoid. On the other hand, well-known bio-inspired kernel functions are Fisher, spectrum, mismatch oligomer distances multi-scale Gaussian and bi-cubic interpolation. Bio-inspired kernel functions are capable of classifying protein with more biological meaning and produce low error rate since it includes biological information in the classification process. These kernels are mainly applied in a multi-class classification strategy that applies a multi-class classification on each kernel score space and combines the decisions of multi-class classifiers. Experimentally, it has been shown that the kernel scores of one classes provide discriminative information for the other classes as well. However, the bio-inspired kernels have never been applied in any enzyme functional or sub-functional class study.

This study proposed a holistic classification named Bio-TWSVM to solve the imbalance classification problem in enzyme sub-functional prediction. The Bio-TWSVM is named based on TWSVM and consists of two components, which are two nonparallel hyperplanes classifier and Bio-kernel that is capable to classify protein sequences from main functional class up to sub-functional class level corresponding to the ENZYME database and incorporate biological information in the classification process using APH as feature representation. In this study, the results of prediction were analyzed based on three different bio-kernels namely Fisher, spectrum and mismatch. This is because different kernels function will lead to different discriminative functions and different performance.

## Materials and Methods

The solution as per shown in Fig. 1 has three major elements to attain: (i) the optimal rate of sequence similarity; (ii) the most significant feature and (iii) the distinguishable enzyme subclasses. In the first element, we managed to counter the adversary of the low sequence similarity rate based on different datasets. The second element was proposed to devise a hybrid feature, labelled APH that integrates AAC, DPC, hydrophobicity and hydrophilicity scores. In the third element, TWSVM was utilized with bio-inspired kernel function to solve the existing imbalance class distribution.

**Extraction of amino acid sequences:** The datasets used in this study were retrieved from ENZYME database at <ftp://ftp.expasy.org/databases/enzyme/> (Release of 19-Oct-2011) where the sub-



**Fig. 1 :** Bio-TWSVM embodies steps of preparation of datasets and features (top), determination of the most significant feature (best feature and classifier) and the optimal sequence identities (bottom).

**Table 1:** Datasets used in the study

Rate of sequence similarities	10%(DS <sub>A</sub> )	20%(DS <sub>B</sub> )	25%(DS <sub>C</sub> )	30%(DS <sub>D</sub> )	40%(DS <sub>E</sub> )
<b>Dataset</b>					
Number of sequences	1,570	3,121	4,732	5,698	10,442
Number of functional class	6	6	6	6	6
Number of sub-functional class	22	48	58	44	50

functional classes were classified into each of the six main enzyme classes, based on the accession numbers extracted. The corresponding protein sequences represented by their EC number were taken from the databank of Uniprot/Swiss-Prot at <http://www.ebi.ac.uk/swissprot/> (Release of 21-Sept-2011). The

experiment used five different pairwise sequence identities (ID) datasets, which were 10%, 20%, 25%, 30% and 40%; used to evaluate the performance of imbalance class distribution. For instance, 25% pairwise sequence ID produced 5,588 protein sequences, grouped into 6 functional classes and 58 sub-

functional classes. The dataset was further reduced by keeping only the desired protein sequences where as those with less than 50 amino acids were excluded to avoid fragment data and the enzymes consisting of multi-domain proteins with multiple enzymatic functions were removed which then produces 4,732 sequences. Meanwhile, 40% sequence ID dataset was derived from (Wang *et al.*, 2010) for comparison purpose. In Bio-TWSVM, for each dataset, we split into training and testing sets accordingly and run the experiments using 10 cross validation processes and repeat each of the cross validation for 10 times for consistency and to find the best prediction value. To allow fair prediction, we ensure that the dataset had different functional classes. Table 1 summarizes the distribution of sequences across all the main classes and subclasses based on different sequence ID datasets.

**Quantification of datasets with various sequence identities (Ids):** In order to determine the most optimal dataset for given enzyme class, different percentage of sequence IDs were

experimented: 10%, 20%, 25%, 30% and 40% ( $DS_A-DS_E$ ) measured with BLAST (Ye *et al.*, 2012) as shown in Fig. 2. For each (query) protein  $q$  in the dataset, BLAST search was run against the dataset after retaining the setting to be default. Protein hits were then considered according to BLAST sorting (from the lowest to the highest E-values) and for each protein  $q$  in the dataset only the best-hit protein was retained from the BLAST list which proved the activity.

**Generation of input features:** Before the sequences could be fed into the TWSVM, it was converted into a feature vector which was applied for each dataset ( $DS_A-DS_E$ ). Feature vectors for every dataset was represented with AAC ( $f_{AAC}$ ), dipeptide composition ( $f_{DPC}$ ) and hydrophobicity/hydrophilicity ( $f_p$ ) which was the unification of amino acid and dipeptide composition, as well as hydrophobicity and hydrophilicity. Fig. 3 summarizes the process in obtaining the proposed feature vector using one protein sequence from EC.1.3 as example.

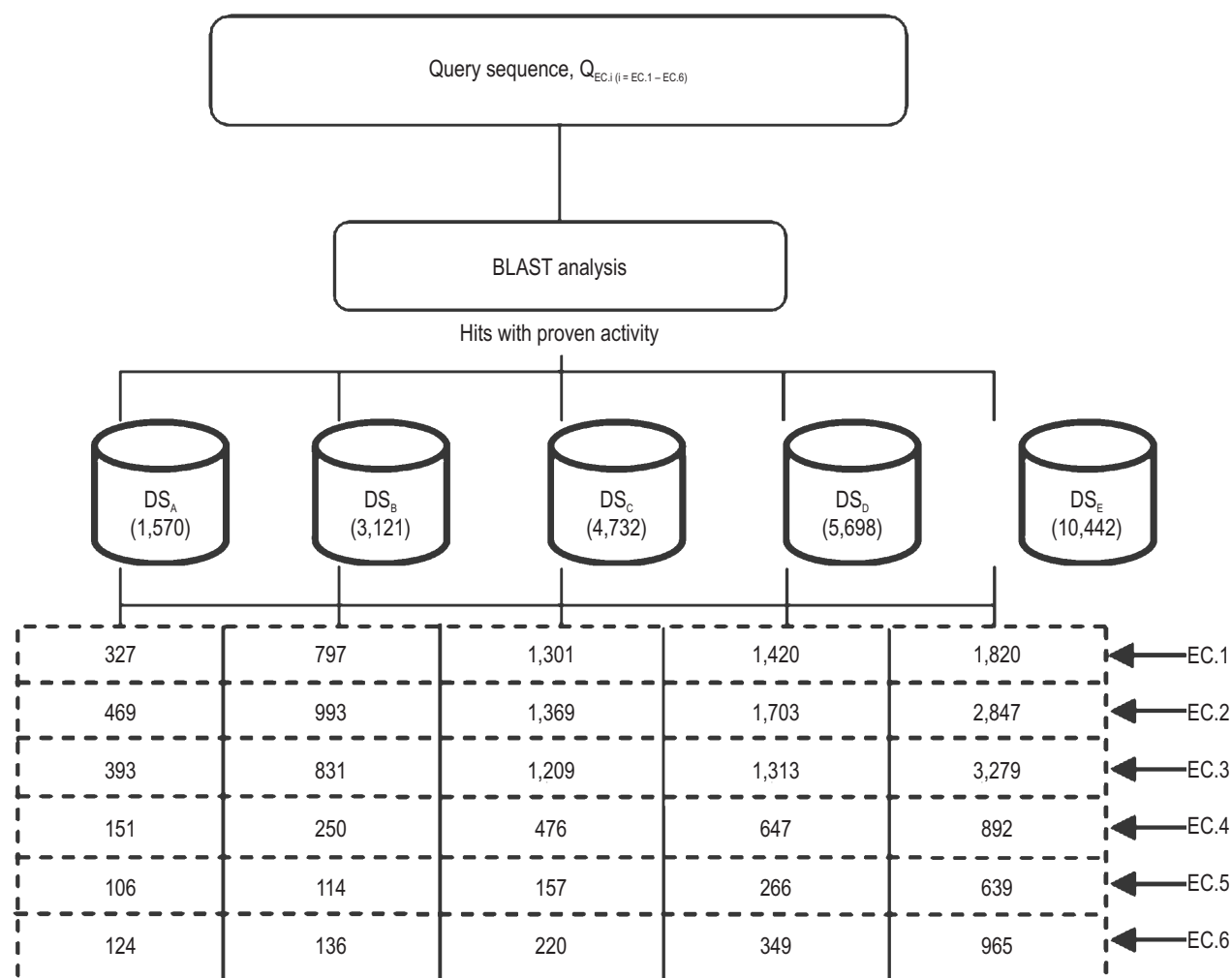


Fig. 2 : Quantification of datasets with different similarities using BLAST.

