GPCR PROTEIN FEATURE REPRESENTATION USING DISCRETE WAVELET TRANSFORM AND PARTICLE SWARM OPTIMISATION ALGORITHM

Nor Ashikin Mohamad Kamal[1], Azuraliza Abu Bakar[2] and Suhaila Zainudin[2]

[1]School of Computing Sciences, College of Computing, Informatics and Media, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia.
[2]Centre for Artificial Intelligence Technology (CAIT), Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia.

ABSTRACT

Features play an important role in representing classes in the hierarchy structure, and using unsuitable features will affect classification performance. The discrete wavelet transform (DWT) approach provides the ability to create the appropriate features to represent data. DWT can produce global and local features using different wavelet families and decomposition levels. These two parameters are essential to obtain a suitable representation for classes in the hierarchy structure. This study proposes using a particle swarm optimisation (PSO) algorithm to select the suitable wavelet family and decomposition level for G-protein coupled receptor (GPCR) hierarchical class representation. The results indicate that the PSO algorithm mostly selects Biorthogonal wavelets and decomposition level 2 to represent GPCR protein. Concerning the performance, the proposed method achieved an accuracy of 97.9%, 85.9%, and 77.5% at the family, subfamily, and sub-subfamily levels, respectively.

KEYWORDS

Decomposition level, GPCR, hierarchical classification, particle swarm optimisation, wavele tfamily

1. INTRODUCTION

Proteins play an important role in the structure and function of all living cells and viruses. Proteins have several general functions: structural, defence, receptor, signal, transport, movement, hormone, and storage. Receptors are protein molecules receiving chemical signals from outside the cell and allowing certain molecules to enter and leave the cell. G-protein coupled receptors (GPCR) exist on the surface of every cell [1]. GPCRs generate signals in cells to regulate key physiological processes, such as hormone signalling, neurotransmission, cognition, vision, taste, pain perception, and others. It is also known as the seven transmembrane domain (7TM) receptor because there are seven transmembrane segments in which three loops are outside the cell, and three loops are inside the cell, with the N-terminal position being inside the cell and the C- terminal being outside the cell. GPCR consists of three levels: family, subfamily, and sub- subfamily. The family level consists of five classes: families A, B, C, D, and E. The subfamily level consists of 38 classes, and the sub-subfamily level consists of 87 classes. Since GPCR has very complicated relationships between classes, classifying GPCR has proven to be very difficult [2]. In addition, many protein sequences in the same family share homology with protein sequences in other families, increasing classification difficulty [3]. GPCR classification not only depends on sequence order but also includes structural, functional, and evolutionary

characteristics, such as chemical and pharmacological factors [4]. GPCRs are one of the most challenging data sets to classify based on these factors.

The discrete wavelet transform (DWT) method is suitable for representing features of biological data ([5]; [6]; [7]; [8]; [9]). With the nature of the multire solution analysis, DWT can provide information on protein sequences more effectively and allow biological signals to be analysed in the frequency domain and time domain ([9]; [6]). This property is not found in signal processing methods, such as the Fourier transform, which can only study signals in the frequency domain ([10]; [11]; [8]; [12]). Therefore, the advantages of this DWT method can provide more information than other feature representation methods ([13]; [14]; [15]; [16]; [17]). DWT can produce global and local features in various decomposition levels to be analysed and produce features that do not have overlaps [18]. DWT can also decompose global features, such as PseAAC and AAC, into coefficients at different decomposition levels, producing global and local features for a protein sequence ([9]). Global features are obtained from approximation coefficients, while local features are obtained from detailed DWT coefficients [6]. [19] used DWT Coiflet 4 family with decomposition level 3 to obtain global and local features. Although this method has achieved high accuracy for all three levels of the GPCR protein hierarchy, the dataset only includes classification at the family, subfamily A, and sub-subfamily Amine levels.

Selecting the appropriate DWT family type and decomposition level is important in data analysis. It is because an accurate representation will preserve the important data features to be analysed and further help to understand its organisation and complexity ([20]; [21]). The selected DWT family needs to meet the characteristics of orthogonality, symmetry, and shape similarity with the studied data signal ([22]; [23], [25]). Nevertheless, most studies chose the family type and DWT decomposition level selection based on experience or manually ([20]; [24]). Different family types and decomposition levels can be used in DWT. Types of DWT families are Haar, Daubechies, Coiflets, Symlets, Discrete Meyer, and Biorthogonal. The decomposition level is related to the number of global and local features produced from DWT [26]. Using a high decomposition level on a short sequence will produce overlapping information, while using a low decomposition level on a long sequence will ignore much detailed information about the sequence ([14]; [27]). Therefore, selecting the appropriate family type and DWT decomposition level for the feature representation of a class is important because these two parameters affect the classification performance ([28]; [29];[30]; [26]; [31]). Some studies used metaheuristic methods to optimise the selection of family types and DWT decomposition levels, such as genetic algorithms ([32]; [33]), particle swarm optimisation ([34]; [26], [32]), whale optimisation algorithm [35], and evolutionary quantum swarm algorithm [36]. Many researchers used metaheuristic methods to obtain the optimal wavelet family and the decomposition level for research in engineering fields. However, no research has been done to choose the optimal DWT family type and decomposition level for protein feature representation in hierarchical classes.

The organisation of the remaining paper is as follows. The methodology section discusses the proposed methods, the particle swarm optimisation (PSO) algorithm to select the wavelet family and decomposition level to represent each parent class in the G-protein coupled receptors (GPCR) protein hierarchy. This section also explains the hierarchical classification and classifier used in this study. Lastly, the results and discussions section contains the experiment's findings, and conclusions are presented in the last section.

2. METHODOLOGY

Figure 1 shows the flowchart of the overall GPCR protein hierarchical classification using the features generated from the optimal wavelet family and decomposition level. This process begins by representing the GPCR protein sequence using the pseudo amino acid composition algorithm (PseAAC). This algorithm converts each protein sequence consisting of letters into a vector with 170 numerical features. Next, these numerical features must be converted to wavelet coefficients using the suitable wavelet family, *m*, and decomposition level, *n*. There are 82 wavelet families, and seven decomposition levels can be chosen. The PSO algorithm has been used to determine each class's wavelet family and decomposition level at the parent node of the GPCR protein hierarchy.

