

Hemolytic-Pred: A machine learning-based predictor for hemolytic proteins using position and composition-based features

DIGITAL HEALTH
Volume 9: 1–19
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076231180739
journals.sagepub.com/home/dhj



Gulnaz Perveen¹, Fahad Alturise² , Tamim Alkhalifah² 
and Yaser Daanial Khan¹

Abstract

Objective: The objective of this study is to propose a novel in-silico method called Hemolytic-Pred for identifying hemolytic proteins based on their sequences, using statistical moment-based features, along with position-relative and frequency-relative information.

Methods: Primary sequences were transformed into feature vectors using statistical and position-relative moment-based features. Varying machine learning algorithms were employed for classification. Computational models were rigorously evaluated using four different validation. The Hemolytic-Pred webserver is available for further analysis at <http://ec2-54-160-229-10.compute-1.amazonaws.com/>.

Results: XGBoost outperformed the other six classifiers with an accuracy value of 0.99, 0.98, 0.97, and 0.98 for self-consistency test, 10-fold cross-validation, Jackknife test, and independent set test, respectively. The proposed method with the XGBoost classifier is a workable and robust solution for predicting hemolytic proteins efficiently and accurately.

Conclusions: The proposed method of Hemolytic-Pred with XGBoost classifier is a reliable tool for the timely identification of hemolytic cells and diagnosis of various related severe disorders. The application of Hemolytic-Pred can yield profound benefits in the medical field.

Keywords

Hemolysis, hemolytic proteins, XGBoost, statistical moments, machine learning, computational biology, mathematical model

Submission date: 30 December 2022; Acceptance date: 22 May 2023

Introduction

Hemolysis is a process that causes the premature destruction of red blood cells (RBCs) in the bloodstream prematurely before reaching their expected lifespan. The natural lifespan of RBCs is approximately 120 days. After that, they naturally break down and are removed from the blood by the spleen, as the spleen is found in all vertebrates.¹ Anemia is a common blood disorder resulting from a reduction in the production of RBCs and an increase in their destruction. This imbalance causes a decline in the hemoglobin level and oxygen-carrying capacity of the blood. The main causes of hemolysis are acquired or

hereditary² as explained in Figure 1. Certain medications sometimes cause acquired hemolysis. Medications can cause oxidative damage to RBCs, leading to the release

¹Department of Computer Science, School of Systems and Technology, University of Management and Technology, Lahore, Punjab, Pakistan

²Department of Computer, College of Science and Arts in Ar Rass Qassim University, Buraidah, Qassim, Saudi Arabia

Corresponding author:

Fahad Alturise, Department of Computer, College of Science and Arts in Ar Rass, Qassim University, Ar Rass, Qassim, 51452, Qassim, Buraidah 52571, Saudi Arabia.

Email: falturise@qu.edu.sa



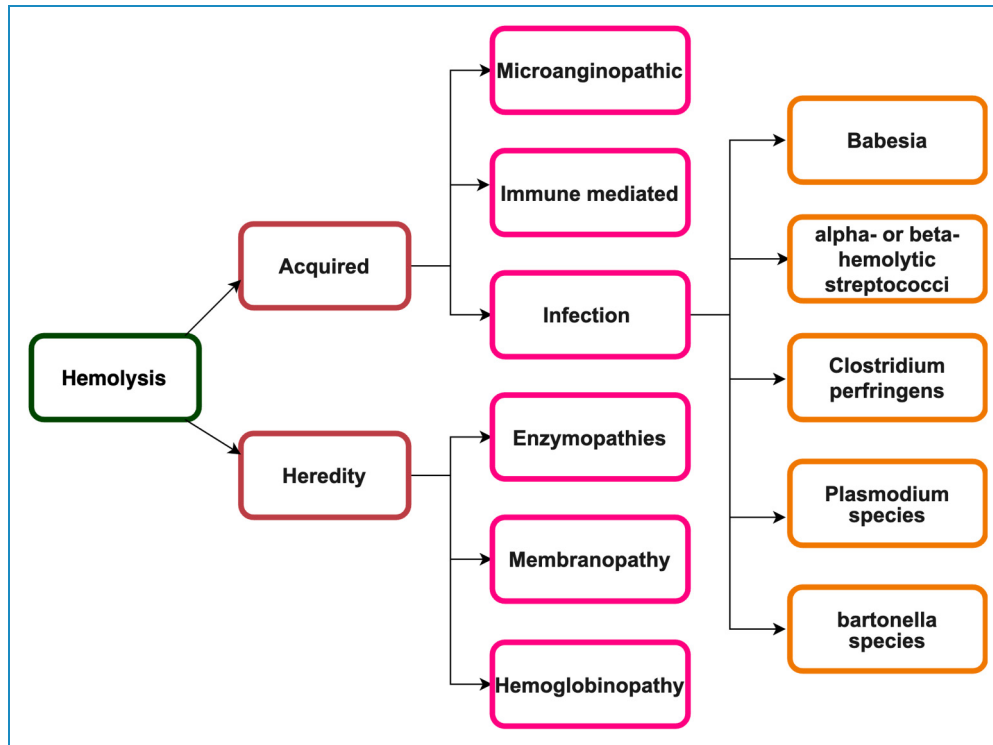


Figure 1. Classification of hemolysis.

of redox-active Hb into the fluid compartment, which is a toxic. Therefore, carefully monitoring patients taking medications that can cause hemolysis and making appropriate dosage adjustments, or discontinuing such medications is critical to prevent further damage.³ Acquired conditions most significantly occur due to pathogens. The best examples of inherited conditions are narratively described as spherocytosis and sickle cell.⁴ Infectious pathogens may cause hemolytic anemia by the direct action of toxins, including alpha- or beta-hemolytic *Streptococci*, *Clostridium perfringens* (*C. welchii*, or *Bacillus welchii*), and *Meningococcus*. Occasionally, RBCs are destroyed by various pathogens of different species such as *Plasmodium* species, *Bartonella* species, and *Babesia* species. Sometimes antibody production against pathogens causes hemolytic anemia that like in the case of Epstein-Barr virus and mycoplasma.^{5–15}

These pathogens produce a toxin in the human body that cause numerous diseases. Researchers, working on these pathogens, need to identify which protein causes hemolysis. A study regarding G-6-P-D scarcity shows RBCs have a “Bitten Apple” like appearance. By using a variety of image processing techniques, an anemic blood sample has been represented. The Wiener filter and Sobel edge detection method are applied for its detection.¹⁶ α 1-microglobulin (A1M) is a ubiquitous reductase and Kristiansson et al.¹⁷ proves via in vitro and in vivo experiments that A1M function in the blood is linked with RBC stability.¹⁸

Subsequently, artificial intelligence provides numerous computational intelligence models that help to assimilate models for predicting the properties of proteins using a primary sequence which can further aid drug discovery and development. Various approaches in bioinformatics have been developed by scientists to identify functional attributes of proteins. Such techniques have been successfully used for sequence analysis of antioxidant proteins,¹⁹ B-cell epitopes,²⁰ anticancer peptides,²¹ anti-CRISPR proteins,²² adaptor proteins²³ and many other such properties that are helpful in finding treatment for varying disorders. Experimental approaches have certain limitations (e.g. time-consuming, laborious, and expensive), so a computational model that can distinguish Hemolytic proteins (HLPs) from non-Hemolytic proteins (non-HLPs) would be highly desirable. HemoPI²⁴ and HemoPred,²⁵ have previously investigated this issue. Chaudhry et al.²⁴ proposed HemoPI which is a computational framework for predicting the hemolytic activities of peptides. It incorporated support vector machines (SVMs) and position-specific compositional information. HemoPred, a method that uses random forests with hybrid features, has been developed by Win et al. However, in all these tools previously proposed, there is a huge gap for improvement in terms of performance. The use of peptide-based drugs in medicine is often limited by their hemotoxic or hemolytic effects. For example, many antimicrobial peptides (AMPs) that are currently in preclinical or clinical applications are only applied

topically, as they can cause hemolysis if administered systemically. AMPs can be highly effective in fighting bacterial infections without the development of bacterial resistance, making them a promising avenue for the treatment of antibiotic-resistant infections.²⁶

The current study aims to identify Hemolytic proteins (HPs) employing the use of various features and different machine learning algorithms. Various rigorous validation techniques were used for evaluation such as self-consistency testing, independent set test, 10-fold cross-validation, and jackknife test are used. In the rest of the article, the “Materials and methods” section contains the proposed approach, the details of the data set used, and explains the statistical moment-based feature extraction. In the “Training machine learning classifiers” section, the feature set is tested on different classifiers to ascertain the best classifier. In the “Results” section, the results of the experimentation are described. The “Discussion” section contains a comparison and discussion in terms of the performance of all employed classifiers. The “Webserver” section contains the details and guidance related to the Web Server. Finally, in the “Conclusions and future work” section, we conclude the article. This study was conducted in the University of Management and Technology and approximately took six months. Major time was spent on data collection, the development of computationally intelligent models, and their validation and testing.

Materials and methods

The proposed approach for detecting HPs is described in the methodology given in Figure 2(a) and (b). The primary purpose of the proposed system is to provide an assiduous model capable of accurately identifying HP patterns. Various features are computed including frequency-dependent features, position-relative features, and statistical moments. Computing is characterized by its simplicity, ability to reach reasonable solutions with minimum knowledge, and efficiency in composite and non-linear problems relating to a process of gradual change and development. Next, the classification was performed using different classifiers, such as AdaBoost (adabst), decision tree (DT), k-nearest neighbors (kNN), neural network (NN), random forest (RF), SVM, and XGBoost (xgboost) (see Figure 3). After evaluation through various techniques, the best performing model was selected which was used for web server deployment.

