scientific reports

Check for updates

OPEN

Prediction of antifreeze proteins using machine learning

Adnan Khan[1], Jamal Uddin[1], Farman Ali[2,7]✉, Ashfaq Ahmad[3], Omar Alghushairy[4], Ameen Banjar[4] & Ali Daud[5,6]✉

Living organisms including fishes, microbes, and animals can live in extremely cold weather. To stay alive in cold environments, these species generate antifreeze proteins (AFPs), also referred to as ice-binding proteins. Moreover, AFPs are extensively utilized in many important fields including medical, agricultural, industrial, and biotechnological. Several predictors were constructed to identify AFPs. However, due to the sequence and structural heterogeneity of AFPs, correct identification is still a challenging task. It is highly desirable to develop a more promising predictor. In this research, a novel computational method, named AFP-LXGB has been proposed for prediction of AFPs more precisely. The information is explored by Dipeptide Composition (DPC), Grouped Amino Acid Composition (GAAC), Position Specific Scoring Matrix-Segmentation-Autocorrelation Transformation (Sg-PSSM-ACT), and Pseudo Position Specific Scoring Matrix Tri-Slicing (PseTS-PSSM). Keeping the benefits of ensemble learning, these feature sets are concatenated into different combinations. The best feature set is selected by Extremely Randomized Tree-Recursive Feature Elimination (ERT-RFE). The models are trained by Light eXtreme Gradient Boosting (LXGB), Random Forest (RF), and Extremely Randomized Tree (ERT). Among classifiers, LXGB has obtained the best prediction results. The novel method (AFP-LXGB) improved the accuracies by 3.70% and 4.09% than the best methods. These results verified that AFP-LXGB can predict AFPs more accurately and can participate in a significant role in medical, agricultural, industrial, and biotechnological fields.

AFP (Antifreeze protein) is essential for various species like animals, fish, plants, and microorganisms living in highly cold regions[1]. In ice recrystallization, small ice crystals bind with adjacent water molecules and form large ice crystal[2]. This ice recrystallization phenomenon is hazardous for cold-blooded organisms due to the formation of ice in their bodies. AFP interacts with small ice crystals and prevents or retards the ice recrystallization progression that leads to the survival of the cold-blooded living organisms in subzero and low-temperature environments[3]. AFP has other diverse significant applications including food preservation, human cryopreservation and cryosurgery improving, boosting freeze tolerance, ice and yogurt formation[4,5]. AFP possesses the characteristic of reducing the water freezing point without altering melting point. This property of AFP is called thermal hysteresis[6].

In respect of above the significance, accurate identification of AFP is essential. A series of methods have been established for identification of AFP. For example, Kandaswamy et al. developed a method, called AFP-Pred to discriminate AFP from non-AFP. They used short peptides, secondary structure properties, physicochemical features, and RF (Random Forest) as training model[7]. In another approach (AFP-PSSM), these authors utilized evolutionary information in conjunction with SVM[8]. Yu et al. adopted multi-respective several composition features such as TPC, DPC, and AAC. They selected the best patterns via genetic algorithm and prediction was carried out by SVM. They also established a web server, called iAFP[9].

Onward, Mondel et al. proposed AFP-PseAAC predictor employing PseAAC (Pseudo Amino Acid Composition) with SVM[10]. In another TargetFreeze protocol, the features are discovered by AAC, PseAAC, and PsePSSM, fused all patterns, and perform the prediction using SVM[11]. Pratiwi et al. adopted AAC, DPC, and physicochemical properties for feature engineering and RF as a classifier. Their novel predictor is called CryoProtect[12]. In RAFP-Pred predictor, authors split each protein sequence into two sub-sequences. Features from each part were

[1]Qurtuba University of Science and Technology, Peshawar, Khyber Pakhtunkhwa, Pakistan. [2]Department of Elementary and Secondary Education, Peshawar, Khyber Pakhtunkhwa, Pakistan. [3]Department of Computer Science, Abdul Wali Khan University Mardan, Mardan, Pakistan. [4]Department of Information Systems and Technology, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia. [5]Abu Dhabi School of Management, Abu Dhabi, United Arab Emirates. [6]Department of Computer Science and Artificial Intelligence, University of Jeddah, Jeddah, Saudi Arabia. [7]Sarhad University of Science and Information Technology, Mardan, Pakistan. ✉email: farman335@yahoo.com; alimsdb@gmail.com

abstracted by AAC and DPC. Info-Gain algorithm was implemented for selection of optimal features and the model was trained using RF classifier[13]. Usman et al. proposed AFP-LSE predictor. They used autoencoder with Composition of K-spaced amino acid pairs and achieved a balanced accuracy of 0.903[14]. In another work, Usman et al. constructed AFP-SRC improved method[15]. Similarly, PoGB-pred approach is developed by Alim et al. They employed PseAAC, AAC, and DPC as feature descriptors and PCA for reducing the feature dimension[16]. Recently, Miyata et al. designed a novel predictor using new datasets. They applied CD, DC, AAC for feature encoding and Light eXtreme Gradient Boosting machine for model learning[17].