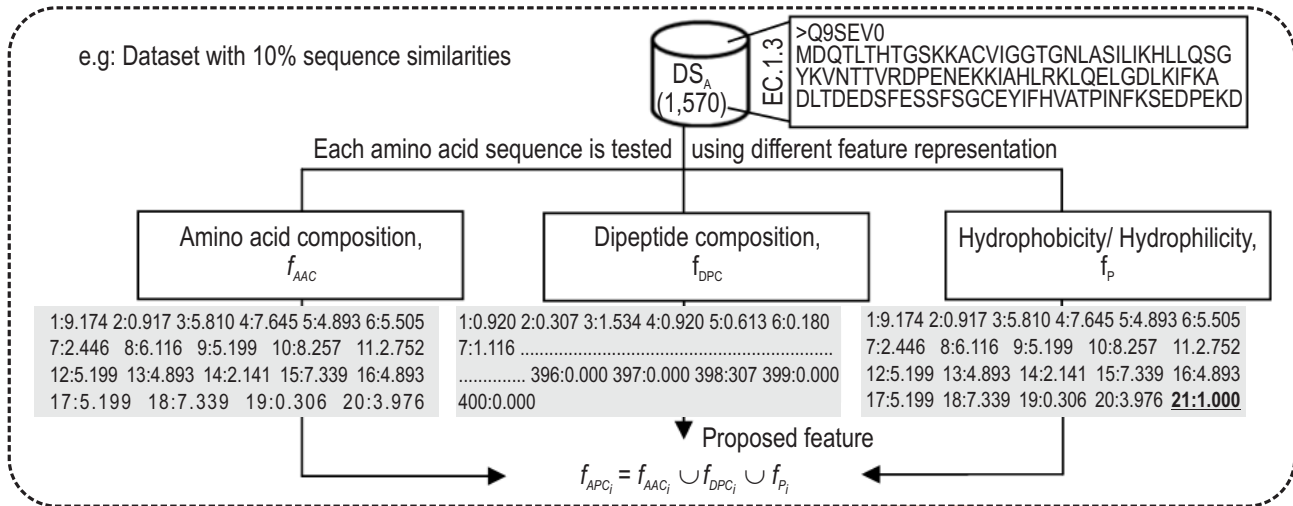


Fig. 3 : Three different feature vector representation used in this study.

**Generation of AAC:** AAC alone performs best with existing yet more complex features indicating the presence of sequence-level information that is predictive of interaction, but which is not necessarily restricted to domains. AAC is a fraction of each amino acid present in the protein sequence. Suppose a protein sequence with  $L$  amino acid residues:

$$R_1 R_2 R_3 R_4 R_5 \dots R_L \quad (1)$$

Where,  $R$  represents the amino acid residue and the subscript number represents the position of amino acid residue of length,  $L$  in a protein sequence. If  $\lambda$  is the length of protein sequence and  $\beta_i$  is the frequency of occurrence of an amino acid  $i$ , then AAC <sub>$i$</sub>  is:

$$AAC_i = \frac{\beta_i}{\lambda} \quad (2)$$

where,  $i$  is any of the 20 amino acids.

**Composition of Dipeptide:** In order to implement information about frequency as well as local order of residues in proteins, a dipeptide composition (DPC) based model was also constructed. DPC is considered as better feature as compared to AAC as it encapsulates global, as well as local information of the sequence. The DPC based model encompasses information about AAC along local order of amino acid. It gives a fixed pattern length of a vector with 400 (20x20) dimensions. The fraction of each dipeptide,  $DPC$ , was computed by the following equation:

$$DPC_i = \frac{\sigma_{ij}}{\omega} \quad (3)$$

where,  $i$  and  $j$  are any of the 20 amino acid residues,  $\sigma_{ij}$  is the fraction of a pair of amino acids ( $i, j = 1, 2, \dots, 20$ ) and  $\omega$  is the total number of all possible dipeptides.

**Generation of APH features:** The concept of Pse-AAC concerning the use of hydrophobicity and hydrophilicity factors was proposed in order to avoid a complete lost in the sequence order information. In contrast with the conventional AAC that contains 20 components with each reflecting the occurrence frequency for one of the 20 native amino acids in a protein, the essence of Pse-AAC is that it includes information beyond AAC where the first 20 represent the components of its conventional AAC, while the additional factors reflects the sequence order effect of a protein through a discrete model. Thus, according to the definition of Pse-AAC, a protein sequence can be expressed as a vector  $P$  which is formulated as follows:

$$P_i = \{P_1, \dots, P_{20}, P_{20+\lambda}\} \quad (4)$$

where the first 20 numbers in Eq. (4) represent the classic AAC, and the next  $\lambda$  discrete numbers describe sequence correlation factor which is the hydrophobicity and hydrophilicity values calculated based on Zhou *et al.* (2007) by the following equation:

$$h^1(i) = \frac{h_i^1(i) - T_1}{\sqrt{\sum_{i=1}^{20} [h_i^1(i) - T_1]^2}}, T_1 = \sum_{i=1}^{20} [h_i^1(i) / 20] \quad (5)$$

$$h^2(i) = \frac{h_i^2(i) - T_2}{\sqrt{\sum_{i=1}^{20} [h_i^2(i) - T_2]^2}}, T_2 = \sum_{i=1}^{20} [h_i^2(i) / 20] \quad (6)$$

where,  $i$  is the indices of amino acid residue; and are the original hydrophobic and hydrophilic values of the  $i^{\text{th}}$  amino acid. Therefore, the equation of APH can be expressed as:

$$f_{APH_i} = f_{AAC_i} \cup f_{DPC_i} \cup f_{P_i} \quad (7)$$

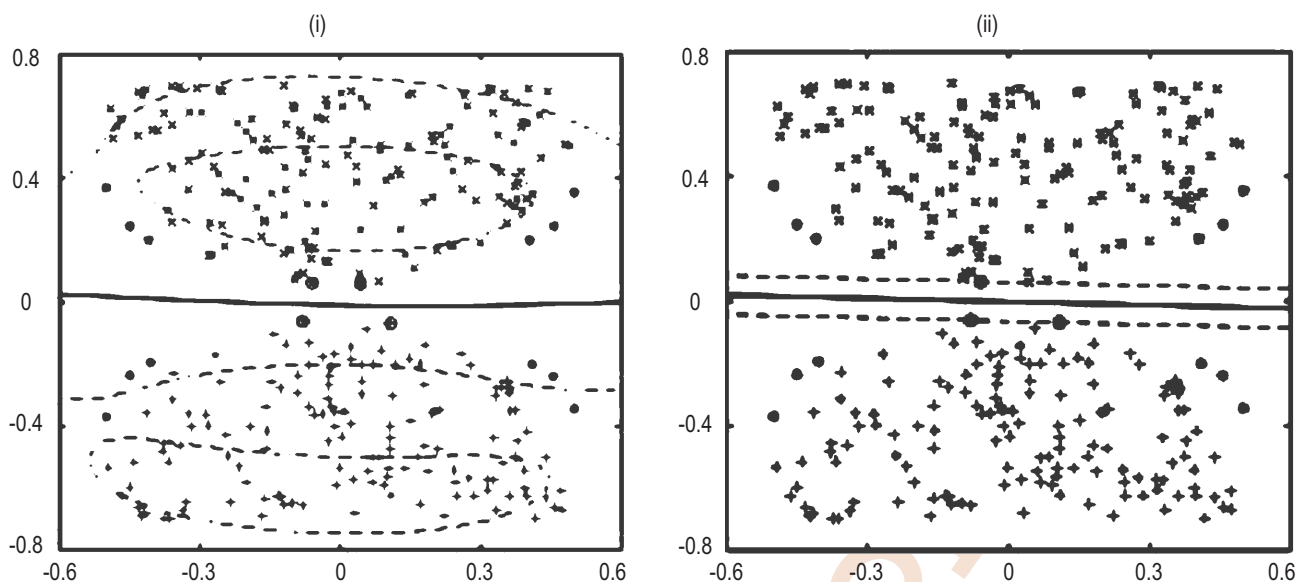


Fig. 4 : Comparison in separation of hyperplanes using (i) TWSVM and (ii) SVM.

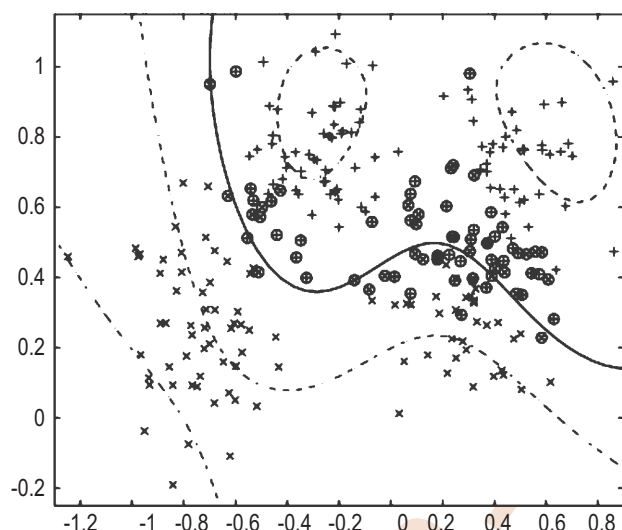


Fig. 5: Two nonlinear kernels generated based on Bio-TWSVM.

TWSVMs are comprised of a pair of quadratic programming problems (QPPs) such that, in each QPP, the objective function corresponds to a particular enzyme class and the constraints are determined by patterns of the other enzyme class. Thus, TWSVMs give rise to two smaller sized QPPs. In TWSVM1, patterns of true class are clustered around the plane  $x^T w^{(1)} + b^{(1)} = 0$ . Similarly, in TWSVM2, patterns of non-true class cluster around the plane  $x^T w^{(2)} + b^{(2)} = 0$ . The algorithm finds two hyperplanes as shown in Fig. 4, one for each enzyme class, and classifies points according to which hyperplane a given point is closest to by solving the following pair of QPPs:

(TWSVM1)

$$\begin{aligned} \text{Min}_{w^{(1)}, b^{(1)}, q} \quad & \frac{1}{2} (Aw^{(1)} + e_1 b^{(1)})^T (Aw^{(1)} + e_1 b^{(1)}) + c_1 e_2^T q \\ \text{s.t.} \quad & -(Bw^{(1)} + e_2 b^{(1)}) + q \geq e_2, q \geq 0, \end{aligned} \quad (8)$$

(TWSVM2)

$$\begin{aligned} \text{Min}_{w^{(2)}, b^{(2)}, q} \quad & \frac{1}{2} (Bw^{(2)} + e_2 b^{(2)})^T (Bw^{(2)} + e_2 b^{(2)}) + c_2 e_1^T q \\ \text{s.t.} \quad & -(Aw^{(2)} + e_1 b^{(2)}) + q \geq e_1, q \geq 0, \end{aligned} \quad (9)$$

where,  $c_1, c_2 > 0$  are the penalty parameters;  $e_1$  and  $e_2$  are column vectors of ones of appropriate dimensions and  $q$  is the slack variable which is used to measure the training loss. Penalty parameter  $c$ , with default value 0.01, is the parameter that controls the balance between training loss and margin. The sum of the linear slack variables,  $\sum q$ , which was adopted from the formulation in (Tsochantaridis et al., 2005) and (Scholkopf and Smola, 2002), were used and divided by  $n$  input length to better capture  $c$ , which scales with the training set size partitioned using 10 fold cross validation (CV) processes 1 partition set as the training set and train it to the other 9 partitions to learn and generate the predictions score on each of it.

**Kernel selection:** For nonlinear case, kernel TWSVM, as shown in Fig. 5, considers the following two kernel-generated nonparallel hyperplanes:

$$w_1^T K(C, x) + b_1 = 0 \quad \text{and} \quad w_2^T K(C, x) + b_2 = 0 \quad (10)$$

where,  $K(x, x')$  is known as a kernel function. In this study, Fisher kernel function was used and was derived from a hidden Markov model (HMM),

$$K(x, x') = e^{-\frac{1}{2\sigma^2}(U_x - U_{x'})^T (U_x - U_{x'})} \quad (11)$$

where,  $T$  is the Fisher information matrix. The Fisher kernel engenders a measure of similarity between two data items  $x_i$  and  $x_j$  by defining a distance between them and can be used with any kernel-based classifier such as a TWSVM. The scaling parameter  $\sigma$  appearing in this kernel was set equal to the median Euclidean distance between the gradient vectors corresponding to the training sequences in the enzyme subclass of interest and the closest gradient vector from a sequence belonging to another functional class.

To summarize, we began with a HMM trained from positive examples to model a given enzyme subclass. HMM was used to map each new protein sequence  $X$  that was to be classified into a fixed length vector, its Fisher score, and compute the kernel function on the basis of the Euclidean distance between the score vector for  $X$ , and the score vectors for known positive and negative examples  $X_i$  of the enzyme subclasses. The resulting discriminant function is given by

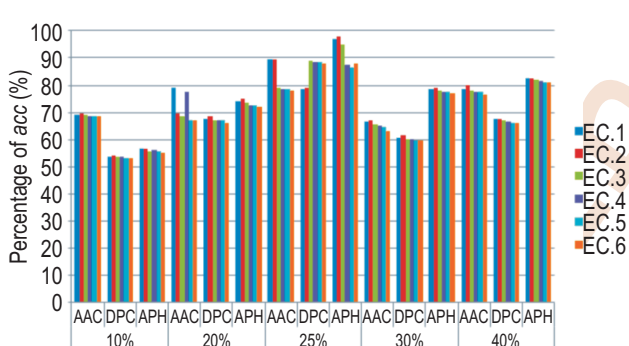
$$f(x) = \sum_{i, x_i \in X_+} \lambda_i K(x, x_i) - \sum_{i, x_i \in X_-} \lambda_i K(x, x_i) \quad (12)$$

where,  $K$  is the kernel function defined above and the  $\lambda_i$  are estimated from the positive ( $X_+$ ) and negative ( $X_-$ ) training examples  $X_i$ .