Formulation of metrics

For performance evaluation, four interlinked performance metrics such as accuracy (Acc), sensitivity (Sn), specificity (Sp), and Matthew’s correlation coefficient (MCC) were employed. Accuracy gives you a broad sense of the

model’s prediction accuracy. Sensitivity refers to the model’s ability to reliably predict positive samples. In the same way, specificity is employed to provide a numerical estimate of the model’s accuracy in predicting negative samples. Taking the imbalanced positive and negative samples in data sets into consideration, MCC is another acceptable amount that evaluates the classification precision of inequality positive and negative samples.

$$Acc = 1 - \frac{F^{\pm} + F^{\mp}}{F^{+} + F^{-}} 0 \leq Acc \leq 1 \quad (1)$$

$$Sn = 1 - \frac{F^{\pm}}{F^{+}} 0 \leq Sn \leq 1 \quad (2)$$

$$Sp = 1 - \frac{F^{\mp}}{F^{-}} 0 \leq Sp \leq 1 \quad (3)$$

$$MCC = \frac{1 - \left(\frac{F^{\pm}}{F^{+}} + \frac{F^{\mp}}{F^{-}} \right)}{\sqrt{\left(1 + \frac{F^{\mp} - F^{\pm}}{F^{+}}\right)\left(1 + \frac{F^{\pm} - F^{\mp}}{F^{-}}\right)}} \quad (4)$$

The total count of non-HP is represented by F^{-} , while the total count of HP is referred to as F^{+} . Those HPs which are inaccurately predicted as non-HPs are F^{\pm} and F^{\mp} are the total count of non-HPs that are inaccurately classified as hemolytic proteins. Thus, equations (1) to (4) refer to the computation of Sp, Sn, Acc, and MCC. MCC value will lay within the range of -1 to 1 . The higher MCC value shows better classifier performance. Various recent research studies^{27–39} have appreciated this set of perceptive metrics shown in equations (1) to (4). Some partitioned tests must be examined to check how well the prediction model has performed. Thus, keeping this in mind, we performed self-consistency test, independent data set testing, 10-fold cross-validation, and jackknife test to extensively evaluate our prediction method.^{40–43}

Data set collection

HP data set was collected from the UniProtKB-SwissProt using the keyword “Hemolysis [KW-0354].” The protein sequences were collected on 2 February 2021. The correctness of data is the key attribute that drives the performance of any desired predictor. A specific well-defined set of rules are used to collect robust and accurate data set as described in previous many studies. Based on these criteria, data set is collected that is best in quality, informative, accurate, and diverse.

We conducted our experiments using both an imbalanced data set (ImBD and a balanced data set (BD). This approach can help you evaluate the performance of the model on different types of data, which can be useful for developing more robust and accurate solutions. This approach can also provide insights into the strengths and weaknesses of your algorithm or model, and help you



Figure 2. Graphical depiction of the whole methodology: (a) from sequences to features and (b) from features to classification.

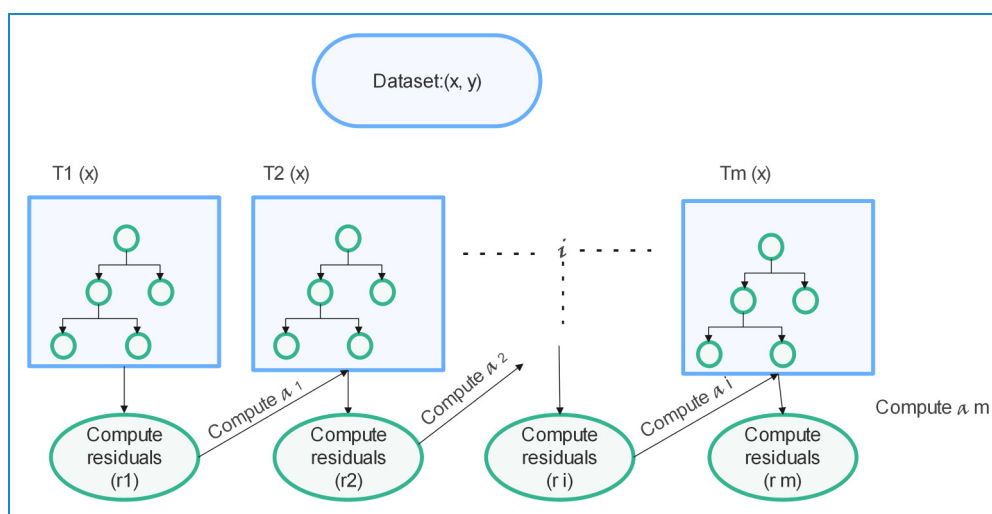


Figure 3. XGBoost illustration.

identify areas for improvement. In the first stage, the sequences with ambiguous annotations like “fragment,” “potential,” “probable,” “probably,” “maybe,” or “by similarity” were excluded from the data set. This yielded a total of 7107 hemolytic proteins out of which, only 946 were reviewed, thus, only these reviewed sequences were included in the positive data set. Similarly, the “non-Hemolysis” proteins were also collected from UniProtKB_SwissProt, by using a converse query, and a total of 980 “non-Hemolysis” reviewed proteins were collected to be used as negative data set. After retrieving data from UniProt, redundancy from data set was reduced by using CD-HIT program with a threshold value of 0.7, that is, sequences having similarity of more than 70% were excluded from data set.⁴⁴ This yielded 329 clusters of positive data set, and 891 clusters of negative data set.

Thus, only 1 representative from each cluster was used, the ImBD comprised 329 positive sequences and 891 negative sequences, and there were 329 positive sequences and 330 negative sequences in the BD. Since the study only involved sequential data collected from well known repositories and not from individual patients, therefore, no patient consent was required for the study.

Feature extraction

The amino acid sequence in a polypeptide chain determines the biophysical characteristics of proteins. An amino acid’s presence or absence does not determine its characteristics. A protein’s behavior is not only affected by amino acid composition, but also by its position in the amino acid cycle. From experience and known data, even the slightest change in the relative positions of amino acids could dramatically impact the entire structure of a protein.^{45,46} Considering these facts, mathematical models that extract characteristics from a primary structure of the protein should not just depend on information about the constituents of these proteins, but should also consider the amino acid positions relative to each other.⁴⁷

Statistical moments. For pattern recognition, statistical moments are used to centralize some key arrangements in the collected data.^{48–56} To describe different properties of data, various statistical moments are utilized such as polynomial and distribution functions. In this study, raw, Hahn, and central moments are employed, as reported in various previous studies such as.^{57–59}

The raw moments are used to evaluate the asymmetry, mean, and variance of a probability distribution. The central moments calculate similar statistics but using the data centroid. With reference to the centroid, the central moments are considered to be as location-invariant whereas these moments are scale-variant. The Hahn moments are based on the Hahn polynomial,^{60,61} and these moments are considered to be as scale-invariant and location-invariant. The above-mentioned moments are

important to calculate the feature vector for protein sequence data because of their sensitivity towards biological sequences. Each moment calculates its own measurements given the data. Moreover, the alteration in the data attributes infers the change in the measurements generated from the moments. In the present study, we used two-dimensional (2D)-matrix representation for the linear composition of protein sequences as

$$A = (a_1, a_2, a_3, \dots, a_t) \quad (5)$$

Using row-major order we get equation (6)

$$p = \lfloor \sqrt{t} \rfloor \quad (6)$$

where p refers to the dimensions of the 2D-square matrix while t is the length of the sample sequence. From (6), matrix A' is generated with m rows and n columns as shown in (7).

$$A' = \begin{bmatrix} b_{1 \rightarrow 1} & b_{1 \rightarrow 2} & \cdots & b_{1 \rightarrow j} & \cdots & b_{1 \rightarrow n} \\ b_{2 \rightarrow 1} & b_{2 \rightarrow 2} & \cdots & b_{2 \rightarrow j} & \cdots & b_{2 \rightarrow n} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ b_{i \rightarrow 1} & b_{i \rightarrow 2} & \cdots & b_{i \rightarrow j} & \cdots & b_{i \rightarrow n} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ b_{m \rightarrow 1} & b_{m \rightarrow 2} & \cdots & b_{m \rightarrow j} & \cdots & b_{m \rightarrow n} \end{bmatrix} \quad (7)$$

A' is the 2D matrix associated with A . The matrix A is transformed into A' using a mapping function ω .

$$\omega(a_m) = b_{xy} \quad (8)$$

Any formation of moments can be evaluated through a matrix or a vector, depicting a sequence pattern. For raw moments, the evaluation A' square matrix is employed. The 2D continuous function order for moments calculation is expressed in equation (7)

$$L_{xy} = \sum_{i=1}^n \sum_{j=1}^n i^x, j^y, b_{ij} \quad (9)$$

For the calculation of raw moments, the center of the data is employed as a reference point. The degree of raw moments is calculated through $x + y$ and refers to those where the sum of $x + y$ is less than or equal to 3. Centroids of data are similar to the gravity center. In relation to the weighted average, a centroid represents the point in data where the data is distributed equally in each direction. After computing raw moments, it can easily be computed.

$$\begin{aligned} \bar{n} &= \frac{L_{10}}{L_{00}}, \\ \bar{m} &= \frac{L_{01}}{L_{00}}. \end{aligned} \quad (10)$$