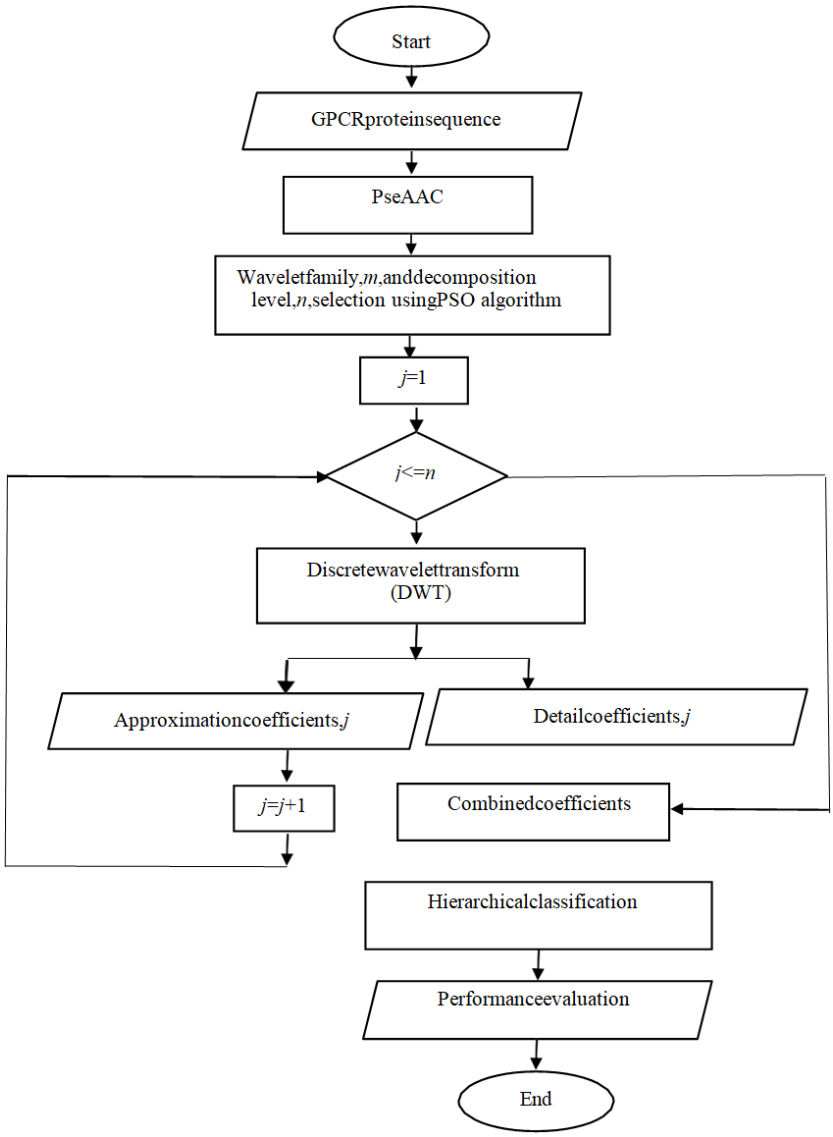

Figure 1. The GPCR protein hierarchical classification flowchart

The DWT process at each decomposition level, *j*, will produce approximation and detail coefficients; they will be used for the transformation process for the next level. The number of coefficients generated depends on the decomposition level, *j*, chosen by the PSO algorithm. This

feature vector is then classified using a top-down hierarchical classification algorithm using the SVM classifier. The following subsections describe each step involved in the GPCR protein hierarchical classification, as shown in Figure 1.

2.1. Data Collection

This study uses the G-Protein Coupled Receptor (GPCR) protein hierarchy benchmark dataset collected by [37], known as GDS. The entire protein sequence has a length measurement of no more than 280 residues. Only classes with a number of data greater than ten at the third hierarchical level are used. It makes a total of five classes at the first level, 36 at the second level, and 87 at the third level of the hierarchy. The total number of protein sequences used in this study is 8,222 protein sequences. The GDS dataset is available at *http://www.cs.kent.ac.uk/projects/biasprofs/down-208loads.html*. This data has been used by past researchers, such as [38], [39], [40], [41], [42], and [43].

2.2. Feature Representation

This study uses two feature representation methods: pseudo-amino acid composition (PseAAC) and discrete wavelet transform (DWT).

2.2.1. Pseudo Amino Acid Composition (PseAAC)

The GPCR protein sequences are first represented using the pseudo amino acid composition (PseAAC) method. This method transforms the protein sequences into numerical features. Here are the steps implemented to transform protein sequences into numerical features using the PseAAC method. In the PseAAC method, protein P can be written as:

$$\text{PseAAC}=P_1,P_2,...,P_{20},...,P\wedge \qquad (1)$$

Where

$$\wedge = 20 + n\lambda \qquad (2)$$

The first twenty elements from P1 to P20 in Equation (1) represent the frequency of the amino acid in the sequence. The symbol λ is the amino acid correlation rank used where λ= 1,…, m. This rank value is a non-negative integer value smaller than the protein sequence's length. The n value represents the number of physiochemical properties of the amino acid used. This study uses six physiochemical properties of the amino acid as in [44], namely the mass properties, the pK group from pK from the α-COOH group, and the pI group at 25∘C. The weighting factor, w, is created to emphasise PseAAC compared to protein features using the amino acid composition. PseAAC protein features are obtained directly through the PseAAC web server ([45]). There are two types of PseAAC: type I and type II. The study of [45] found that PseAAC type II is more suitable to be used as a protein feature. PseAAC type II considers the contribution of physiochemical properties during the calculation ([46]). Therefore, the features generated for each protein sequence from the parameters determined above according to Equation (2) is = 20+6*25, which is 170 features.

The feature normalisation process is next on the PseAAC features, in which each feature is scaled to the limit range [0,1] so that the feature value is in a small range. Normalising features can prevent features with large values from dominating features with small values. In addition, it can also simplify the calculation process. The formula for feature normalisation is in Equation (3):

$$x_N = \frac{x_0 - min_0}{max_v - min_v} \qquad (3)$$

where x_N is the value of the feature after the normalisation process x_0 is the original value of the feature, max_v and min_v are the maximum value and minimum value of the entire features.