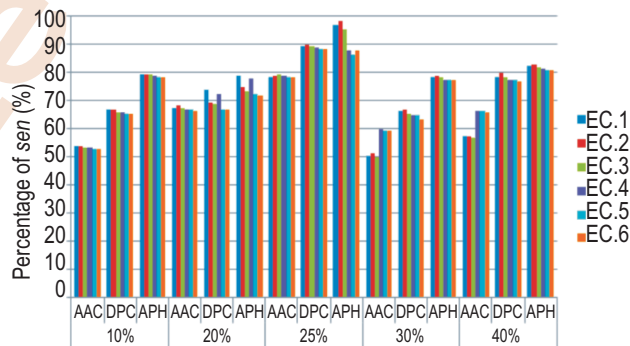
**Evaluation measures:** To assess the performance of the tested methods we counted the validation of the number of true positives (TP) number of correctly predicted sequences of enzyme class  $i$ , true negatives (TN) number of correctly predicted sequences not of enzyme class  $i$ , false positive (FP) number of incorrectly predicted sequences of enzyme class  $i$ , and false negatives (FN) number of incorrectly predicted sequences not of enzyme class  $i$ , where,  $i$  represents either one of the six enzyme classes. Then, we computed the following indices:

$$\text{Accuracy: } acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (13)$$

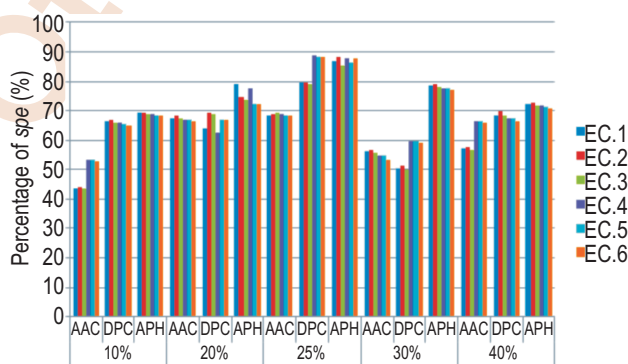
$$\text{Sensitivity: } sen = \frac{TP}{TP + FN} \quad (14)$$



(i) Rate of sequence similarities using different features by acc



(ii) Rate of sequence similarities using different features by sen



(iii) Rate of sequence similarities using different features by spe

**Fig. 6 :** Results for enzyme subclasses prediction in terms of (i) *acc*, (ii) *sen* and (iii) *spe* using different rate of sequence similarities.

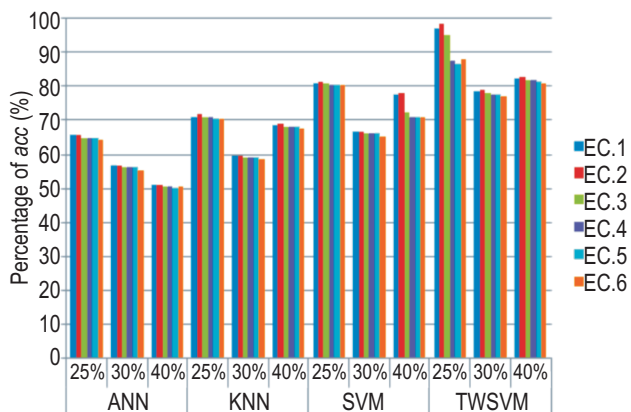


Fig. 7 : Results based on classifiers in terms of acc for various rate of sequence similarities.

Specificity: 
$$spe = \frac{TP}{TP + FP} \tag{15}$$

Besides, *t*-test was applied at 95% confidence level to revalidate the performance of the prediction outcome in term of statistical significance in order to assure that the prediction performance over *acc*, *sen* and *spe* is not just better but also substantial. A *t*-test is any statistical hypothesis test in which the test statistic follows a student's *t* distribution, if the null hypothesis is supported. In this research, *t*-test was applied on two samples of result which represents different features in every enzyme sub-functional class. Thus, if the table of confidence interval analysis reveals that the range between low and up bounds excludes 0, then the compared features are significantly different. Otherwise, the improvement gained is rendered as insufficient or meaningless.

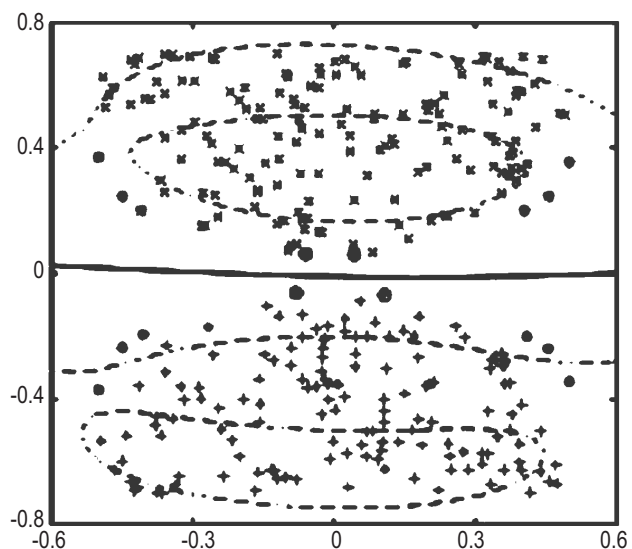
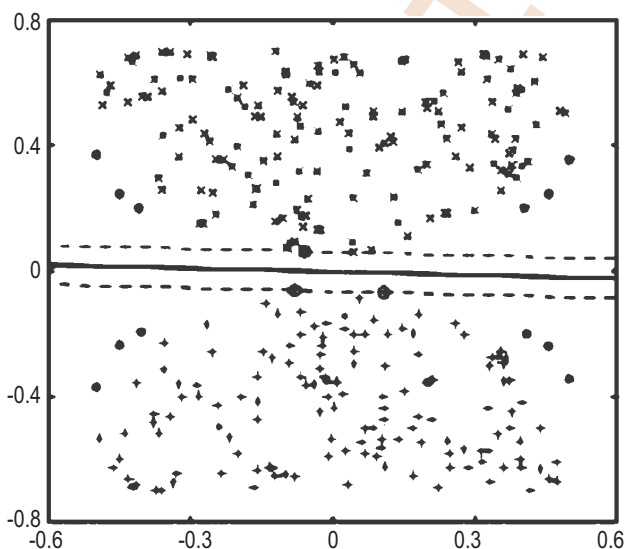
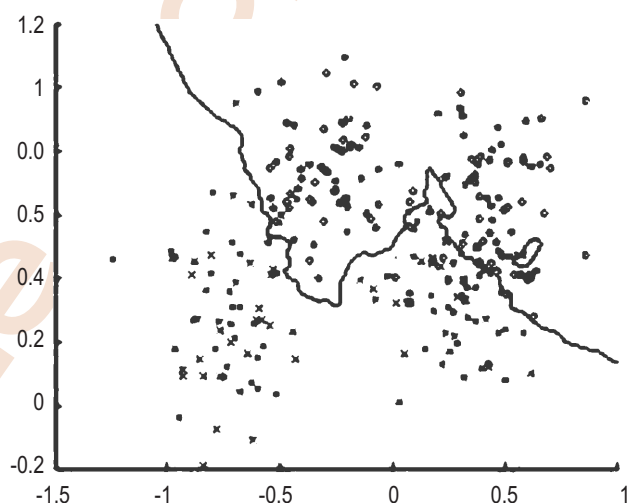
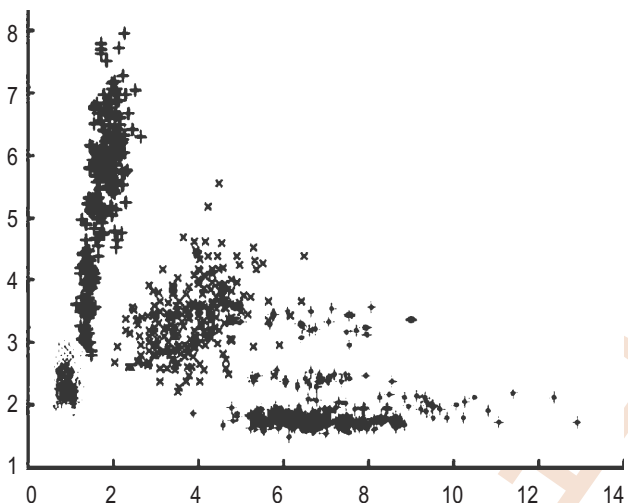
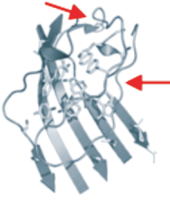
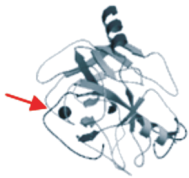
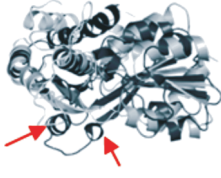