For central moments, the following relation is used to compute these moments as

$$\mu_{xy} = \sum_{i=1}^n \sum_{j=1}^n (i - \bar{n})^x, (j - \bar{m})^y, b_{ij} \quad (11)$$

A square matrix notation AT has been derived from the one-dimensional notation A . Hahn moments in 2Ds are reversible, which means they provide another leverage. The inverse of these functions of discrete Hahn moments enables one to reconstruct the original data. Additionally, this implies that compositional and positional information of the sequence preserves through these moments. We can calculate the Hahn polynomial order of p as follows

$$h_p^{\mu,v}(r, N) = (N + v - 1)_p (N - 1)_p \times \sum_{t=0}^p (-1)^t \frac{(-p)_t (-r)_t (2N + \mu + v - p - 1)_t}{(N + v - 1)_t (N - 1)_t t!} \quad (12)$$

The pochhammer symbol used in the equations above is generalized as

$$(a)_t = a(a + 1) \cdots (a + t - 1) \quad (13)$$

The Gamma operator is used to simplify

$$(a)_t = \frac{\Gamma(a + t)}{\Gamma(a)} \quad (14)$$

A weighted Hahn moment is generally scaled using a square norm and weighting function

$$\begin{aligned} \widetilde{h}_p^{\mu,v}(r, N) &= h_p^{\mu,v}(r, N) \sqrt{\frac{\rho(r)}{d_p^2}}, p \\ &= 0, 1, 2, 3, \dots, N - 1. \end{aligned} \quad (15)$$

However,

$$\rho(r) = \frac{\Gamma(\mu + r + v) \Gamma(v + r + 1) (\mu + v + r + 1)_N}{(\mu + v + 2r + 1) n! (N - r - 1)!} \quad (16)$$

Furthermore, we compute the normalized orthogonal Hahn moments for 2D matrices by

$$H_{xy}^{\$} = \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} \beta_{xy} \widetilde{h}_x^{\mu,v}(n, N) \widetilde{h}_y^{\mu,v}(m, N), \quad (17)$$

$\$, m, n = 0, 1, 2, 3, \dots, N - 1.$

Thus, for all sequences, raw central and Hahn moments of order 3 are computed which sum up to a total of 30 moments. Data-driven machine learning algorithms have the capability of adapting to uncover patterns hidden within obscure data set.

Computation of position relative incidence matrix (PRIM). To analyze proteins computationally requires a mathematical model that uses sequence order information. The relative positions of the residues are key determinants of protein properties. Furthermore, it is necessary to quantify the relative positions of residues in polypeptide chains. It extracts

the relative positional information of amino acid components from the sequence and forms the PRIM as a matrix of 20×20 elements. PRIM is structured as

$$S_{PRIM} = \begin{bmatrix} G_{1 \rightarrow 1} & G_{1 \rightarrow 2} & \cdots & G_{1 \rightarrow j} & \cdots & G_{1 \rightarrow 20} \\ G_{2 \rightarrow 1} & G_{2 \rightarrow 2} & \cdots & G_{2 \rightarrow j} & \cdots & G_{2 \rightarrow 20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ G_{i \rightarrow 1} & G_{i \rightarrow 2} & \cdots & G_{i \rightarrow j} & \cdots & G_{i \rightarrow 20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ G_{N \rightarrow 1} & G_{N \rightarrow 2} & \cdots & G_{N \rightarrow j} & \cdots & G_{N \rightarrow 20} \end{bmatrix} \quad (18)$$

In the above matrix, the sign of the score of the i th position excess is determined by $G_{i \rightarrow j}$. This value is substituted in the biological evolution change by amino acid type j . The values of $j = 1, 2, 3, 4, \dots, 20$ are shown in the alphabetical order of 20 native amino acids.

Computing reverse position relative incidence matrix (RPRIM).

Machine learning algorithms are either efficient or accurate based on how thoroughly and precisely relevant aspects of data have been extracted. Data mining algorithms can adapt themselves in discovering obscure patterns buried within data by understanding and uncovering them. By using the PRIM, information can be derived about the amino acid positions within poly-peptide chains. Through the introduction of RPRIM, ambiguities in proteins shared by seemingly resembling poly-peptide sequences can be relieved, uncovering further hidden patterns. In RPRIM, there are 400 elements, each having a 20×20 dimension. It is stated as

$$S_{RPRIM} = \begin{bmatrix} R_{1 \rightarrow 1} & R_{1 \rightarrow 2} & \cdots & R_{1 \rightarrow j} & \cdots & R_{1 \rightarrow 20} \\ R_{2 \rightarrow 1} & R_{2 \rightarrow 2} & \cdots & R_{2 \rightarrow j} & \cdots & R_{2 \rightarrow 20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ R_{i \rightarrow 1} & R_{i \rightarrow 2} & \cdots & R_{i \rightarrow j} & \cdots & R_{i \rightarrow 20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ R_{N \rightarrow 1} & R_{N \rightarrow 2} & \cdots & R_{N \rightarrow j} & \cdots & R_{N \rightarrow 20} \end{bmatrix} \quad (19)$$

By computing the raw, central, and Hahn moments of the large RPRIM, the dimension of the matrix is reduced to 24 coefficients, resulting in a feature vector.

Computing frequency vector. Within the primary structure of an amino acid residue, a frequency vector is obtained. The vector is defined as

$$\zeta = (\vartheta_1, \vartheta_2, \vartheta_3, \dots, \vartheta_{20}) \quad (20)$$

Here the number ϑ_i gives the overall count for the i th amino acid. PRIM already contains sequence information related to positions. Protein composition can be determined from the frequency matrix and there is no sequence information

in the frequency matrix.

Computing AAPIV and RAAPIV. AAPIV and RAAPIV are referred to as absolute accumulative position incidence vectors and reverse absolute accumulative position incidence vectors, respectively. This feature extraction model has the capability of extracting all factors related to the protein sequence structure and order. As the frequency vector provides information about the frequency of the presence of each nucleotide base. AAPIV vector can be denoted as

$$C = (\partial_1, \partial_2, \partial_3, \dots, \partial_{20}) \quad (21)$$

In this case, the i th component of AAPIV will be as follows

$$\partial_i = \sum_n^{c=1} s_c \quad (22)$$

Here s_c stands for the arbitrary position in which a nucleotide occurs. Reverse sequencing reveals hidden patterns in protein sequences. RAAPIV is computed the same as AAPIV, but using the reversed input sequence. It is denoted as

$$D = (e_1, e_2, e_3, \dots, e_{20}) \quad (23)$$

the i th component of RAAPIV will be as follows

$$e_i = \sum_n^{D=1} s_D \quad (24)$$

Training machine learning classifiers

All the computed features in previous phases are merged into a feature vector for each data sample. It formulated a feature input matrix of fixed size for any arbitrary sequence irrespective of its length. The matrix formed by combining the feature vector obtained for each sample is further input to the machine learning classifiers for training and further evaluation.

Random forest

Random forest⁶² is a supervised learning algorithm that is used for classification and regression.⁶³ From the name of Random Forest, it represents multiple decision trees that compute their own results. The results from all the trees are collected and based on votes, a random forest makes its own decision for the given data-point. Table 1 shows the hyper-parameters of all the classifiers used in this study.

Support vector machines

SVM⁶⁴ is a robust, flexible supervised algorithm used for classification.⁶⁵ They are the most credible, as well as easy-to-use machine learning methods. They have their rare deployment and are most famous just due to their ability to easily tackle versatile and arranged different

Table 1. Details of hyperparameters of machine learning classifiers.

Classifiers hyper parameter	
Random forest	Number of estimators = 110
	Minimum slip = 3
	Decision trees
Decision trees	Criterion = Gini coefficient
	Minimum split = 2
	Minimum leaf = 1
Support vector machines	Kernel = linear
	Probability = true
	Gamma = scale
AdaBoost	Number of estimators = 60
	Learning rate = 0.1
K-nearest neighbor	Number of neighbors = 5
	Weights initialization = uniform
	Leaf size = 35
Neural network	Hidden layer sizes = (15, 15, 15)
	Maximum iterations = 500
XGBoost	Maximum depth = 5
	Learning rate = 0.4
	Number of estimators = 50

functions. For many scopes, they [SVM] are the face of various classes in a hyperplane. The hyperplane would be made in a repetitive knack by SVM so that the chance of error can be lessened. The goal of SVM is to distinguish the data sets into classes to generate a maximum marginal hyperplane.