2.2.2. Discrete Wavelet Transform (DWT)

The next step is calculating the wavelet transform on the normalised PseAAC features. There are two important functions in DWT: the scaling and wavelet functions. The scaling function is used to get low-frequency features for the normalised PseAAC features, and these low-frequency features are known as the approximation coefficients containing global information. The definition of the scaling functionis $\phi(x)$ as in Equation (4) [47]:

$$\phi_{a,b}(x) \equiv \frac{1}{\sqrt{|a|}} \phi\left(\frac{x-b}{a}\right) \tag{4}$$

where *x*, *a*, and *b* denote the data, scaling, and translation parameters, respectively. The wavelet function $\psi(x)$ is used to obtain detail coefficients or local features. The definition of the wavelet function is in Equation (5) [47]:

$$\psi_{a,b}(x) \equiv \frac{1}{\sqrt{|a|}} \psi\left(\frac{x-b}{a}\right) \tag{5}$$

This study combines approximation and detail coefficients depending on the wavelet family selection and decomposition level selection to form the coefficient vector. The detail coefficients are retained since they may contain valuable information. DWT can be implemented using different wavelet family types. Each wavelet family type has its scaling and wavelet functions [31]. The right family type should be selected so DWT can be used efficiently [18]. In this study, 82 wavelet families are available to represent GPCR proteins. These wavelet families include 20 from the Daubechies family (Db1-Db20), 20 from the Symlet family (Sym1-Sym20), five from the Coiflet family (Coif1-Coif5), six from the Fejer-Korovk in family, 15 from the Biorthogonal family, 15 from the Reverse Biorthogonal family, and one from the Meyer family. Each family differs based on compact support, vanishing moment, symmetry, and orthogonality [48].

DWT can analyse the PseAAC feature vectors at various resolutions, which can be done by decomposing the PseAAC feature vector into several decomposition levels. The maximum level that can be implemented is based on the formula log2 (*N*), where *N* is the size of the PseAAC feature. Since the feature size of PseAAC is 170, the maximum possible decomposition is seven levels.

2.2.3. Particle Swarm Optimisation (PSO) Algorithm

Particle swarm optimisation (PSO) is an intelligence-oriented, stochastic computing technique introduced by Kennedy and Eberhart in 1995 [49]. PSO has been widely used for many optimisation problems because of its unique search mechanism and simple implementation. It was inspired by the social behaviour of birds looking for food. In this algorithm, the birds in the flock are a solution in a high-dimensional space with four vectors: the current position, the best position obtained so far, the best position of the other particles, and the velocity. According to [50, 51], the PSO algorithm does not require many calculations and fast convergence. Two important parameters in PSO are *pbest* and *gbest*. The *pbest* value is the best accuracy value for each particle up to iteration, *t*, while the *gbest* value is the best global accuracy value. The fitness function for each particle is calculated using Equation (6) where *c(i)* is the number of protein sequences correctly classified in class i and *Total (i)* is the total number of protein sequences in class i. Each particle updates its position and velocity in each iteration based on Equations (7) and (8):

$$Fitness\ function = \frac{c(i)}{Total\ (i)} \quad (6)$$

$$x_{k+1}^{i} = x_{k}^{i} + v_{k+1}^{i} \quad (7)$$

$$v_{k+1}^{i} = v_{k}^{i} + c_1 r_1 \left(p_{k}^{i} - x_{k}^{i}\right) + c_2 r_2 \left(p_{k}^{g} - x_{k}^{i}\right) \quad (8)$$

where x_k^i is the particle position, v_k^i is the particle velocity, p_k^i is the best-remembered position, and c_1 c_2 are cognitive and social parameters, r_1 and r_2 are random numbers between 0 and 1.

Figure 2 shows the PSO algorithm for calculating each parent class's approximation and detail coefficients in the GPCR protein hierarchy.

Step 1: Assign two random positions value to each particle. The first position resembles the wavelet family, and the second position is the decomposition level.
Step 2: Generate the wavelet approximation and detail coefficients based on the selected wavelet family and decomposition level.
Step 3: Evaluate the classification accuracy for each particle using the fitness function in Equation (6).
Step 4: Compare the classification accuracy for each particle to the accuracy from the *pbest* position. If the particle's classification accuracy value is better than the *pbest* value, update the *pbest* value using the latest classification accuracy value.
Step 5: Identify the particle with the highest classification accuracy value and assign as *gbest*.
Step 6: Update the position and velocity of each particle using Equations (7) and (8).
Step 7: Repeat Steps 2 to 5 until the number of iterations does not exceed the maximum iteration or has found a high accuracy value.

Figure 2. Pseudo code of the Particle Swarm Optimisation Algorithm

The parameter values for the PSO algorithm set in this study are presented in Table 1. The parameters c_1 and c_2 are based on the parameters set in [50].

Table 1. The parameter list for the PSO algorithm

Parameter	Value
Particle size	30
Number of iteration	100
c_1 c_2	1.49445

The coefficients or features generated from the selected wavelet family and decomposition level that produced the highest accuracy for each class are selected as features representing the class in the hierarchy. The next step is to classify the features in the hierarchical structure.

2.3. Hierarchical Classification

In this study, the type of hierarchy used is a tree, and the type of hierarchical classification used is Local Classification at the Parent Node (LCPN). Figure 3 shows an example of a tree hierarchy consisting of three levels. This hierarchy has five parent nodes: the root node, nodes 1, 2, 2.1, and 2.2. A leaf or terminal node consists of nodes 1.1, 1.2, 2.1.1, 2.1.2, 2.2.1, and 2.2.2. The classification model is trained using the exclusive sibling policy, as described in [52]. For example, in Figure 3, the shaded box indicates node 2.1. The positive data set for node 2.1 is node 2.1, while the negative data set for node 2.1 is node 2.2.

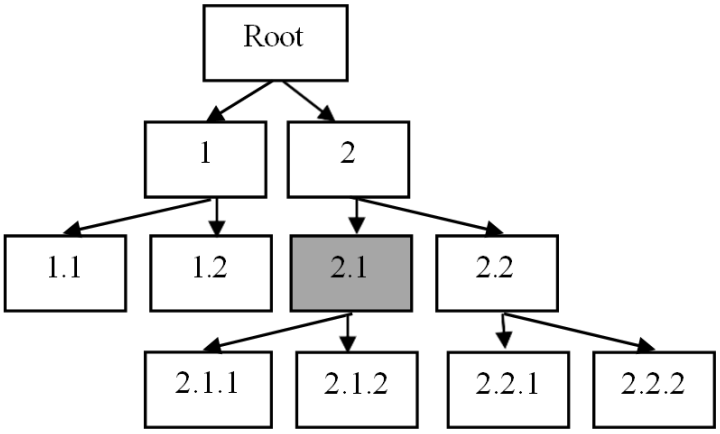

Figure 3. Hierarchical classification using Local Classification on the Parent Node (LCPN)

Local classification methods are divided into three types: local classification at each node (LCN), local classification at each hierarchical level (LCL), and local classification at each parent node (LCPN). Different methods are used to train hierarchical classification depending on the type of local classification. In this study, the support vector machine (SVM) classifier is located on each parent node except the leaf node. The SVM classifier performs well using the PseAAC feature representation method [53]. Therefore, in this study, the classifier used in each parent node in the hierarchy is the SVM classifier. Each SVM model's training is done using a positive training dataset from the parent node and a negative dataset from the parent's sibling nodes.