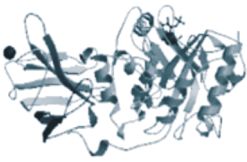
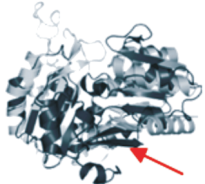


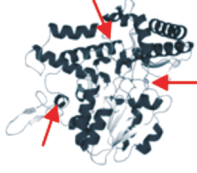
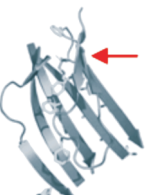
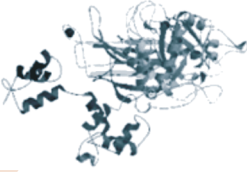
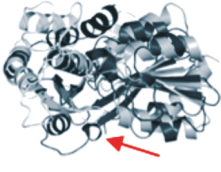

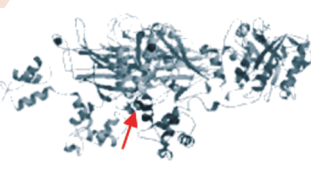
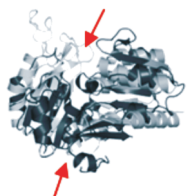


Fig. 8 : Difference in classification process using (i) ANN, (ii) KNN, (iii) SVM and (iv) TWSVM classifier based on sequence from EC.3.2.



**Table 2** : Samples of sequence structure from EC. 3.2 in different sequence similarities represented by three input features

Input features	AAC	DPC	APH
Sequence similarity			
10%			
20%			
25%			
30%			
40%			

### Results and Discussion

The performance and stability of Bio-TWSVM was tested using five different sequence similarities versions of datasets extracted from ENZYME and Uniprot/Swiss-Prot database, which were 10%, 20% 25% 30% and 40%. As a comparison, the results produced, based on accuracy, sensitivity and specificity was compared with other existing classifiers such as the conventional SVM (Hu and Wang, 2011; Wang and Hu, 2011), KNN (Huang *et al.*, 2007) and ANN (Naik *et al.*, 2007).

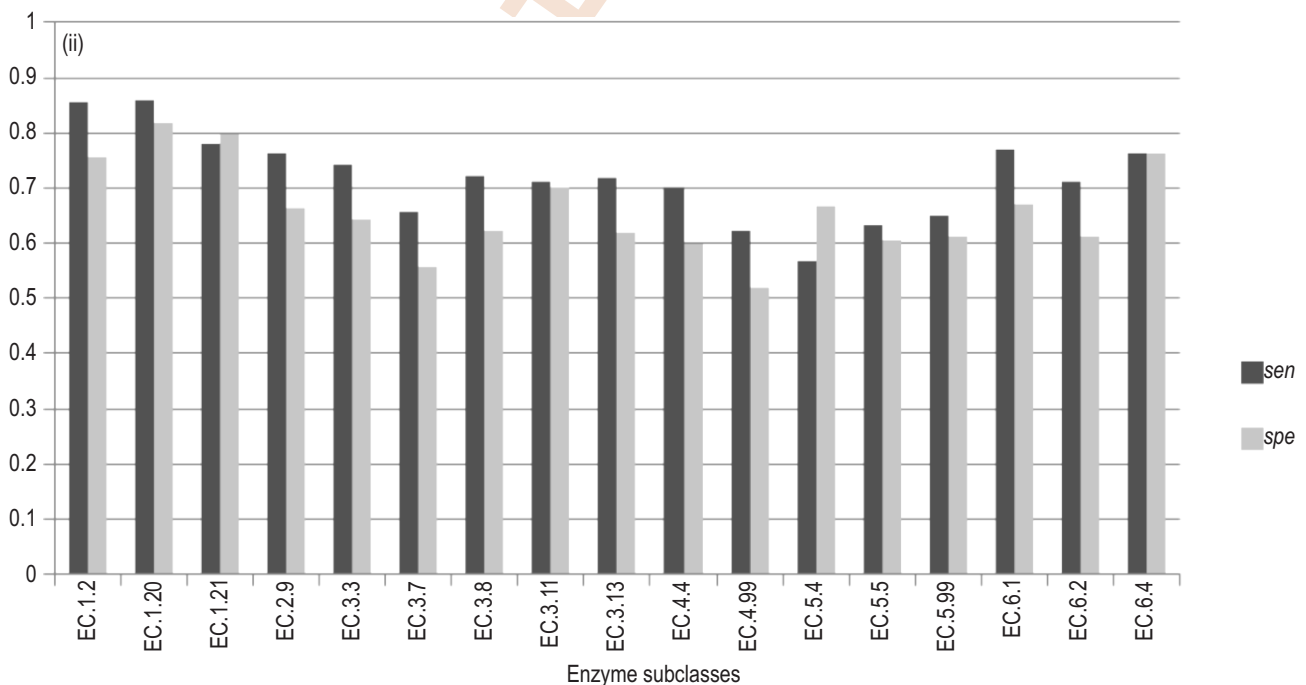
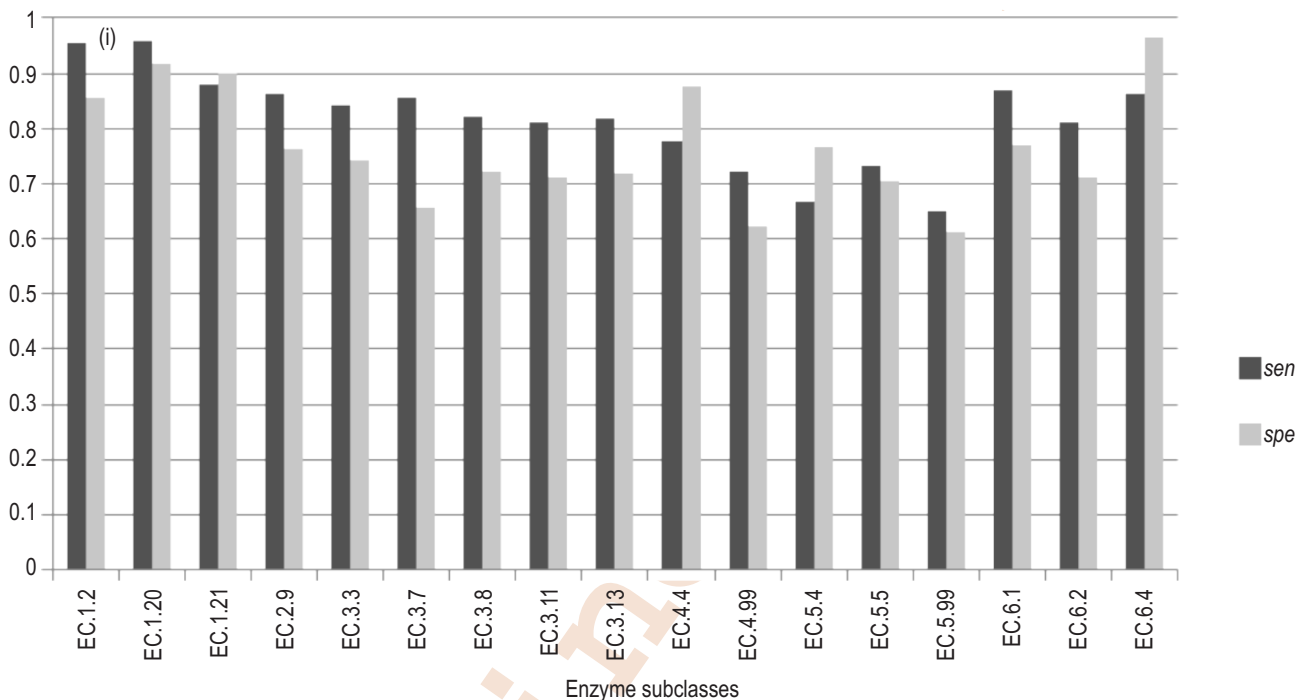
**Assessment on the most significant feature using different rate of sequence similarities:** The most significant feature is assessed using three measurements: *acc* is used to assess the

degree of correctly predicted subclasses with respect to the ground truth; *sen* is used to assess the degree of correctly predicted subclasses with respect to the actual positive label; *spe* is used to assess the degree of correctly predicted subclasses with respect to the actual negative label; where positive is the class under consideration while negative is the complementary counterpart. Moreover, highest *acc*, *sen* as well as *spe* are also highlighted. Fig. 6 presents the prediction performance that has been achieved using the discussed measurements. Meanwhile, Table 2 shows its related active sites.