Decision trees

Decision tree⁶⁶ is an ensemble learning method for classification algorithms that belongs to supervised learning algorithms. It can be used for classification and the best regression tasks.⁶⁷ It decides the target value by utilizing the features available in the data set. Several metrics are used in decision tree building, such as entropy, information gain, and the Gini index. It is also referred to as a greedy classifier because it attempts to decrease cost at every split. The fundamental working mechanism for

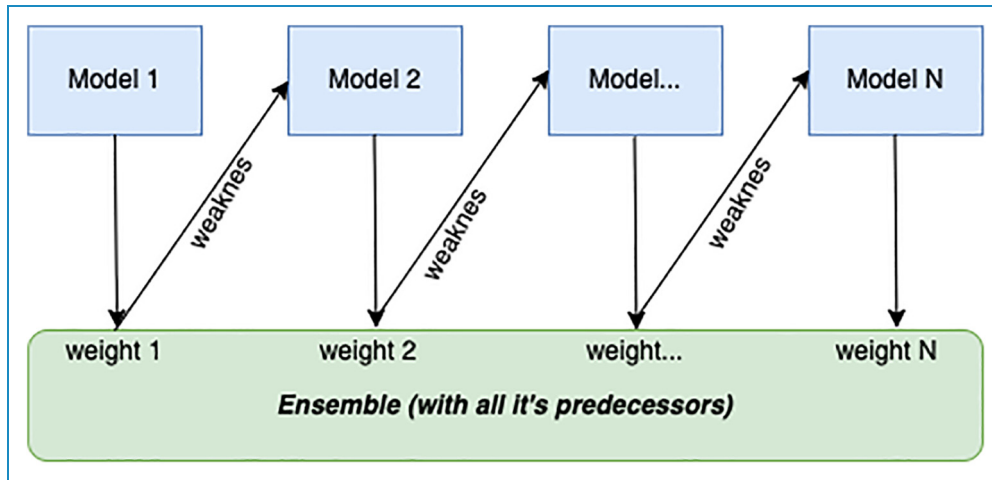


Figure 4. Adaboost illustration.

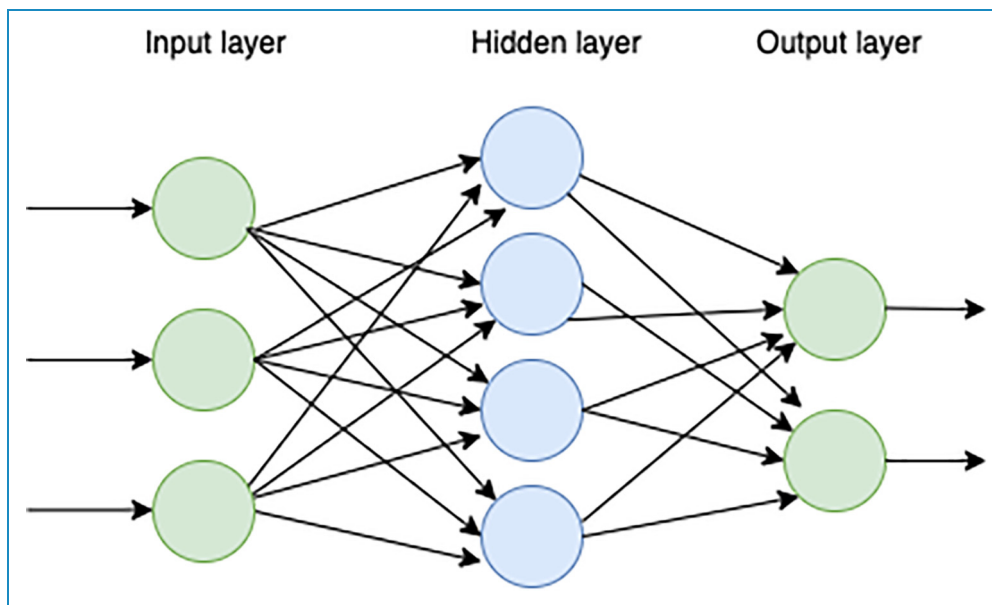


Figure 5. Neural network illustration.

decision trees is to predict the target class variable by learning decision rules from the training data.

AdaBoost

The adaptive boosting algorithm was developed by Freund and Schapire.⁶⁸ Boosting has many other problems that will be overcome by using Adaboost.

Adaboost classifier takes training part $(x_1, y_1), \dots, (x_m, y_m)$ and these features of the data set assigned some weight that is $D(i) = 1/m$, where and create base learner sequentially, if one learner is classified incorrectly, then it is passed to the next classifier to

increase the weight for the incorrectly classified data set and decrease the correctly classified data set. After that, normalize the value using the algorithm for the next learner data set with updated weights and use those parts of the data set for training the incorrectly classified data set from the first learner. This process is repeated for all sequential learners. Figure 4 shows a brief illustration of how the adaptive boosting algorithm works.

K-nearest neighbor

KNN⁶⁹ method is used for both classification and regression and a non-parametric classifier.⁷⁰ It does not consist of any

Table 2. Comparison of XGBoost with existing classifiers in terms of percentage accuracy on self-consistency.

Classifiers	KNN	SVM	RF	Adabst	DT	NN	XGBoost
TP	259	261	298	296	298	287	294
FP	39	37	0	2	0	11	4
TN	790	756	800	794	800	790	797
FN	10	44	0	6	0	10	3
MCC	0.88	0.81	1	0.98	1	0.95	0.98
AUC	0.99	0.97	1	0.99	1	0.99	0.99
F1	0.91	0.87	1	0.99	1	0.96	0.99
Specificity%	95.3	95.33	100	99.75	100	98.63	99.5
Sensitivity%	96.28	85.57	100	98.01	100	96.63	98.99
ImBD accuracy%	95.54	92.62	100	99.27	100	98.09	99.54
BD accuracy%	90.56	90.01	100	98.31	100	98.02	100

adabst: AdaBoost; DT: decision tree; kNN: k-nearest neighbours; NN: neural network; RF: random forest; SVM: support vector machine; xgboost: XGBoost; MCC: Matthews correlation coefficient; AUC: area under the curve; ImBD: imbalanced data set; BD: balanced data set. The best accuracy values attained among all the classifier is highlighted in bold in Table 2.

specific training session; it keeps in touch with the entire data for training during classification. The load needs the value of K meaning the closest data points while it takes training and data sets. K can measure the gap between test data and each racket of training data organized based on distance values, and class them in climbing demand.

Neural network

NN⁷¹ comprises neurons, fixed in layers, that transform an input vector into some outcome. Every neuron carries an input, deploys a function to it, and then hands over the output to the next layer. Networks are defined as feedforward: a unit forages its outcome to entire units on the next layer, but there is no response to the preceding layer. Weights are deployed to the signals passing from one unit to the next, and it is such weightings transformed in the training session to acclimate a NN to the specific issue at hand. An illustration of the NN algorithm can be seen in Figure 5.

XGBoost

eXtreme Gradient Boosting⁷² is an incline enhanced method, that is most famous and resourceful in its operation. It accurately predicts a corresponding variable by combining predictions from weaker samples and models. When used for regression, poor seekers are regression trees,

and each regression tree maps an input data point to one of its leaves consisting of a consistent score. A convex loss function (consisting of the deviation between expected and actual outputs) and a model difficulty reduce the associated organized utility function with a penalty term (which from another perspective can be called tree functions regression). Iterative training inputs, including new trees waiting for residuals or outputs from previous trees, are then combined with previous trees to perform the final computation. It is known as Gradient Boost as it implements a downhill method to reduce losses while incorporating the latest algorithms. Figure 3 is a brief illustration of how gradient tree boosting works.

Results

One of the major steps⁷³ in developing a new prediction technique is how its predicted success rate can be objectively assessed. We look at the following two aspects to fix this:

1. Which measure should be utilized to reflect the predictor quantitatively quality?
2. What type of test method should be used to measure the result?

Thus, herein, the performance of RF, SVM, KNN, DT, AdaBst, XGBoost, and NN is evaluated using statistical

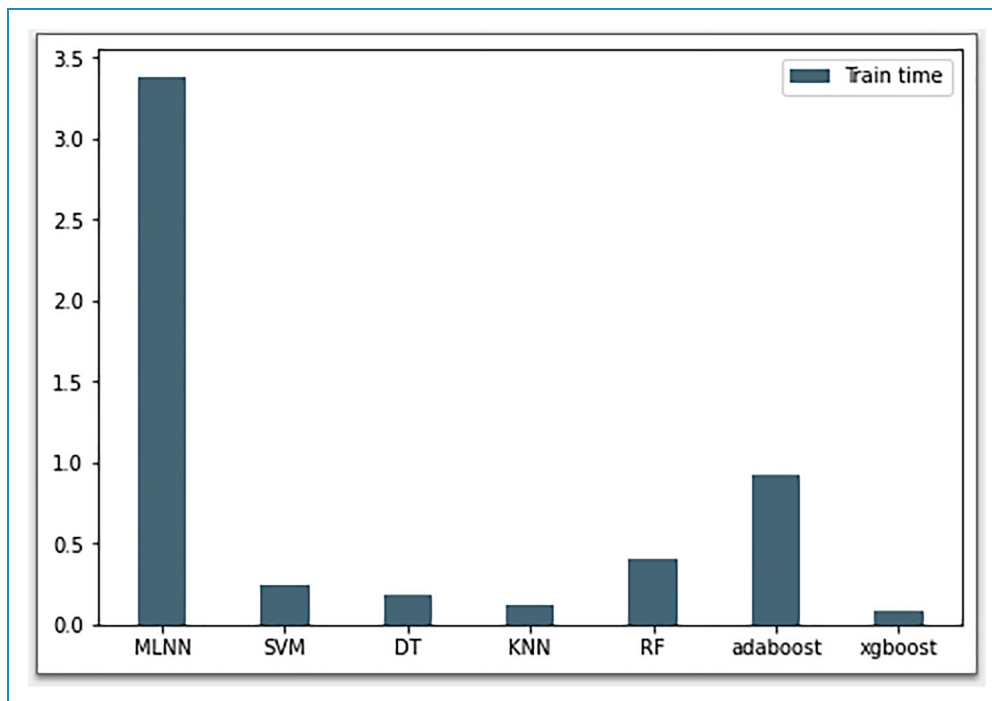


Figure 6. Training time for all classifiers.