The top-down hierarchical classification method is used to classify the test data for the classification phase. Referring to Figure 3, the classifier in the root node will classify the data to the node in the first level, which is class 1 or 2 only. This data will then be sent to the second level according to the classification results from the root node. For example, if the SVM classifier classifies new data into class 2, this data will be classified by the classifier into class 2.1 or 2.2. If the data is classified as node 2.1, the next classification is to the leaf nodes, which are nodes 2.1.1 or 2.1.2. The SVM classifier is used to predict GPCR since it is a robust classifier in multiple areas of biological analysis [44]. This study uses the binary SVM classifier and the ECOC model available in MATLAB R2019 a software. The coding method used in this study is one-to-one classification (OVO), while the type of kernel used is Radial Basis Function (RBF).

2.4. Performance Evaluation

The five-fold cross-validation method is used for the evaluation of the proposed algorithm. Data is divided into five subsets that have almost the same size. After training the SVM classifier using four subsets, the classifier's performance is tested using the fifth subset. This process is repeated five times so that each subset can be used as test data.

The effectiveness of the proposed techniques is tested based on four performance testing criteria: accuracy (A), precision (P), recall (R), and F-score (F). These measurement values are used to evaluate the output classification performance resulting from the developed method. Accuracy, precision, recall, and F-score values are defined in Equations (9), (10), (11), and Equation (12).

Equation (9) is used to obtain the accuracy value for each class in the hierarchy, where *c(i)* is the number of protein sequences correctly classified in class *i* and *Total (i)* is the total number of protein sequences in class *i*. In Equations (10), (11), and (12), the values of true positive(TP), false negative (FN), true negative (TN), and false positive (FP) have been used. A true positive (TP) is a situation when a positive case is successfully classified as a positive class, and a false negative (FN) is a positive case classified as a negative case. A true negative (TN) is a negative case successfully classified as negative, and a false positive (FP) is a case that is negative but classified as positive.

$$\text{Accuracy (A)} = \frac{c(i)}{Total\ (i)} \tag{9}$$

$$\text{Precision (P)} = \frac{TP}{TP+FP} \tag{10}$$

$$\text{Recall (R)} = \frac{TP}{TP+FN} \tag{11}$$

$$\text{F} - \text{score (F)} = 2 \text{ x } \frac{P\,X\,R}{P+R} \tag{12}$$

3. RESULTS AND DISCUSSIONS

This paper used the optimised DWT features and SVM classifier in the GPCR protein hierarchical classification model. The results were analysed using three hierarchy levels: family, subfamily, and sub-subfamily. The family, subfamily, and sub-subfamily levels consist of 5, 38, and 87 classes, respectively.

Figure 4(a) shows the PseAAC features vector graph for one of the class A GPCR proteins. The DWT transformed the PseAAC features into approximation and detail coefficients using the Coiflet 4 wavelet family and three decomposition levels as an example. Graphs 4(b), 4(c), and 4(d) show the detailed wavelet coefficients for decomposition levels 1, 2, and 3 containing local features of the protein. In contrast, graph (d) shows the wavelet approximation coefficient with the protein's global features or rough characteristics.

For further discussion, the results are divided into two parts. The first part discusses the selected wavelet family and decomposition level for each parent node in the GPCR protein hierarchy, and the second part discusses the proposed algorithm performances.

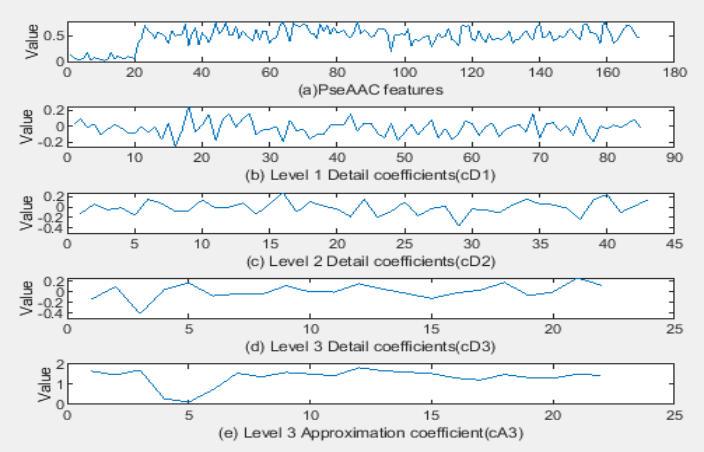

Figure 4. Representation of PseAAC features for class A GPCR proteins using the Coiflet4 wavelet family with three decomposition levels

3.1. Selection of Wavelet Family and Decomposition Level

Table 2 shows the types of wavelet families selected by the PSO algorithm for five folds of the dataset. Each data set fold contained 11 wavelet families to represent 11 parent node classes in the GPCR protein hierarchy, which are the root, family A, family B, family C, subfamily A Amine, subfamily A Hormone, subfamily A Nucleotide, subfamily A Peptide, subfamily A Thyro, subfamily Prostanoid, and subfamily C CalcSense. Therefore, the total amount wavelet families for the five folds of the data set are 55 wavelet families. The wavelet family and decomposition level selected bythe PSO method at eleven hierarchy nodes are analysed. It is clear from Table 2 that the wavelet family and decomposition level selected at different parent nodes differ considerably. It corroborates the use of PSO, which automatically determines each class node's wavelet family and decomposition level.

Table 2. Selected wavelet family and decomposition level at each GPCR node

Class node	Wavelet family and decomposition level
Root	(RBior 3.7,1), (Bior4.4,3), (Bior5.5,3), (Bior2.2,3), (Bior3.7,1)
Family A	(RBior4.4,2), (FK14,3), (Db1,1), (Db8,2), (Db3,1)
Family B	(RBior4.4,2), (RBior3.3,4), (Db3,1), (Bior2.8,2), (Sym5,2)
Family C	(Db8,2), (Db1,1), (Bior5.5,2), (FK4,2), (Rbio1.4,2)
Subfamily A A mine	(Bior3.7,1), (Sym10,3), (Db1,1), (Bior1.5,4), (RBio3.3,2)
Subfamily A Hormone	(Bior2.8,1), (FK4,5), (Bior4.4,5), (Bior2.2,7), (RBio2.2,6)
Subfamily A Nucleotide	(Sym10,7), (Sym5,6), (Sym8,1), (Sym10,2), (Db8,3)
Subfamily A Peptide	(RBio6.8,1), (Bior2.2,1), (Db1,4), (Rbio1.1,3), (Db8,1)
Subfamily A Thyro	(Bior3.9,6), (Bior3.1,2), (RBio3.7,5), (Sym9,6), (Bior5.5,5)
Subfamily A Prostanoid	(Bior1.5,3), (Bior4.4,2), (RBio1.5,2), (Sym8,6), (RBio2.2,6)
Subfamily C CalcSense	(Bior3.3,5), (Bior2.2,4), (RBio1.5,2), (Sym5,2), (Rbio1.3,7)