Although, each prediction system made efforts to predict antifreeze proteins. However, due to the variant behavior of AFP structure and sequences, it is still highly desirable to predict AFP more accurately. Considering this, we developed a protocol, named AFP-LXGB for accurate prediction of antifreeze proteins.

Proposed method. In the design of AFP-LXGB predictor, we carried out the following contribution.

- Extracted the sequential patterns via GAAC, DPC, and evolutionary features by PSSM.
- To extract the local information, segmentation notion is extended into PSSM and split each PSSM into three segments. Further, the autocorrelation transformation (ACT) strategy is applied to each segment and finally combines all segments. Thus, a novel feature descriptor is introduced named Sg-PSSM-ACT.
- Developed another feature representative method named PseTS-PSSM. In this method, PSSM of the each sequence is decomposed into three slices. Further, the sequence-order patterns are computed using Pseudo strategy by extending to each slice and combined all the slices into one super set.
- Concatenated feature vectors into different groups and provided to RF, ERT, and LXGB for model training.
- A novel feature selection method namely ERT-RFE is introduced for the selection of optimal features.

The schematic view of the proposed work has been described in Fig. 1 and detailed in the following subsections.

Materials and methods

Benchmark datasets. To construct a promising method, we implemented datasets widely utilized by previous works such as AFP-Pred[7], AFP-PSSM[8], iAFP[9], AFP-PseAAC[10], and CryoProtect[12]. The AFPs (positive) set comprises 481 AFPs sequences. Similarly, the negative set containing 9193 non-AFPs instances was collected from Pfam protein families as explained in[7]. The datasets are provided in supplementary file.

Feature formulation techniques. Discovering the discriminative features by appropriate schemes is an important step in the design of an effective computational model[20]. In this regard, GAAC, DPC, Sg-PSSM-ACT, and PseTS-PSSM are used for exploring the salient patterns from primary sequences of AFPs.

Grouped amino acid composition. The simple Amino Acid Composition (AAC) comprises 20 amino acids that compute the frequency of each amino acid[21]. GAAC classifies the 20 amino acids into five groups using the physicochemical properties. The five classes contain aliphatic group (G1: AGLIMV), negative charged group (G2: DE), aromatic group (G3: FWY), positive charge group (G4: HRK), and uncharged group (G5: CNPQST). GAAC calculates the frequency of each group using the following equation:

$$F(G) = \frac{n(G)}{n}, \quad G \in (G1, G2, G3, G4, G5), \tag{1}$$

$$n(G_a) = \sum n(a), \quad a \in G, \tag{2}$$

where $n(G)$ represents the amino acids in a group G, $n(a)$ indicates the amino acid type a, and n shows the length of sequence. GAAC extracts 5 features.

Dipeptide composition. DPC formulates frequencies of two connected amino acids of a protein sequence[22]. It explores the partial local information by computing consecutive sequence-order patterns and generates 400 (20×20) dimensional vector. DPC is formulated by following equation:

$$A(t) = \frac{n(t)}{C}, \tag{3}$$

where t = 1,2,3,…,0.400, n represents the dipeptide t, and C indicates the total number of possible dipeptides.

Position specific scoring matrix. It has been reported that evolutionary information performs a crucial role in the construction of many predictors[23–26]. Considering this, the evolutionary features are explored by PSSM employing PSI-BLAST tool by aligning each protein sequence of the dataset with homogenous sequences in the NCBI[27]. The following equation is utilized for normalization of each PSSM.