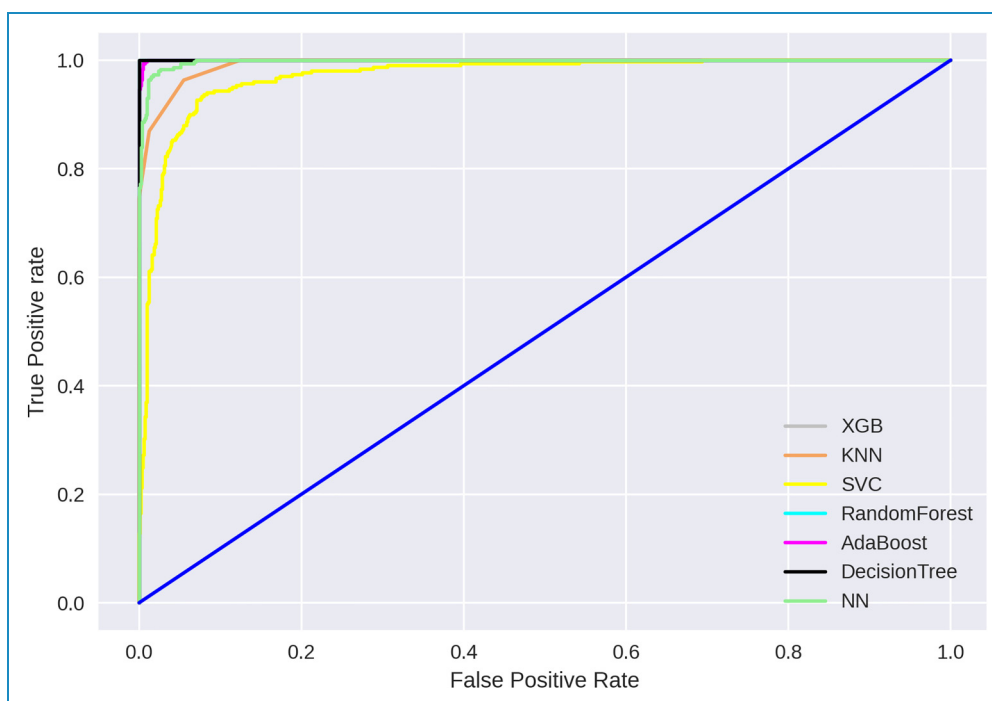


Figure 7. Self-consistency test for all predictors.

Table 3. Comparison of XGBoost with existing classifiers in terms of percentage accuracy on independent test.

Classifiers	KNN	SVM	RF	Adabst	DT	NN	xgboost
TP	59	68	63	67	63	65	67
FP	17	8	13	9	13	11	9
TN	165	159	166	164	160	162	166
FN	3	9	2	4	8	6	2
MCC	0.80	0.83	0.85	0.87	0.79	0.83	0.89
AUC	0.97	0.96	0.97	0.98	0.89	0.96	0.98
F1	0.85	0.89	0.90	0.91	0.85	0.89	0.93
Specificity%	90.66	95.21	92.74	94.8	92.49	93.64	94.86
Sensitivity%	95.16	85.31	96.92	94.37	88.73	91.55	97.1
ImBD accuracy%	91.8	93.03	93.85	94.67	91.39	93.03	95.49
BD accuracy%	89.01	87.01	91.09	90.03	84.02	88.09	91.51

adabst: AdaBoost; DT: decision tree; kNN: k-nearest neighbours; NN: neural network; RF: random forest; SVM: support vector machine; xgboost: XGBoost; MCC: Matthews correlation coefficient; AUC: area under the curve; ImBD: imbalanced data set; BD: balanced data set. Most convincing accuracy metric achieved is depicted in bold in Table 3.

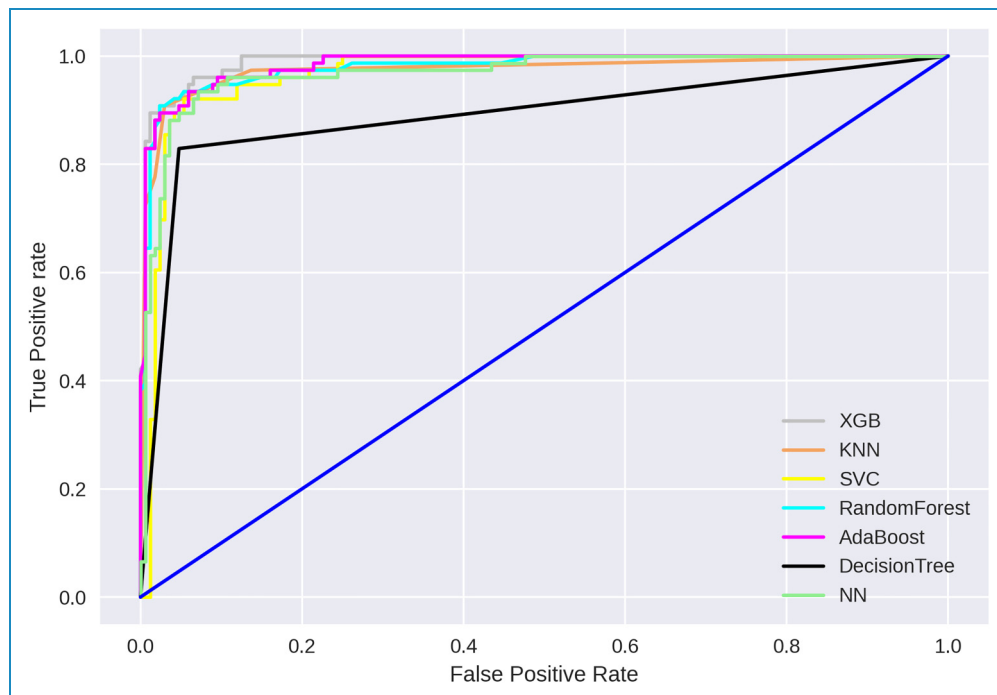


Figure 8. Independent test for all predictors.

Table 4. Comparison of XGBoost with other six classifiers for k -fold cross-validation.

Classifiers	KNN	SVM	RF	Adabst	DT	NN	xgboost
TP	262	285	255	297	270	284	283
FP	67	44	74	50	59	45	46
TN	861	836	878	867	832	850	862
FN	30	55	13	24	59	41	29
MCC	0.79	0.79	0.81	0.84	0.75	0.82	0.84
AUC	0.95	0.96	0.97	0.97	0.87	0.96	0.98
F1	0.84	0.85	0.85	0.88	0.82	0.87	0.89
Specificity%	92.78	95.0	92.23	94.55	93.38	94.97	94.93
Sensitivity%	89.73	83.82	95.15	92.08	88.07	87.38	90.71
ImBD accuracy%	92.05	91.89	92.87	93.93	90.33	92.95	93.9
BD accuracy%	85.1	86.50	90.00	89.92	85.28	88.16	90.12

adabst: AdaBoost; DT: decision tree; KNN: k-nearest neighbours; NN: neural network; RF: random forest; SVM: support vector machine; xgboost: XGBoost; MCC: Matthews correlation coefficient; AUC: area under the curve; ImBD: imbalanced data set; BD: balanced data set. Most assiduous results obtained are illustrated in bold in Table 4.

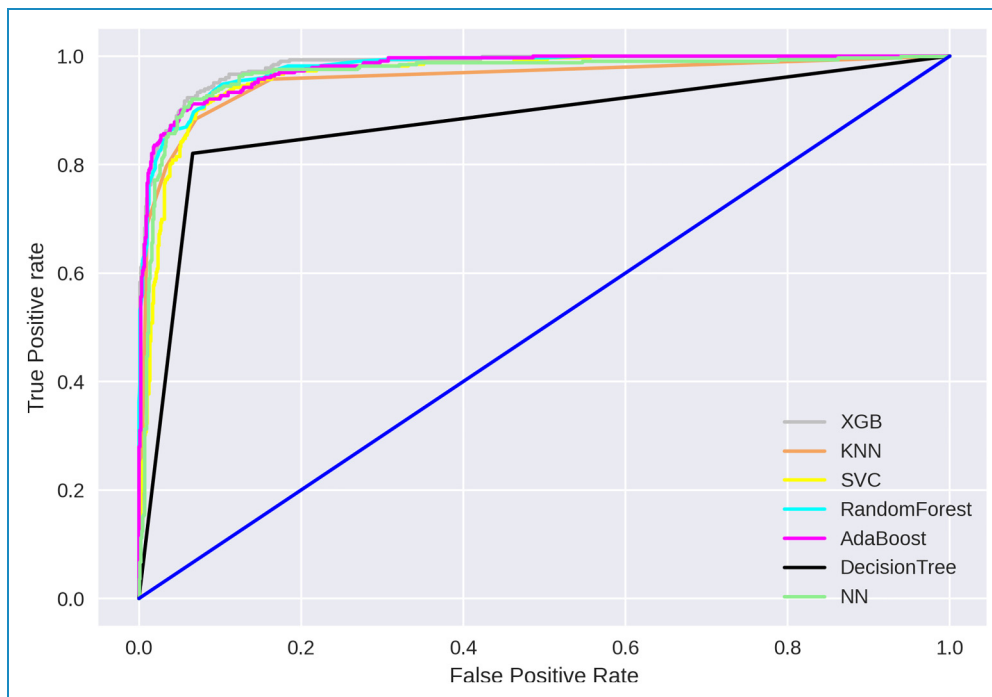
**Figure 9.** Receiver operating characteristics (ROCs) for k -fold cross-validation for all predictors.