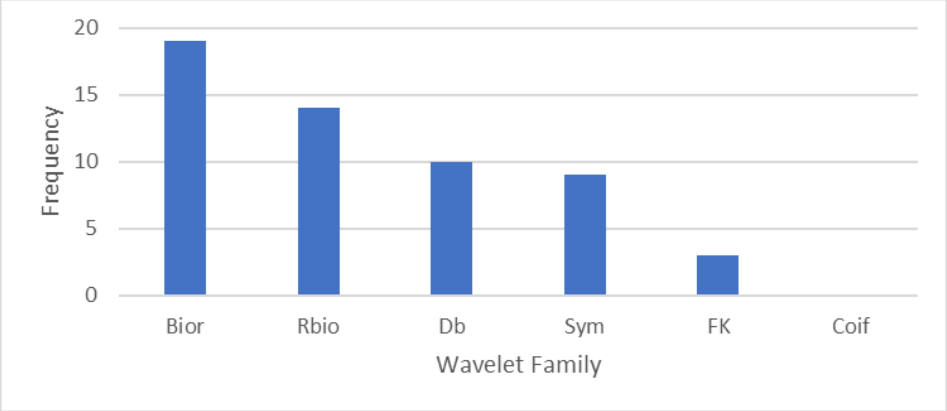

Figure 5. Selection frequency of each wavelet family at each node of the GPCR hierarchy

Figure 5 shows the selection frequency of the wavelet family at each parent node of the GPCR hierarchy. Out of 55 wavelet families, 19 wavelet families are selected from the Biorthogonal (Bior) family, 14 from the family Reverse Biorthogonal (Rbio), ten from the Daubechies (Db) family, nine from the Symlet(Sym) family, and three from the Fejer-Korovkin (FK) family. The Biorthogonal wavelet family is the choice for representing most protein classes because most scaling and wavelet functions in this family have a sudden shape change. It is consistent with PseAAC features as in Figure 4(a), for example, which have a rough shape and many sharp variations. Additionally, according to [54], the Biorthogonal wavelet can eliminate redundant information from protein sequences, minimise feature information leakage and aliasing, allow feature vectors to represent the original sequence information, and enhance prediction performance. In addition, PSO selects Daubechies and Fejer-Korovkin wavelets at least nine times. In contrast, the Coiflet wavelet family is not an option to express the PseAAC feature because the scaling and wavelet functions did not resemble the PseAAC feature in this study.

Figure 6 shows the frequency of selected decomposition levels to represent protein classes. Decomposition level 1 produced half of the 170 PseAAC features as 85 approximation and 85 detail coefficients. Therefore, decomposition level 1 allowed the preservation of global and local information. Decomposition level 2 produced half of the 85 PseAAC features as 42 approximation and 42 detail coefficients. This process continues for the rest of the decomposition levels. Out of 55 wavelet families, 16 used decomposition level 2, the highest decomposition level used in feature representation. It is followed by 13 wavelet families using decomposition level 1 and eight wavelet families using decomposition level 3. Four wavelet families used decomposition level 4, five families chose decomposition level 5, and six families chose decomposition level 6. The least decomposition level selected for the feature representation is decomposition level 7, which is 3. The greater decomposition level will increase the number of detail coefficients and decrease the number of approximation coefficients. The protein class requires more local information than global information to represent it.

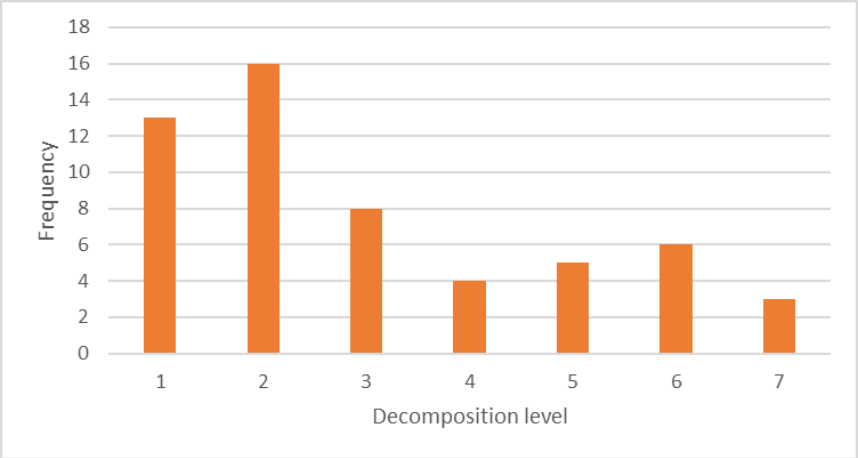

Figure 6. Selection frequency of each decomposition level at each node of the GPCR hierarchy

3.2. GPCR Classification Performance Without Virtual Class

Table 3 shows the performance for classifying GPCR using SVM at the family level. It can be observed from the results show that, at the family level, DWT+PSO is the best feature extraction strategy. The performance achieved by DWT+PSO is 97.9% for accuracy, precision, and recall, and the F-score achieved is 0.979.

Table 4 shows the performance for classifying GPCR using SVM at the subfamily level. As shown in the table, DWT+PSO is the best feature extraction strategy at the subfamily level at this hierarchy level. The performance achieved by DWT+PSO is 85.9%, 88.6.7%, 88.4%, and 0.885% for accuracy, precision, recall, and F-score, respectively. Nevertheless, the DWT+PSO algorithm achieved almost the same accuracy as PseAAC, which is 85%.

Table 3. GPCR classification performance for family level

Feature extraction method	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F-score**
PseAAC	97.7	97.7	97.7	0.977
DWT+PSO	97.9	97.9	97.9	0.979

Table 5 shows the performance for classifying GPCR using SVM at the sub-subfamily level. It can be seen from this table that, at the sub-subfamily level, the best accuracy comes from DWT+PSO, which is 77.5%. The best precision, recall, and F-score come from DWT+PSO, 87.7%, 82.3%, and 0.849, respectively.

Tables 4 and 5 show that the precision value is higher than the accuracy value for the classification results using PseAAC and PSO at the subfamily and sub-subfamily levels. However, the accuracy value is not a good metric for imbalanced data sets, and it is because high accuracy can also be achieved by successfully classifying the dominant negative class. Since it has been confirmed that the GPCR protein has imbalanced data for each class **(**[55]**,** [40]**),** getting a high precision value is more important ([56])**.** It is the same case in Table 5, whereby the recall value is higher than the accuracy value. It also denotes that DWT+PSO performed well in an imbalanced GPCR protein dataset.