Based on Fig. 6, it shows APH as the top performer in all functional classes with 25% sequence ID in the range of *acc* from

86.39% to 98.10%. On the other hand, using 10% sequence ID, AAC achieved highest *acc* in all classes with 69.33% as compared to 53.04% and 55.11% using AAC and APH, respectively. Assessment based on *acc* alone may be biased towards the majority class. Thus, *sen* and *spe* is needed in order to describe the degree of true prediction tendency. APH depicts a more stable behaviour where the *sen* and *spe* were 96.12% and 88.70%, respectively, in 25% sequence ID dataset. The dominance of APH is attributable to the scores amalgamation

between AAC, DPC and hydrophobicity and/or hydrophilicity, where one complements the other, owing to the strong inheritance between the composition of amino acid and amphiphilic pseudo amino acid properties of protein. Both were used to assign the enzyme functional and sub-functional class were designed based on the native information on sequence-order information using the correlation factors (Zhou *et al.*, 2007), while DPC was used to define the sequence length information. In Table 2, we have illustrated the existence of active sites where



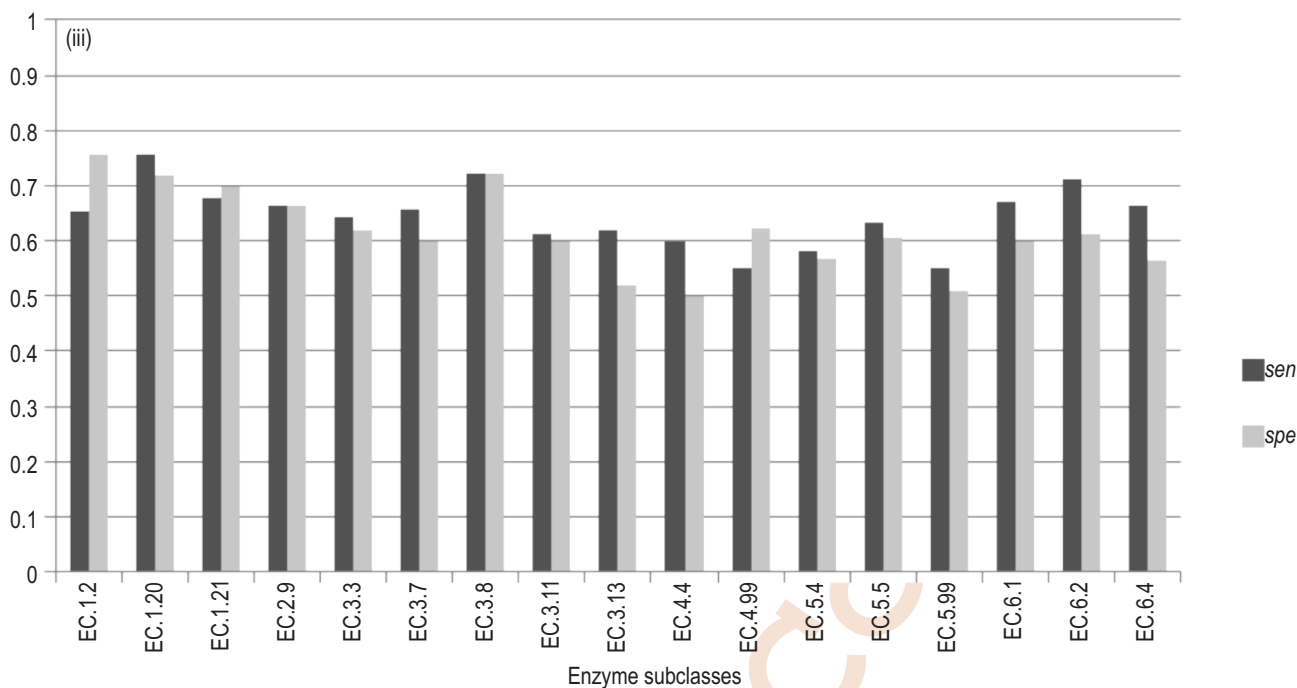


Fig. 9 : Performance comparison in terms of sensitivity and specificity for selected subclasses using (i) Fisher, (ii) Mismatch and (iii) Spectrum kernels.

these sites in an enzyme structure contribute to the alignment scores. Therefore, sequence with highest scores is functionally related. For instance, there were active sites in EC.3.1 using 25% similarity, thus increasing the functionality prediction as well. (i) Rate of sequence similarities using different features by *acc*; (ii) Rate of sequence similarities using different features by *sen*; (iii) Rate of sequence similarities using different features by *spe*.

**Assessment on the effects of classifiers by using different rate of sequence similarities:** The standard SVM (one versus one binary SVM) is introduced for comparison. We plot the distributions of *acc* with respect to the rate of sequence ID for each classifier on the six main functional classes, respectively, in Fig. 7. TWSVM obtains the best accuracies, followed by standard SVM, KNN and ANN becomes the last in identifying all sub-functional classes of all six main families, mainly due to both TWSVM and standard SVM take the imbalance property into account, perform better than KNN and ANN. These results suggest that the more properties of dataset itself are incorporated into the predictive model, the better results can be expected. Fig. 7 shows that TWSVM outperforms the standard SVM in identifying all sub-functional classes of all six main families, i.e., although both TWSVM and standard SVM are especially designed for the imbalance problem, TWSVM seems more reliable.

Khemchandani, Jayadeva and Chandra, (2009), demonstrated that by designing parameter adjusting strategies; TWSVM algorithm with bio-inspired kernel component

demonstrates better generalization performance than standard SVM on imbalanced classification problems, i.e.,  $K$  was adjusted in each TWSVM algorithm iteration. However, to reduce the computational complexity, parameter  $K$  was fixed in each iteration, so the *acc* from TWSVM did not increase dramatically as expected. Fig. 8 shows the output of classifiers used for comparison. The proposed classifier, TWSVM has a better separation plane than the others thus able to handle the imbalance class problem in enzyme sub-functional class prediction. Besides, the neural network-based methods is complex as compared to SVM based methods where the former requires huge amount of data for classification whereas the latter are simple, more efficient and achieves better results.