Table 5. Comparison of XGBoost with other six classifiers for jackknife test.

Classifiers	KNN	SVM	RF	Adabst	DT	NN	xgboost
TP	265	287	249	286	270	284	292
FP	64	42	80	43	59	45	37
TN	857	837	878	854	828	860	853
FN	34	54	13	37	63	31	38
MCC	0.79	0.80	0.80	0.83	0.74	0.84	0.84
AUC	0.96	0.95	0.96	0.93	0.82	0.95	0.97
F1	0.84	0.85	0.84	0.87	0.81	0.87	0.89
Specificity%	93.05	95.22	91.65	95.21	93.35	95.03	95.84
Sensitivity%	88.63	84.16	95.04	98.54	81.08	90.16	88.48
ImBD accuracy%	91.97	92.13	92.38	93.44	90.0	93.77	93.85
BD accuracy%	86.05	86.65	89.99	89.09	87.43	88.98	90.21

adabst: AdaBoost; DT: decision tree; kNN: k-nearest neighbours; NN: neural network; RF: random forest; SVM: support vector machine; xgboost: XGBoost; MCC: Matthews correlation coefficient; AUC: area under the curve; ImBD: imbalanced data set; BD: balanced data set. The best accuracy values attained are highlighted in bold in Table 5.

moment-based feature extraction of the hemolytic data set.

Self-consistency test

Self-consistency test is one of the most basic test that is usually used to determine the quality of the predictor in terms of training accuracy. Same data set is for training and testing of classifiers, evaluating the training accuracy.

The accurate estimated number of samples for each of the classifiers is used to compute the measured accuracy indicated in Table 2. The time required by each classifier for training is also reported in Figure 6. A comparison of accuracy shown by each predictor displays the classifier receiver operating characteristic (ROC) curve. The performance of the RF and DT predictors is gone good in comparison with other predictors.

In Figure 7, the area under the DT and RF predictor is maximum and Figure 7 depicts that the proposed approach (XGBoost) have AUC, that is, 0.99 but RF and DT have 1.0 AUC.

Independent test

An Independent test is a standard method of working that is used to measure accuracy. In this strategy, we split the entire data set

into two portions. One portion is a large data set used for training, and a small portion of the data set is used for testing.

For this study, we divided the data set into 20% for testing and 80% for training. Accuracy values derived from every predictor shown in Table 3. The test indicates that the XGBoost predictor exceeds the accuracy of other classifiers. The ROC curve in Figure 8 shows the comparison between all machine learning models and great AUC of 0.98 that are xgboost and Adaboost.

k-fold cross-validation

Cross-validation is a most common and well-known approach that is widely used to perform an exhaustive evaluation of a prediction model, as compared to a single independent data set test. Data set is split into k-disjoint folds where training and testing are performed k-times, with k – 1 folds used for training and 1 fold used for testing. Thus, we used 10 folds on a our benchmark data set. One-fold is saved for testing, and the remaining nine folds of the data set is used for training and then repeated for all folds.

We compared the classifier's performance, as shown in Table 4 and Figure 9 depicts that the proposed approach have the highest AUC, that is, 0.98.

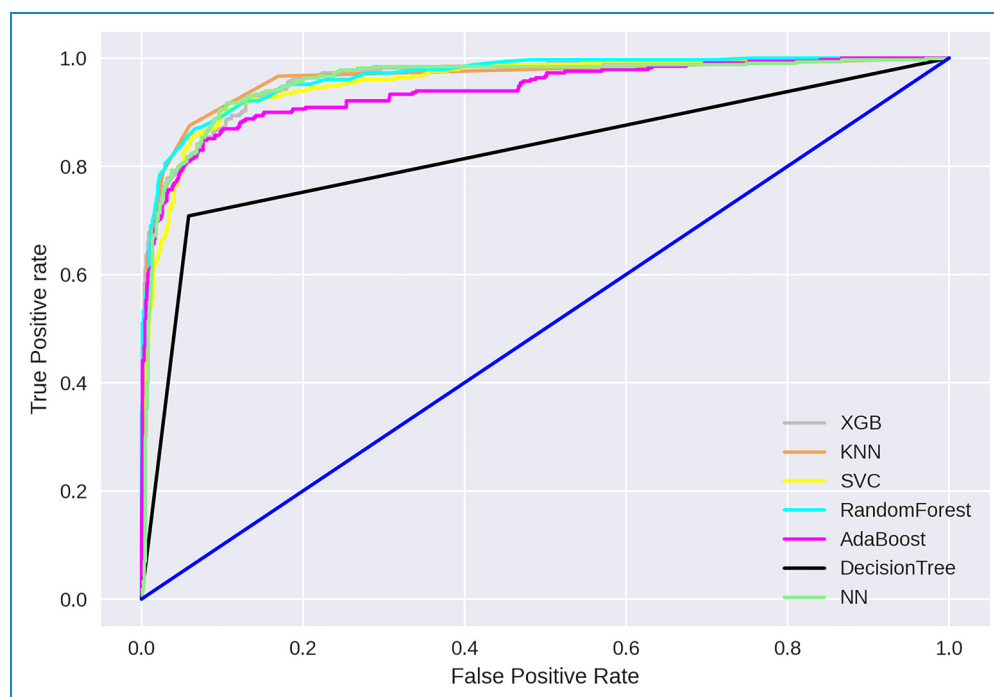


Figure 10. Receiver operating characteristics (ROCs) of jackknife test for all predictors.

Although cross-validation testing is performed on 10-fold partitioned data. It is still possible that some data is missed throughout the test. The jackknife test is a more demanding test that alleviates this issue.

Jackknife test

Jackknife test is also referred as leave-one-out cross-validation test. In this assessment, data is split into no of data folds(n). One entity from the data used for testing and the remaining the data set part are trained by the machine learning classifiers. And this whole process is performed no of data set(n) times, when we got the results of all the n predictions are obtained and take their mean. In cross-validation testing, jackknife is the most crucial and operational cost test because it takes a lot of time. The only downside of choosing the jackknife test for assessment is its feasibility. The computational cost of the testing process is high. This test showed that to validate the quality classifier that is summed up in Table 5. Figure 10 illustrates the comparison of these techniques because it shows that the proposed model (XGBoost) outperforms the other model as it yields the largest AUC (area under the curve) of 0.97 for the curated data set.

Discussion

The identification of hemolysis driver proteins is of paramount importance in precision and customized oncology. To this end, bioinformatics tools that can accurately and efficiently

Table 6. Comparison of Hemolytic-Pred with the state-of-the-art existing methods.

Classifiers	Acc	Sp	Sn	MCC	AUC-ROC
Hemolytic-Pred	0.96	0.94	0.97	0.89	0.98
HLPpred-Fuse ⁷⁴	0.905	0.964	0.845	0.82	0.905
Gradient boosting ⁷⁵	0.95	NA	NA	0.90	0.951
HemoPred ²⁵	0.709	0.728	0.652	0.34	NA
HemoPI ²⁴	0.873	0.94	0.80	0.75	0.952

Acc: accuracy; MCC: Matthew's correlation coefficient; AUC: area under the curve; ROC: receiver operating characteristic.

evaluate sequencing data are highly sought after. In this study, we proposed and validated a novel in-silico method, Hemolytic-Pred, for identifying HPs based on their sequences using statistical moment-based features, position-relative, and frequency-relative information.

Our approach was systematic and reliable, involving the collection of high-quality data, extraction of significant features, training of machine learning algorithms, and rigorous validation. Among all classifiers, XGBoost consistently outperformed the others, achieving the highest accuracy scores across all evaluation metrics. The ROC curves also showed that the proposed method with the XGBoost model performed better than the other classifiers, with an area under the curve of 0.97.

infers that the proposed method can be considered as the most accurate and reliable predictor till now. Furthermore, its importance is realized as an in-silico model in comparison with in vitro and in vivo analysis which are typically both time-consuming and expensive, due to the need for laboratory facilities, specialized equipment, and trained personnel. In addition, these types of experiments can also be ethically and legally complex, as they often involve testing on live animals or human subjects. In-silico analysis, on the other hand, uses computer simulations and modeling to study biological processes, making it a more cost-effective and efficient alternative. It also eliminates the need for animal testing and can provide a more controlled and reproducible environment for testing. As a result, in-silico analysis has become a popular tool for researchers to study various biological processes and to help guide their experimental design.

Webserver

The availability of a scientific method as a tool or webserver is necessary so that the scientists and research community can take benefit from it, as several recent papers have presented.^{76–88} We developed a web server that is user-friendly and its guidelines are also available. It's free and can be used without any login requirement. It is established on "http://ec2-54-160-229-10.compute-1.amazonaws.com/," that is, developed using Python 3.7 Flask for an XGBoost. Step-by-step instructions are provided for the use of the website.

Step 1

Users can open the website at <http://ec2-54-160-229-10.compute-1.amazonaws.com/> and a menu header will appear that consists of four pages, that is, home, hemolytic protein, server, and sample data. The home page provides an overview of proteins. The hemolytic protein tab leads to information on the hemolytic protein. The server tab is the main page that is the prediction portal and the sample data tab provides some positive and negative FASTA format samples.

Step 2

When the server page loads, the user will notice an empty text box that is also depicted in Figure 11, where the input sequence or multiple sequences that will be in FASTA file will be pasted. The submit sequence button will be clicked to obtain results after pasting the sequence. The findings will pop up at the next screen after 0.66 or 1 s shown in Figure 12, which time is dependent on the number of sequences and length of sequences also matter.