Table 4. GPCR classification performance for sub family level

Feature extraction method	Accuracy (%)	Precision (%)	Recall (%)	F-score
PseAAC	85.3	87.7	87.7	0.877
DWT+PSO	85.9	88.6	88.4	0.885

Table 5. GPCR classification performance for sub-sub family level

Feature extraction method	Accuracy (%)	Precision (%)	Recall (%)	F-score
PseAAC	76.9	86.2	77.4	0.816
DWT+PSO	77.5	87.7	82.3	0.849

Table 6 compares GPCR protein classification accuracy performance between previous studies using the GDS data set and this study. It was found that this study produced an accuracy value of 97.9% at the family level, which is almost the same as the accuracy value in the study of [42], [40], [41], [57], and [39]. This study obtained an accuracy value of 86% for the second level of the hierarchy. However, [40] has obtained higher accuracy values for the second level of the hierarchy, which is 89.2%. For the third level, the accuracy value obtained from this study is 78.3%. The highest accuracy value for the third level was obtained from the study of [40], with an accuracy value of 90.95%.

Table 6. Comparison of GPCR protein classification using the GDS dataset

Authors	Super Family	Family	Subfamily	Sub-subfamily
[37]	-	90.59%	73.77%	58.08%
[43]	-	96.97%	82.72%	70.46%
[42]	99.75%	97.38%	81.91%	73.34%
[41]	-	97.41%	84.97%	75.60%
[40]	98%	97.5%	**89.2%**	**90.95%**
[57]	-	97.40%	87.78%	81.13%
[39]	-	97.17%	86.82%	81.17%
This study	-	**97.9%**	86%	78.3%

Although the accuracy of this study is lower than [40] at the subfamily and sub-subfamily level, their research focused on optimising the GPCR protein classifiers. In contrast, this study tried to optimise the representation of GPCR proteins at the feature level. Hence, compared to other studies, it was found that this study's results are comparable to [57] and [39] at the family and subfamily levels, which used the current trending deep learning method.

4. CONCLUSIONS

In this study, the particle swarm optimisation (PSO) algorithm has helped to identify the wavelet family and the appropriate decomposition level for each parent class representation in the GPCR

protein hierarchy. There is still much room for improvement and development in this study. The imbalance data problem in the data set is always encountered in data mining. The GPCR dataset also suffers from the same problem in which 60% of the known GPCR protein sequences are proteins from the A family. The sampling process must be considered to overcome this problem, such as using the Synthetic Minority Over sampling Technique (SMOTE) algorithm. SMOTE is an oversampling technique generating synthetic samples from minority classes, and it is used to obtain a synthetically balanced training set for each class and then train the classifier.

Future research is welcome to study the appropriate wavelet family and decomposition levels for GPCR protein class representation to perform better. The experiments conducted in this study utilised parameter values for the PSO algorithm based on previous studies. This study did not examine whether the parameters are appropriate values for PSO. Therefore, there is a need for a comprehensive study to determine the appropriate parameter values required for the PSO algorithm as an algorithm for selecting the wavelet family and the decomposition level.

There are several approaches to hierarchical classification that can be implemented. It includes the global hierarchical classification method that does not suffer the propagated errors problem. This method generates a complex classification model from a training set for global classification andconsiders the class hierarchy as a whole during classification. Each test data will be classified based on the classification model produced for the test phase. It is hoped that this global hierarchical classification method can help improve the hierarchical classification performance of GPCR proteins.

ACKNOWLEDGEMENT

This study was financially supported by the Malaysia Ministry of Higher Education (MOHE).

REFERENCES

[1] K. Alhosaini, A. Azhar, A. Alonazi, and F. Al-Zoghaibi, "GPCRs: The most promiscuous druggable receptor of the mankind," *Saudi Pharm. J.*, no. May, 2021, doi: 10.1016/j.jsps.2021.04.015.

[2] M. Li, C. Ling, and J. Gao, "An Efficient CNN-based Classification on G-protein Coupled Receptors Using TF-IDF and N-gram," *2017 IEEE Symp. Comput. Commun.*, pp. 924–931, 2017.

[3] M. Davies, A. Secker, and A. Freitas, "Optimising amino acid groupings for GPCR classification, "*Bioinformatics*, vol. 24, no. 18, pp. 1980–1986, 2008, doi: 10.1093/bioinformatics/btn382.

[4] R. Karchin, K. Karplus, and D. Haussler, "Classifying G-protein coupled receptors with support vector machines,"*Bioinformatics*, vol. 18, no. 1, pp. 147–159, 2002, doi: 10.1093/bioinformatics/18.1.147.

[5] S. Saini and L. Dewan, "Comparison of Numerical Representations of Genomic Sequences: Choosing the Best Mapping for Wavelet Analysis, "*Int. J. Appl. Comput. Math.*, vol.3, no.4, pp. 2943–2958, 2017, doi: 10.1007/s40819-016-0277-1.

[6] T. T. Gayathri and S. A. Christe, "Wavelet Analysis in Prediction and Identification of Cancerous Genes, "*Int. J. Sci. Eng. Res.*, vol. 8, no. 3, pp. 720–727, 2017.

[7] W. Hou, Q.Pan, Q. Peng, and M. He, "A new method to analyse protein sequence similarity using Dynamic Time Warping, "*Genomics J.*, vol. 109, no. 2, pp. 123–130, 2017.

[8] T. Meng*etal.*, "Wavelet analysis in current cancer genome research: A survey, "*IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 10, no. 6, pp. 1442–1459, 2013, doi: 10.1109/TCBB.2013.134.

[9] J.-D. Qiu, X.-Y. Sun, J.-H. Huang, and R.-P. Liang, "Prediction of the types of membrane proteins based on discrete wavelet transform and support vector machines., "*Protein J.*, vol. 29, no. 2, pp. 114–9, 2010, doi: 10.1007/s10930-010-9230-z.

[10] B. Chen, Y. Li, and N. Zeng, "Centralized Wavelet Multiresolution for Exact Translation Invariant Processing of ECG Signals, "*IEEE Access*, vol. 7, pp. 42322–42330, 2019, doi: 10.1109/ACCESS.2019.2907249.

[11] A. Elbir, H. O. Ilhan, G. Serbes, and N. Aydin, "Short Time Fourier Transform based music genre

classification,"*2018 Electr. Electron. Comput. Sci. Biomed. Eng. Meet. EBBT2018*, no. June, pp. 1– 4, 2018, doi: 10.1109/EBBT.2018.8391437.