**Assessment on the effect of bio-inspired kernels on TWSVM:** The overall results produced by the proposed method, Bio-TWSVM, in various versions of sequence similarities datasets were better in terms of *acc* than the previous proposed method. Fig. 9 summarizes the results obtained from the method using three different bio-inspired kernels combined with TWSVM classifier. When class distributions are imbalanced traditional classification algorithms can be biased towards the majority class due to its over-prevalence. Thus, to overcome this, the proposed approaches were categorized to deal with imbalanced class distributions in enzyme sub-functional class prediction, which modify existing algorithms to take the class imbalance into consideration. The basic thought of TWSVM was to construct a hyperplane for each class of samples and making each hyperplane as close as possible to one class of samples, and as

**Table 3** : Examples of classification between previous and this study based on Gene Ontology

Enzyme subclasses	This study	Previous study (Wang et al., 2010)	GO molecular function
<b>Unclassified subclasses</b>			
EC.1.20 [Swiss-Prot:P44589]	EC.1	Unknown	GO:0016491 Oxidoreductase activity
EC.1.21 [Swiss-Prot:Q47878]	EC.1	Unknown	GO:0016491 Oxidoreductase activity
EC.2.9 [Swiss-Prot:A9KW78]	EC.2	Unknown	GO:0030699 Glycine reductase activity GO:0016740 Transferase activity GO:0016785 Transferase activity, transferring selenium-containing groups
EC.3.3 [Swiss-Prot:B3KUA0]	EC.3	Unknown	GO:0016787 Hydrolase activity
EC.3.7 [Swiss-Prot:E5ALX2]	EC.3	Unknown	GO:0016787 Hydrolase activity GO:0047621 Acylpyruvate hydrolase activity
EC.3.8 [Swiss-Prot:Q9UYC9]	EC.3	Unknown	GO:0016787 Hydrolase activity
EC.3.11 [Swiss-Prot:O31156]	EC.3	Unknown	GO:0016787 Hydrolase activity GO:0050194 Phosphonoacetaldehyde hydrolase activity
EC.3.13 [Swiss-Prot:P54997]	EC.3	Unknown	GO:0016787 Hydrolase activity
<b>Correctly classified subclasses</b>			
EC.4.4 [Swiss-Prot:B4E1R2]	EC.4	EC.4	GO:0016829 Lyase activity
EC.4.99 [Swiss-Prot:E5AN32]	EC.4	EC.4	GO:0016829 Lyase activity GO:0016852 Sirohydrochlorin cobaltochelataase activity
EC.5.5 [Swiss-Prot:P04982]	EC.5	EC.5	GO:0016853 Isomerase activity GO:0048029 Monosaccharide binding
EC.5.99 [Swiss-Prot:B4DLV2]	EC.5	EC.5	GO:0003918 DNA topoisomerase (ATP-hydrolyzing) activity GO:0016853 Isomerase activity
EC.6.2 [Swiss-Prot:B3KUV2]	EC.6	EC.6	GO:0016874 Ligase activity
EC.6.4 [Swiss-Prot: P11498]	EC.6	EC.6	GO:0016874 Ligase activity GO:0046872 Metal ion binding
<b>Misclassified subclasses</b>			
EC.1.2 [Swiss-Prot:B4DVF1]	EC.1(EC.1.2)	EC.1(EC.1.3)	GO:0016491 Oxidoreductase activity
EC.5.4 [Swiss-Prot:B3KRB2]	EC.5(EC.5.4)	EC.5(EC.5.1)	GO:0009982 Pseudouridine synthase activity GO:0016853 Isomerase activity
EC.6.1 [Swiss-Prot:B7Z840]	EC.6(EC.6.1)	EC.6(EC.6.3)	GO:0016874 Ligase activity

far as possible from the other class of samples. The new samples would be assigned to one of the classes depending upon its proximity to the two nonparallel hyperplanes, and with the aid of bio-inspired kernels the biological information carried on the sequence level could be transmitted for prediction.

**Validation on the unclassified enzyme subclasses using Gene Ontology:** The result of the predicted enzyme sub-functional class based on its main functional class was analyzed with the Gene Ontology (GO) molecular function annotation Revision: 1.487 of 19–June–2012 (<http://www.geneontology.org/external2go/ec2go>) of the interacting protein sequences involved in the experimental dataset. The sub-functional class contains miscellaneous enzymes and includes several reactions for which the classification may have to be reviewed further by incorporating knowledge-based context of functional groups. Table 3 shows some of the functions that were predicted by the proposed method, thus act as additional functional terms to the existing GO annotated enzyme functional class.

The data in table was partitioned into three sections: (i) Unclassified subclasses, where the existing subclasses in enzyme database was not classified in previous studies, but in this work it was successfully annotated using its GO function; (ii) Correctly classified subclasses, the comparison between previous and current work on accurate subclass prediction was proven using GO function; and (iii) Misclassified subclasses, some of the previously predicted subclasses was found to be incorrect. For instance, from Table 3 the subclass EC.5.4 was predicted as EC.5.1 by Wang et al. (2010) where our work has predicted the subclass as EC.5.4 which corresponds to its GO function.

**Comparison to other related works:** According to Table 4, comparing the prediction performance with existing work is difficult due to different specification of computational framework and sequences. However, it is obvious that previous researchers have used large number of sequences compared to current method followed in this study. To the best of our knowledge, the best  $acc_{all}$  was reported by Shi and Hu (2010) with 91.72%, using

**Table 4:** Performance comparison with other methods in predicting enzyme sub-functional classes

Sequence Similarity/Features Vector/Method	References	acc <sub>all</sub> (%)
<40ED/PseAAC+CTF/AM-SVM	Wang et al. (2010)	91.51
d" 40ED/LFD+ID/SVM	Shi and Hu (2010)	90.72
d" 40ED/Top-Down Approach/KNN	Shen and Chou (2007)	91.40
<40ED/Am-Pse-AAC/AFK-NN	Huang et al. (2007)	90.10
<40ED/Functional Domain+PseAAC/ISort	Cai and Chou (2005)	90.82
d" 25ED/GO-PseAAC	Chou and Cai (2004)	86.53
<40ED/Am-PseAAC/Augmented CDA	Chou (2005)	76.60
<40ED/AAC/CDA	Chou and Elrod (2003)	63.61
<100ED/EMBOSS-PEPSTAT/Random Forest	Kumar and Choudhary (2012)	91.08
d" 40ED/Scoring Function and ID/SVM	Hu and Wang (2011)	89.16
d" 40ED/Auto-correlation function/SVM	Wang and Hu (2011)	88.47
<b>25ED/APH/Bio-TWSVM</b>	<b>This study</b>	<b>92.01</b>

\*ED: Percentage of sequence ID; AM-SVM: Arithmetic mean (AM) offset SVM; LFD+ID: Low-frequency power spectral density and increment of diversity; AFK-NN: Adaptive fuzzy KNN; CDA: Covariant-discriminant algorithm.

low-frequency power spectral density and increment of diversity as features. Notwithstanding of utilizing only 4,732 sequences, the proposed method surpasses others at 92.01%. We believe this was due to the efficiency of Bio-TWSVM to precisely solve the imbalance classification problem and in turn use it to predict the enzyme sub-functional classes. This was directly accredited to the hybridization of different features to form APH, which cancelled out each other weaknesses. Furthermore, most of other works used fixed parameter for classification. Thus, their prediction performances could be aggravated by the non-optimal feature or classifier. The notion was supported by our experimentation in finding the optimal feature and classifier.

In this study, we can conclude that enzyme sub-functional classes are essential protein fold prior to the prediction of protein structure and function as we proposed Bio-TWSVM in order to overcome the weaknesses of imbalance data distribution in subclasses prediction. We devised APH, which is hybridized from different features in predicting low homologous sequence similarities. In a nutshell, our prediction performance is significantly better than other related works. The larger amount of sequences and further exploring the sub-functional classes for more latent information also will be explored.

### Acknowledgments

We would like to thank the Universiti Tun Hussein Onn for supporting this research under the Contract Grant Vot number W004. We very grateful for Universiti Teknologi Malaysia for supporting this research under Fundamental Research Grant Scheme (FRGS) Grant Vot number 4F973 and 5F037 awarded by Ministry of Education Malaysia. The authors also thank the anonymous viewers for the feedback.

### References

Arjunan, S.P., D.K. Kumar and G.R. Naik: A machine learning based method for classification of fractal features of forearm sEMG using

Twin Support vector machines. *International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4821–4824 (2010).

Borro, L.C., S.R.M. Oliveira, M.E.B. Yamagishi, A.L. Mancini, J.G. Jardine, I. Mazoni, E.H. Santos, Higa, R.H. Kuser and Neshich: Predicting enzyme class from protein structure using Bayesian classification. *Genet. Mol. Res.*, **5**,193–202 (2006).