Step 3

Click on the sample data tab to find some hemolytic and non-hemolytic sequences for an experiment.

Conclusions and future work

We have proposed a novel tool for the identification of hemolytic proteins as a supplement to experimental approaches. Our model, which extracts features based on statistical moments and position relative features incorporated of proteins and exploits efficient feature selection, was shown to be robust and high performing according to cross-validation and jackknife testing. It is associated with various diseases, mainly tumors, autoimmune, leukemia, sickle cell anemia, lymphoma, and thalassemia. In the present study, out of all the classifiers evaluated, XGBoost outperformed others, giving the most accurate results for the prediction of the HP. The system's overall accuracy on the jackknife testing is 93.85%, sensitivity value is 91.6%, and specificity is 95.03%. It is concluded that the proposed model has the capability of more improvement in the computational result. XG-Boost is an iterative learning method where the model self-analyzes its mistakes and gives more weightage to misclassified data points in the next iteration, making it a reliable model. XGBoost uses a similarity score to prune trees and prevent overfitting. It is a better option for unbalanced data sets and more efficient in optimizing hyperparameters compared to Random Forest and other classifiers. XGBoost is a more preferable choice in situations like Poisson regression and rank regression.

As the biological sequence data increases day-by-day at high speed in a different type of database like the Swiss Prot database, in the future, the space to improve efficiency in this field still exists due to the increasing number of data sets. A number of diseases are associated with HPs, including sickle cell anemia, glucose-6-phosphate dehydrogenase (G6PD) deficiency, hemolytic uremic syndrome, thalassemia, autoimmune hemolytic anemia, pyruvate kinase deficiency, spherocytosis, G6PD deficiency, and paroxysmal nocturnal hemoglobinuria. In order to treat or cure the aforementioned diseases, scientists would need to determine whether a protein is hemolytic or non-hemolytic. The researcher could benefit from our reliable and accurate model in this case. Using this model, it might be possible to design and discover drugs for the diseases listed above.

Acknowledgements: Researchers would like to thank the Deanship of Scientific Research, Qassim University for funding the publication of this project.

Contributorship: GP conducted experimentation and validation, FA contributed in data collection, analysis and write up, TA was responsible for conceptualization and implementation, YDK supervised the work.


Declaration of conflicting interests: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: The study did not involve any human or animal tests, further, no ethical approvals were required.

Funding: The author(s) received no financial support for the research, authorship, and/or publication of this article.

Guarantor: The submitting author FA is the guarantor of the article.

ORCID iDs: Fahad Alturise  <https://orcid.org/0000-0001-9176-7984>

Tamim Alkhalifah  <https://orcid.org/0000-0001-8407-2068>

Supplemental material: Supplemental material for this article is available online. All data generated or analyzed during this study are included in this published article (and its Supplemental Material File S1).

References

- Dhaliwal G, Cornett PA and Tierney Jr LM. Hemolytic anemia. *Am Fam Physician* 2004; 69: 2599–2606.
- Faghih MM and Sharp MK. Modeling and prediction of flow-induced hemolysis: a review. *Biomech Model Mechanobiol* 2019; 18: 845–881.
- Gladwin MT, Kanas T, Kim-Shapiro DB, et al. Hemolysis and cell-free hemoglobin drive an intrinsic mechanism for human disease. *J Clin Invest* 2012; 122: 1205–1208.
- Manciu S, Matei E and Trandafir B. Hereditary spherocytosis-diagnosis, surgical treatment and outcomes. A literature review. *Chirurgia (Bucur)* 2017; 112: 110–116.
- Barcellini W, Giannotta J and Fattizzo B. Autoimmune hemolytic anemia in adults: primary risk factors and diagnostic procedures. *Expert Rev Hematol* 2020; 13: 585–597.
- Duineveld C, Verhave JC, Berger SP, et al. Living donor kidney transplantation in atypical hemolytic uremic syndrome: a case series. *Am J Kidney Dis* 2017; 70: 770–777.
- Uzal FA, Freedman JC, Shrestha A, et al. Towards an understanding of the role of *Clostridium perfringens* toxins in human and animal disease. *Future Microbiol* 2014; 9: 361–377.
- Pichichero ME. Group A beta-hemolytic *Streptococcal* infections. *Pediatr Rev* 1998; 19: 291–302.
- Cunningham MW. Pathogenesis of group A *Streptococcal* infections. *Clin Microbiol Rev* 2000; 13: 470–511.
- Rafiqdoust H, Ahangarzadeh S, Yarian F, et al. Bioinformatics prediction and experimental validation of VH antibody fragment interacting with *Neisseria meningitidis* factor H binding protein. *Iran J Basic Med Sci* 2020; 23: 1053.
- Bakhraibah AO. Malaria in modern day: a review article. *Aus J Basic Appl Sci* 2018; 12: 73–75.
- Maurin M, Birtles R and Raoult D. Current knowledge of *Bartonella* species. *Eur J Clin Microbiol Infect Dis* 1997; 16: 487–506.
- Barbosa COS, Garcia JR, Fava NMN, et al. Babesiosis caused by *Babesia vogeli* in dogs from Uberlândia State of Minas Gerais, Brazil. *Parasitol Res* 2020; 119: 1173–1176.
- Kimura H, Hoshino Y, Kanegane H, et al. Clinical and virologic characteristics of chronic active Epstein-Barr virus infection. *Blood* 2001; 98: 280–286.
- Masukagami Y, Nijagal B, Mahdizadeh S, et al. A combined metabolomic and bioinformatic approach to investigate the function of transport proteins of the important pathogen *Mycoplasma bovis*. *Vet Microbiol* 2019; 234: 8–16.
- Rakshit P and Bhowmik K. Detection of abnormal findings in human RBC in diagnosing G-6-P-D deficiency haemolytic anaemia using image processing. In *2013 IEEE 1st International Conference on Condition Assessment Techniques in Electrical Systems (CATCON)*, (pp.297–302, 2013).
- Kristiansson A, Bergwik J, Alattar AG, et al. Human radical scavenger α 1-microglobulin protects against hemolysis in vitro and α 1-microglobulin knockout mice exhibit a macrocytic anemia phenotype. *Free Radical Biol Med* 2021; 162: 149–159.
- Richmond CM and et al. Rapid identification of biallelic SPTB mutation in a neonate with severe congenital hemolytic anemia and liver failure. *Mol Syndromol* 2020; 11: 50–55.
- Ahmed S, Arif M, Kabir M, et al. PredAODP: accurate identification of antioxidant proteins by fusing different descriptors based on evolutionary information with support vector machine. *Chemometr Intell Lab Syst* 2022; 228: 104623.
- Alghamdi W, Attique M, Alzahrani E, et al. LBCEPred: a machine learning model to predict linear B-cell epitopes. *Brief Bioinformatics* 2022; 23: bbac035.
- Arif M, Ahmed S, Ge F, et al. StackACPred: prediction of anticancer peptides by integrating optimized multiple feature descriptors with stacked ensemble approach. *Chemometr Intell Lab Syst* 2022; 220: 104458.
- Dao F-Y, Liu M-L, Su W, et al. AcrPred: a hybrid optimization with enumerated machine learning algorithm to predict anti-CRISPR proteins. *Int J Biol Macromol* 2023; 228: 706–714.
- Wang Y-H, Zhang Y-F, Zhang Y, et al. Identification of adaptor proteins using the ANOVA feature selection technique. *Methods* 2022; 208: 42–47.
- Chaudhary K, Kumar R, & Tuknait A, et al. A web server and mobile app for computing hemolytic potency of peptides. *Scientific Reports* 2016; 6: 22843.
- Win TS, Malik AA, Prachayasittikul V, et al. HemoPred: a web server for predicting the hemolytic activity of peptides. *Future Med Chem* 2017; 9: 275–291.
- Haney EF, Straus SK and Hancock REW. Reassessing the host defense peptide landscape. *Front Chem* 2019; 7: 43.
- Wei L, Ye X, Sakurai T, et al. ToxIBTL: prediction of peptide toxicity based on information bottleneck and transfer learning. *Bioinformatics* 2022; 38: 1514–1524.
- Liu B, Wang S, Long R, et al. iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics* 2017; 33: 35–41.
- Zhang C-J, Tang H, Li W-C, et al. iOri-Human: identify human origin of replication by incorporating dinucleotide physicochemical properties into pseudo nucleotide composition. *Oncotarget* 2016; 7: 69783.
- Liu B, Yang F, Huang D-S, et al. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics* 2018; 34: 33–40.
- Ehsan A, Mahmood K, Khan YD, et al. A novel modeling in mathematical biology for classification of signal peptides. *Sci Rep* 2018; 8: 1–16.