[12] C. C. Aggarwal, "On effective classification of strings with wavelets, " *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 163–172, 2002, doi: 10.1145/775069.775071.*

[13] T. D. Mai, T .D. Ngo, D. D. Le, D .A. Duong, K. Hoang, and S. Satoh, "Using node relationships for hierarchical classification, "*Proc. - Int. Conf. Image Process. ICIP*, vol. 2016-Augus, pp. 514–518, 2016, doi: 10.1109/ICIP.2016.7532410.

[14] B. Yu *et al.*, "Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising,"*J. Mol. Graph. Model.*, vol. 76, no. July, pp. 260– 273, 2017, doi: 10.1016/j.jmgm.2017.07.012.

[15] C. de Trad, Q. Fang, and I. Cosic, "An overview of protein sequence comparisons using wavelets, "*Proc. IEEE-EMBS*, 2001, Accessed: Mar. 29, 2014. [Online]. Available: http://www.eng.monash.edu/non-cms/ecse/ieee/ieeebio2001/trad.pdf.

[16] P. Liò, "Wavelets in bioinformatics and computational biology: State of art and perspectives," *Bioinformatics*, vol. 19, no.1, pp. 2–9, 2003, doi:10.1093/bioinformatics/19.1.2.

[17] A. D. Haimovich, B. Byrne, R. Ramaswamy, and W. J. Welsh, "Wavelet analysis of DNA walks, "*J. Comput. Biol.*, vol. 13, no. 7, pp. 1289–1298, 2006, doi: 10.1089/cmb.2006.13.1289.

[18] Z. Germán-Salló and G. Strnad, "Signal processing methods in fault detection in manufacturing systems, "*Procedia Manuf.*, vol. 22, pp. 613–620, 2018, doi: 10.1016/j.promfg.2018.03.089.

[19] J.-D. Qiu, J.-H. Huang, R.-P. Liang, and X.-Q. Lu, "Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform., "*Anal. Biochem.*, vol. 390, no. 1, pp. 68–73, Jul. 2009, doi: 10.1016/j.ab.2009.04.009.

[20] Y. I. Jang, J. Y. Sim, J. R. Yang, and N. K. K won, "The optimal selection of mother wavelet function and decomposition level for denoising of dcg signal, "*Sensors*, vol. 21, no. 5, pp. 1–17, 2021, doi: 10.3390/s21051851.

[21] S. Saini and L. Dewan, "Performance comparison of first generation and second generation wavelets in the perspective of genomic sequence analysis, "*Int. J. Pure Appl. Math.*, vol. 118, no. 16, pp. 417–442, 2018.

[22] H. He, Y. Tan, and Y. Wang, "Optimal base wavelet selection for ECG noise reduction using a comprehensive entropy criterion, "*Entropy*, vol. 17, no. 9, pp. 6093–6109, 2015, doi: 10.3390/e17096093.

[23] W. K. Ngui, M. S. Leong, L. M. Hee, and A. M. Abdelrhman, "Wavelet analysis: Mother wavelet selection methods, "*Appl. Mech. Mater.*, vol. 393, no. January 2014, pp. 953–958, 2013, doi: 10.4028/www.scientific.net/AMM.393.953.

[24] T. Wang, L. Li, Y. A. Huang, H. Zhang, Y. Ma, and X. Zhou, "Prediction of protein-protein interactions from amino acid sequences based on continuous and discrete wavelet transform features, "*Molecules*, vol. 23, no. 4, 2018, doi: 10.3390/molecules23040823.

[25] D. Chen, S. Wan, J. Xiang, and F. S. Bao, "A high-performance seizure detection algorithm based on Discrete Wavelet Transform (DWT) and EEG, "*PLoS One*, vol. 12, no. 3, Mar. 2017, doi: 10.1371/journal.pone.0173138.

[26] C. Guarnizo, a a Orozco, and M. a Alvarez, "Optimal sampling frequency in wavelet-based signal feature extraction using particle swarm optimisation., "*Conf. Proc. IEEE Eng. Med. Biol .Soc.*, vol. 2013, pp. 993–6, Jan. 2013, doi: 10.1109/EMBC.2013.6609670.

[27] J.-D. Qiu, J.-H. Huang, S.-P. Shi, and R.-P. Liang, "Using the Concept of Chous Pseudo Amino Acid Composition to Predict Enzyme Family Classes: An Approach with Support Vector Machine Based on Discrete Wavelet Transform, " *Protein Pept. Lett.*, vol. 17, no. 6, pp. 715–722, 2010, doi: 10.2174/092986610791190372.

[28] F. M. Albkosh, M. S. Hitam, W. N. J. H. Wan Yussof, A. A. K. Abdul Hamid, and R. Ali, "Optimisation of discrete wavelet transform features using artificial bee colony algorithm for texture image classification, "*Int. J. Electr. Comput. Eng.*, vol. 9, no. 6, pp. 5253–5262, 2019, doi: 10.11591/ijece.v9i6.pp5253-5262.

[29] C. Caramia, C. De Marchis, and M. Schmid, "Optimising the scale of a wavelet-based method for the detection of gait events from a waist-mounted accelerometer under different walking speeds, "*Sensors (Switzerland)*, vol. 19, no. 8, 2019, doi: 10.3390/s19081869.

[30] Z. Zhang, Q. K. Telesford, C. Giusti, K. O. Lim, and D. S. Bassett, "Choosing wavelet methods, filters, and lengths for functional brain network construction, " *PLoS One*, vol. 11, no. 6, pp. 1–24, 2016, doi: 10.1371/journal.pone.0157243.

[31] N. Ahuja, L. Lertrattanapanich, and N. K. Bose, "Properties determining choice of mother wavelet, "*IEE proceedings. Vision, image signal Process.*, vol. 152, no. 5, pp. 205–212, 2005, doi: 10.1049/ip-vis.

[32] Z. A. A. Alyasseri, A. T. Khader, M. A. Al-Betar, A. K. Abasi, and S. N. Makhadmeh, "EEG Signals Denoising Using Optimal Wavelet Transform Hybridized with Efficient Metaheuristic Methods, "*IEEE Access*, vol. 8, pp. 10584–10605, 2020, doi: 10.1109/ACCESS.2019.2962658.

[33] G. Oltean and L. N. Ivanciu, "Computational intelligence and wavelet transform based metamodel for efficient generation of not-yet simulated waveforms, " *PLoS One*, vol. 11, no. 1, pp. 1–30, 2016, doi: 10.1371/journal.pone.0146602.