Cai, Y.D. and K.C. Chou: Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. *J. Prot. Res.*, **4**, 967–971 (2005).

Chou, K.C. and D.W. Elrod: Prediction of enzyme family classes. *J. Prot. Res.*, **2**, 183–190 (2003).

Chou, K.C. and Y.D. Cai: Using GO-PseAA predictor to predict enzyme sub-class, *Biochemical and Biophysical Research Communications. Academic Press*, **325**, 506–509 (2004).

Chou, K.C.: Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinform. Oxf. Univ. Press*, **21**, 10–19 (2005).

Daya, B. and K. Pant: Biomonitoring of wetland using macrophytes and macroinvertebrates. *Malays. J. Sustain. Agric.*, **1**, 11-14 (2017).

Ding, S., J. Yu, B. Qi and H. Huang: An overview on twin support vector machines. *Artificial Intelligence Review. Springer Netherlands*, **42**, 245–252 (2014).

Getreuer, P.: Linear Methods for Image Interpolation. *Image Proc. Line*, **1**, 238–259 (2011).

H'ng, S.Y., M.S. Anuar and M.Z. Mohd Nor: Drying, colour and sensory characteristics of 'Berangan' Banana (*Musa accuminata*) flesh dried using a microwave oven. *Malays. J. Halal Res.*, **1**, 10-14 (2018).

Hanedar, A., E. Guneş, G. Kaykiog'lu and E. Cabi: Determination of polychlorinated Biphenils in the soil, atmospheric deposition and bioindicator samples in the Meric-Ergene River Basin. Turkey. *Acta Sci. Malays.*, **1**, 11-13 (2017).

Hasan, M.M.: Bioaugmentation approach in rhizospheric microbiome research: A lesson from arsenic remediation. *Malays. J. Halal Res.*, **1**, 15-16 (2018).

Hu, X. and T. Wang: Prediction of enzyme subclass by using support vector machine based on improved parameters, in 2011 Seventh International Conference on Natural Computation. *IEEE*, 593–598

- (2011).
- Huang, W.L., H.M. Chen, S.F. Hwang and S.Y. Ho: Accurate prediction of enzyme subfamily class using an adaptive fuzzy k-nearest neighbor method. *BioSystems. Elsevier*, **90**, 405–413 (2007).
- Jayadeva, R. Khemchandani and S. Chandra: Twin support vector machines for pattern classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **29**, 905–910 (2007).
- Khemchandani, R. Jayadeva and S. Chandra: Optimal kernel selection in twin support vector machines'. Optimization Letters. *Springer-Verlag*, **3**, 77–88 (2009).
- Kumar, C. and A. Choudhary: A top-down approach to classify enzyme functional classes and sub-classes using random forest. *EURASIP J. Bioinform. Syst. Biol.*, **1**, 1 (2012).
- Kumar, C., G. Li and A. Choudhary: Enzyme function classification using protein sequence features and random forest in 2009 3rd International Conference on Bioinformatics and Biomedical Engineering. *IEEE*, 1–4 (2009).
- Lee, B.J., H.G. Lee, J.Y. Lee and K.H. Ryu: Classification of enzyme function from protein sequence based on feature representation, 2007 IEEE 7th International Symposium on Bio Informatics and Bio Engineering. *IEEE*, 741–747 (2007).
- Leslie, C., E. Eskin and W. Noble: The Spectrum Kernel: A String Kernel for SVM protein clarification. *Pacif. Sympos. Biocomput.*, **7**, 564–575 (2001).
- Leslie, C.S., E. Eskin, A. Cohen, J. Weston and W.S. Noble: Mismatch string kernels for discriminative protein classification. *Bioinformatics. Oxford University Press*, **20**, 467–476 (2004).
- Lotan, I. and F. Schwarzer: Approximation of protein structure for fast similarity measures. *J. Computational Biol.: A J. Comput. Molec. Cell Biol.*, **11**, 299–317 (2004).
- Naik, P.K., V.S. Mishra, M. Gupta and K. Jaiswal: Prediction of enzymes and non-enzymes from protein sequences based on sequence derived features and PSSM matrix using artificial neural network, *Bioinformation*, **2**, 107–12 (2007).
- Ong, S.Q., B.B. Lee, G.P. Tan and S. Maniam: Capacity of black soldier fly and house fly larvae in treating the wasted rice in Malaysia. *Malays. J. Sustain. Agric.*, **1**, 08–10 (2017).
- Scholkopf, B. and A.J. Smola: Learning with kernels: Support vector machines, regularization, optimization and beyond. MIT Press, (2002).
- Shen, H.B. and K.C. Chou: EzyPred: A top-down approach for predicting enzyme functional classes and subclasses. *Biochem. Biophys. Res. Communi.*, Academic Press, **364**, 53–59. (2007).
- Shi, R. and X. Hu: Predicting enzyme subclasses by using support vector machine with composite vectors, protein and peptide. *Letters*, **17**, 599–604 (2010).
- Sonnenburg, S., A. Zien, P. Philips and G. Rätsch: POIMs: Positional oligomer importance matrices - Understanding support vector machine-based signal detectors. *Bioinformatics. Oxford University Press*, **24**, 6–14 (2008).
- Syed, U. and G. Yona: Enzyme function prediction with interpretable models. *Humana Press*, **541**, 373–420 (2009)
- Tian, W. and J. Skolnick: How well is enzyme function conserved as a function of pairwise sequence identity. *J. Molec. Biol.*, **333**, 863–882 (2003).
- Tsochantaridis, I., T. Joachims, T. Hofmann and Y. Altun: Large margin methods for structured and interdependent output variables. *J. Mach. Lear. Res.*, **6**, 1453–1484 (2005).
- Tsuda, K., S. Akaho, M. Kawanabe and K.R. Müller: Asymptotic properties of the Fisher Kernel, Neural Computation. MIT Press 238 Main St., Suite 500, Cambridge, MA 02142-1046 USA, **16**, 115–137 (2004).
- Walder, C., K.I. Kim and B. Schölkopf: Sparse multiscale gaussian process regression. In Proceedings of the 25<sup>th</sup> International Conference on Machine learning - ICML '08, 1112–1119 (2008).
- Wang, P., S. Ownby, Z. Zhang, W. Yuan and S. Li: Cytotoxicity and inhibition of DNA topoisomerase I of polyhydroxylated triterpenoids and triterpenoid glycosides. *Bioor. Medi. Chem. Lett.*, **20**, 2790–2796 (2010).
- Wang, Y. and X. Hu: Predicting of oxidoreductase and lyase subclasses by using support vector machine' in 2011. 10<sup>th</sup> International Conference on Computer and Information Science. *IEEE*, 27–31 (2011).
- Wang, Y.C., X.B. Wang, Z.X. Yang and N.Y. Deng: Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature, protein and peptide letters, **17**, 1441–1449 (2010).
- Webb, E.C.: Enzyme Nomenclature. Recommendations 1984. Supplement 2: corrections and additions. *Europ. J. Biochem.*, **179**, 489–533 (1989).
- Wu, G. and E.Y. Chang: KBA: Kernel boundary alignment considering imbalanced data distribution. *IEEE Transactions on Knowledge and Data Engineering*, **17**, 786–795 (2005).
- Ye, J., G. Coulouris, I. Zaretskaya, I. Cutcutache, S. Rozen and T.L. Madden: Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinform.*, **13**, 134 (2012).
- Yehia, Y.I.: Molecular histological and biochemical effects of tea seed cake on hepatic and renal functions of *Oreochromis niloticus*. *Acta Sci. Malays.*, **1**, 08–10 (2017).
- Zhou, X., C. Chen, Z.C. Li and X.Y. Zou: Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol. Acad. Press*, **248**, 546–551 (2007).