32. Naseer S, Hussain W, Khan YD, et al. Optimization of serine phosphorylation prediction in proteins by comparing human engineered features and deep representations. *Anal Biochem* 2021; 615: 114069.
33. Naseer S, Hussain W, Khan YD, et al. Sequence-based identification of arginine amidation sites in proteins using deep representations of proteins and PseAAC. *Curr Bioinform* 2020; 15: 937–948.
34. Khan YD, Alzahrani E, Alghamdi W, et al. Sequence-based identification of allergen proteins developed by integration of PseAAC and statistical moments via 5-step rule. *Curr Bioinform* 2020; 15: 1046–1055.
35. Malebary SJ and Khan YD. Identification of antimicrobial peptides using Chou's 5 step rule. *CMC-Computers, Materials and Continua* 2021; 67: 2863–2881.
36. Malebary SJ, Khan R and Khan YD. ProtoPred: advancing oncological research through identification of proto-oncogene proteins. *IEEE Access* 2021; 9: 68788–68797.
37. Naseer S, Hussain W, Khan YD, et al. NPalmityl Deep-PseAAC: a predictor of N-palmitoylation sites in proteins using deep representations of proteins and PseAAC via modified 5-steps rule. *Curr Bioinform* 2021; 16: 294–305.
38. Malebary SJ and Khan YD. Evaluating machine learning methodologies for identification of cancer driver genes. *Sci Rep* 2021; 11: 1–13.
39. Awais M, Hussain W, Rasool N, et al. iTSP-PseAAC: identifying tumor suppressor proteins by DIFdelU using fully connected neural network and PseAAC. *Curr Bioinform* 2021; 16: 700–709.
40. Khan YD, Rasool N, Hussain W, et al. iPhosT-pseAAC: identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC. *Anal Biochem* 2018; 550: 109–116.
41. Hussain W, Khan YD, Rasool N, et al. SPrenylC-PseAAC: a sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-prenylation sites in proteins. *J Theor Biol* 2019; 468: 1–11.
42. Awais M et al. iPhosH-PseAAC: identify phosphohistidine sites in proteins by blending statistical moments and position relative features according to the Chou's 5-step rule and general pseudo amino acid composition. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2019; 18: 596–610.
43. Khan YD, Amin N, Hussain W, et al. iProtease-PseAAC (2L): a two-layer predictor for identifying proteases and their types using Chou's 5-step-rule and general PseAAC. *Anal Biochem* 2020; 588: 113477.
44. Huang Y, Niu B, Gao Y, et al. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010; 26: 680–682.
45. Chen J et al. A comprehensive review and comparison of different computational methods for protein remote homology detection. *Brief Bioinformatics* 2018; 19: 231–244.
46. Liu B, Liu F, Fang L, et al. reprNA: a web server for generating various feature vectors of RNA sequences. *Mol Genet Genomics* 2016; 291: 473–481.
47. Brown DP, Krishnamurthy N and Sjölander K. Automated protein subfamily identification and classification. *PLoS Comput Biol* 2007; 3: e160.
48. Lo C-H and Don H-S. 3-D moment forms: their construction and application to object identification and positioning. *IEEE Trans Pattern Anal Mach Intell* 1989; 11: 1053–1064.
49. Zhou J, Shu H, Zhu H, et al. Image analysis by discrete orthogonal Hahn moments. In *International Conference Image Analysis and Recognition* (pp.524–531). Springer, 2005.
50. Hussain W, Khan YD, Rasool N, et al. SPalmitylC-PseAAC: a sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-palmitoylation sites in proteins. *Anal Biochem* 2019; 568: 14–23.
51. Khan YD, Batool A, Rasool N, et al. Prediction of nitrosocysteine sites using position and composition variant features. *Lett Org Chem* 2019; 16: 283–293.
52. Rasool N, Iftikhar S, Amir A, et al. Structural and quantum mechanical computations to elucidate the altered binding mechanism of metal and drug with pyrazinamidase from *Mycobacterium tuberculosis* due to mutagenicity. *Journal of Molecular Graphics and Modelling* 2018; 80: 126–131.
53. Rasool N, Hussain W and Mahmood S. Prediction of protein solubility using primary structure compositional features: a machine learning perspective. *Journal of Proteomics and Bioinformatics* 2017; 10: 324–328.
54. Hussain W, Qaddir I, Mahmood S, et al. In silico targeting of non-structural 4B protein from dengue virus 4 with spiropyrazolopyridone: study of molecular dynamics simulation, ADMET and virtual screening. *VirusDisease* 2018; 29: 1–10.
55. Khan YD, Jamil M, Hussain W, et al. pSSbond-PseAAC: prediction of disulfide bonding sites by integration of PseAAC and statistical moments. *J Theor Biol* 2019; 463: 47–55.
56. Saeed S, Mahmood MK and Khan YD. An exposition of facial expression recognition techniques. *Neural Computing and Applications* 2018; 29: 425–443.
57. Shah AA and Khan YD. Identification of 4-carboxyglutamate residue sites based on position based statistical feature and multiple classification. *Sci Rep* 2020; 10: 1–10.
58. Butt AH and Khan YD. CanLect-Pred: a cancer therapeutics tool for prediction of target cancerlectins using experiential annotated proteomic sequences. *IEEE Access*, vol. 8, pp.9520–9531, 2019.
59. Hussain W, Rasool N and Khan YD. A sequence-based predictor of Zika virus proteins developed by integration of PseAAC and statistical moments. *Comb Chem High Through Screen* 2020; 23: 797–804.
60. Mahmood MK, Ehsan A, Khan YD, et al. iHyd-LysSite (EPSV): identifying hydroxylysine sites in protein using statistical formulation by extracting enhanced position and sequence variant feature technique. *Curr Genomics* 2020; 21: 536–545.
61. Naseer S, Hussain W, Khan YD, et al. IPhosS (deep)-PseAAC: identify phosphoserine sites in proteins using deep learning on general pseudo amino acid compositions via modified 5-steps rule. *IEEE/ACM Trans Comput Biol Bioinformatics* 2020; 19: 1703–1714.
62. Pal M. Random forest classifier for remote sensing classification. *Int J Remote Sens* 2005; 26: 217–222.
63. Salazar A, Safont G, Vergara L, et al. Pattern recognition techniques for provenance classification of archaeological ceramics using ultrasounds. *Pattern Recognit Lett* 2020; 135: 441–450.
64. Stecanella B. An Introduction to Support Vector Machines (SVM), 2017.
65. Vapnik V. *The nature of statistical learning theory*. New York, Berlin, Heidelberg: Springer Science and Business Media, 2013.

66. Swain PH and Hauska H. The decision tree classifier: design and potential. *IEEE Trans Geosci Electron* 1977; 15: 142–147.
67. Safavian SR and Landgrebe D. A survey of decision tree classifier methodology. *IEEE Trans Syst Man Cybern* 1991; 21: 660–674.
68. Freund Y, Schapire R and Abe N. A short introduction to boosting. *J-Japanese Soc Artif Intell* 1999; 14: 1612.
69. Liao Y and Vemuri VR. Use of k-nearest neighbor classifier for intrusion detection. *Comput Secur* 2002; 21: 439–448.
70. Yang JM, Yu PT and Kuo BC. A nonparametric feature extraction and its application to nearest neighbor classification for hyperspectral image data. *IEEE Trans Geosci Remote Sens* 2009; 48: 1279–1293. IEEE. n.ñÁAmino Acids 42, no. 6 (2012): 2447–2460.
71. Denoeux T. A neural network classifier based on Dempster-Shafer theory. *IEEE Trans Syst, Man, Cybernetics-Part A: Syst Humans* 2000; 30: 131–150.
72. Chen T and Guestrin C. Xgboost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp.785–794. 2016.
73. Chou K-C, Wu Z-C and Xiao X. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PLoS ONE* 2011; 6: e18258.
74. Hasan MM, Schaduagratt N, Basith S, et al. HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. *Bioinformatics* 2020; 36: 3350–3356.
75. Plisson F, Ram'irez-S'anchez O and Mart'inez-Hern'andez C. Machine learning-guided discovery and design of non-hemolytic peptides. *Sci Rep* 2020; 10: 1–19.
76. Wei L, Ye X, Xue Y, et al. ATSE: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. *Brief Bioinformatics* 2021; 22: bbab041.
77. Xu Y, Wang Z, Li C, et al. iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. *Med Chem (Los Angeles)* 2017; 13: 544–551.
78. Chen W, Feng P, Yang H, et al. iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. *Oncotarget* 2017; 8: 4208–4200.
79. Chen W, Feng P-M, Lin H, et al. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Res* 2013; 41: e68–e68.
80. Cheng X, Xiao X and Chou KC. pLoc-mEuk: predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC. *Genomics* 2018; 110: 50–58.
81. Cheng X, Zhao S-G, Lin W-Z, et al. pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites. *Bioinformatics* 2017; 33: 3524–3531.
82. Cheng X, Zhao S-G, Xiao X, et al. iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. *Bioinformatics* 2017; 33: 341–346.
83. Cheng X, Lin WZ, Xiao X, et al. pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC. *Bioinformatics* 2019; 35: 398–406.
84. Cheng X, Xiao X and Chou K. pLoc_bal-mGneg: predict subcellular localization of Gram-negative bacterial proteins by quasi-balancing training dataset and general PseAAC. *J Theor Biol* 2018; 458: 92–102.
85. Chou K, Cheng X and Xiao X. pLoc_bal-mHum: predict subcellular localization of human proteins by PseAAC and quasi-balancing training dataset. *Genomics* 2019; 111: 1274–1282.
86. Xiao X, Cheng X, Chen G, et al. pLoc_bal-mGpos: predict subcellular localization of Gram-positive bacterial proteins by quasi-balancing training dataset and PseAAC. *Genomics* 2019; 111: 886–892.
87. Malebary SJ, Rehman MSU and Khan YD. iCrotoK-PseAAC: identify lysine crotonylation sites by blending position relative statistical features according to the Chou's 5-step rule. *PLoS ONE* 2019; 14: e0223993.
88. Akmal MA, Hussain W, Rasool N, et al. Using Chou's 5-steps rule to predict O-linked serine glycosylation sites by blending position relative features and statistical moment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020.