[34] H. Tao, J. M. Zain, M. M. Ahmed, A. N. Abdalla, and W. Jing, "A wavelet-based particle swarm optimisation algorithm for digital image watermarking, "*Integr. Comput. Aided. Eng.*, vol.19, no.1, pp. 81–91, 2012, doi:10.3233/ICA-2012-0392.

[35] H. Aprillia, H. T. Yang, and C. M. Huang, "Optimal decomposition and reconstruction of discrete wavelet transformation for short-term load forecasting, "*Energies*, vol. 12, no. 24, 2019, doi: 10.3390/en12244654.

[36] A. Semnani, L. Wang, M. Ostadhassan, M. Nabi-Bidhendi, and B. N. Araabi, "Time-frequency decomposition of seismic signals via quantum swarm evolutionary matching pursuit, "*Geophys. Prospect.*, vol. 67, no. 7, pp. 1701–1719, 2019, doi: 10.1111/1365-2478.12767.

[37] M. N. Davies, A. Secker, A. a Freitas, M. Mendao, J. Timmis, and D. R . Flower, "On the hierarchical classification of G protein-coupled receptors., "*Bioinformatics*, vol. 23, no. 23, pp. 3113–8, Dec. 2007, doi: 10.1093/bioinformatics/btm506.

[38] A. Secker, M. N. Davies, A. A. Freitas, J. Timmis, M. Mendao, and D. R. Flower, "An Experimental Comparison of Classification Algorithms for the Hierarchical Prediction of Protein Function Classification of GPCRs, "*Proc. 3rd UK Data Min. Knowl. Discov. Symp.*, 2007.

[39] S. Seo, M. Oh, Y. Park, and S. Kim, "DeepFam: Deep learning based alignment-free method for protein family modeling and prediction, " *Bioinformatics*, vol. 34, no. 13, pp. i254–i262, 2018, doi: 10.1093/bioinformatics/bty275.

[40] M. Zekri, K. Alem, and L. Souici-Meslati, "Immunological Computation for Protein Function Prediction, "*Fundam. Informaticae*, vol. 139, no. February 2014, pp. 91–114, 2015, doi: 10.3233/FI-2015-1227.

[41] Z.-U. Rehman, M. T. Mirza, A. Khan, and H. Xhaard, "Predicting G-protein-coupled receptors families using different physio chemical properties and pseudo amino acid composition., "*Methods Enzymol.*, vol. 522, pp. 61–79, Jan. 2013, doi: 10.1016/B978-0-12-407865-9.00004-2.

[42] M. Naveed and A. U. Khan, "GPCR -MPredictor: Multi-level prediction of G protein-coupled receptors using genetic ensemble, "*Amino Acids*, vol. 42, no. 5, pp. 1809–1823, 2012, doi: 10.1007/s00726-011-0902-6.

[43] A. Secker, M. N. Davies, A. A. Freitas, J. Timmis, E. Clark, and D. R. Flower, "An artificial immune system for clustering amino acids in the context of protein function classification,"*J. Math. Model. Algorithms*, vol. 8, no. 2, pp. 103–123, 2009, doi: 10.1007/s10852-009-9107-3.

[44] Q. Bin Gao, X. F. Ye, and J. He, "Classifying G-protein–coupled receptors to the finest subtype level, "*Biochem. Biophys. Res. Commun.*, vol. 439, no. 2, pp. 303–308, 2013, doi: 10.1016/j.bbrc.2013.08.023.

[45] H. Bin Shen and K. C. Chou, "PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition, "*Anal. Biochem.*, vol. 373, no. 2, pp. 386–388, 2008, doi: 10.1016/j.ab.2007.10.012.

[46] F. Y. Dao *et al*, "Recent advances in conotoxin classification by using machine learning methods," *Molecules*, vol. 22, no. 7, 2017, doi: 10.3390/molecules22071057.

[47] A. Shaker, "Comparison Between Orthogonal and Bi-Orthogonal Wavelets, "*J. Southwest Jiatong Univ.*, vol. 55, no. 2, 2020.

[48] A. Dogra, B. Goyal, and S .Agrawal, "Performance Comparison of Different, "*Asian J. Pharm.*, vol. 2016, no. 4, pp. 9–12, 2016.

[49] J. Kennedy and R. Eberhart, "Particle Swarm Optimisation, "Proc. IEEE Int. Conf. Neural Networks, pp. 1942–1948, 1995, doi: 10.1007/978-3-030-61111-8_2.

[50] İ. B. Aydilek, "A hybrid firefly and particle swarm optimisation algorithm for computationally expensive numerical problems, "Appl. Soft Comput. J., vol. 66, no. February 2018, pp. 232–249, 2018, doi: 10.1016/j.asoc.2018.02.025.

[51] T. T. Ngo, A. Sadollah, and J. H. Kim, "A cooperative particle swarm optimiser with stochastic movements for computationally expensive numerical optimisation problems, "J. Comput. Sci., vol. 13, pp. 68–82, 2016, doi: 10.1016/j.jocs.2016.01.004.

[52] C. N. Silla and A. A. Freitas, "Selecting different protein representations and classification algorithms in hierarchical protein function prediction, "Intell. Data Anal. Journal. Vol. 15, No. 6, vol. 44, no. 0, pp. 979–999, 2011.
[53] S. Bekhouche and Y. M. Ben Ali, "Optimising the identification of GPCR function, " ACMInt. Conf. Proceeding Ser., 2019, doi: 10.1145/3314074.3314082.
[54] B. Yu et al., "Prediction subcellular localisation of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition,"Chemom. Intell. Lab. Syst., vol. 167, no. October, pp. 102–112, 2017, doi: 10.1016/j.chemolab.2017.05.009.
[55] Q. Gu, Y.-S. Ding, and T.-L. Zhang, "Prediction of G-Protein-Coupled Receptor Classes in Low Homology Using Chous Pseudo Amino Acid Composition with Approximate Entropy and Hydrophobicity Patterns, "Protein Pept. Lett., vol. 17, no. 5, pp. 559–567, 2010, doi: 10.2174/092986610791112693.
[56] B. Juba and H. S. Le, "Precision-Recall versus accuracy and the role of large data sets, "33rd AAAI Conf. Artif. Intell. AAAI 2019, 31st Innov. Appl. Artif. Intell. Conf. IAAI 2019 9th AAAI Symp. Educ. Adv. Artif. Intell. EAAI 2019, pp. 4039–4048, 2019, doi: 10.1609/aaai.v33i01.33014039.
[57] R. Paki, E. Nourani, and D. Farajzadeh, "Classification of G protein–coupled receptors using attention mechanism," Gene Reports, vol. 21, no. August, p. 100882, 2020, doi: 10.1016/j.genrep.2020.100882.