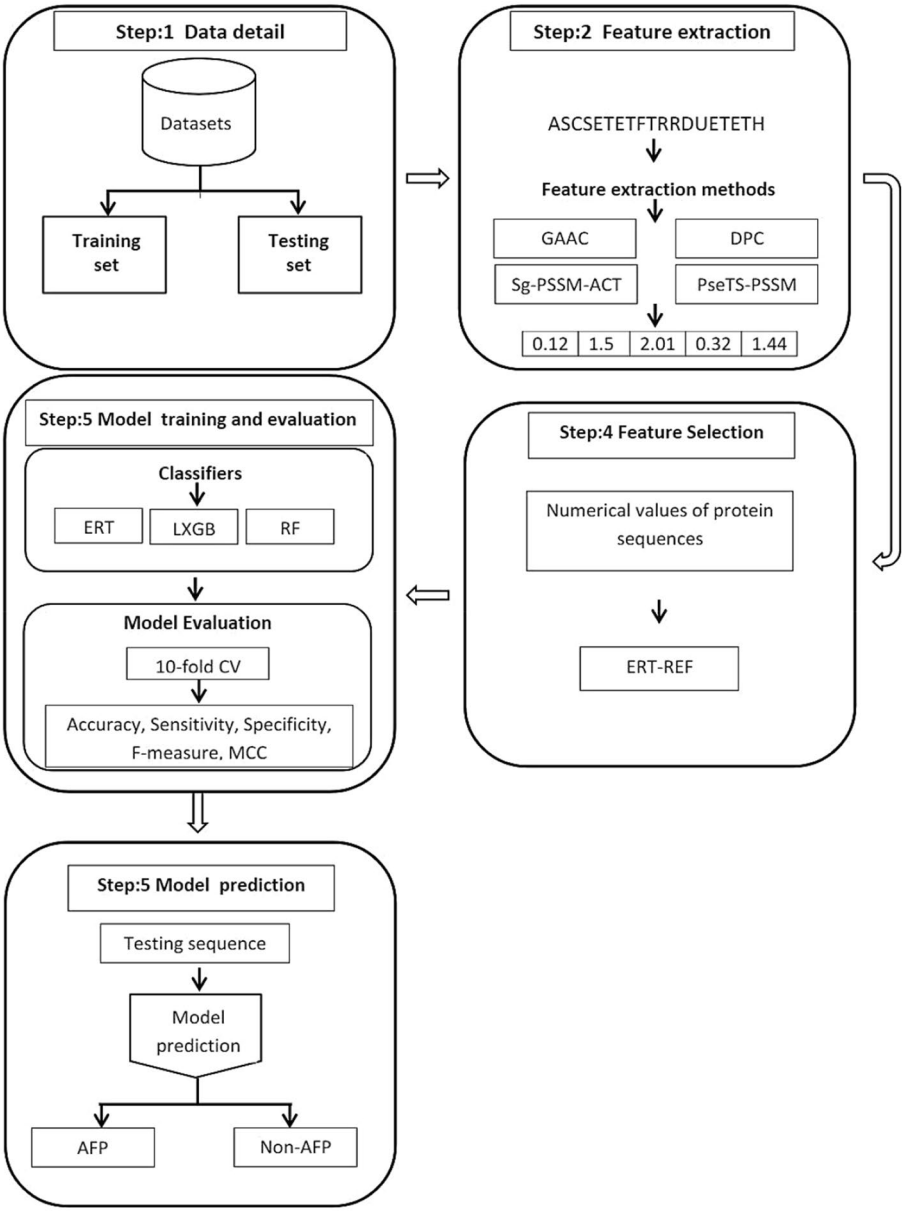

Figure 1. Pipeline of the AFP-LXGB.

$$f(t) = \frac{1}{1 + e^{-t}}, \tag{4}$$

where t represents each element of PSSM.

The PSSM can be denoted as:

$$PSSM = (A_1, A_2, \ldots.., A_j, \ldots.., A_{20})^T, \tag{5}$$

$$A_{m,n} = \left(A_{1,n}, A_{2,n}, \ldots.., A_{L,n}\right), \quad (m = 1, 2, \ldots.., L), \tag{6}$$

where L, T, and $A_{m,n}$ show the length of sequence, transpose operator, and score of the residue in the *mth* position of query sequence replaced with residue of type n, respectively. The dimensional size of PSSM is 20.

Position specific scoring matrix tri-slicing. Recent studies have reported that local regions of PSSM contain more decisive features [26,28,29]. To investigate these features, we incorporated the tri-slicing strategy into PSSM. We split the PSSM into three slices (parts) by row in equivalent dimensions. Each slice (S-PSSM) of the PSSM can be formulated as:first and second rank correlation

$$S-PSSM\left(\acute{\text{Å}}\right)=\begin{bmatrix} P_{b+1,1} & P_{b+1,2} & \ldots. & P_{b+1,20} \\ P_{b+2,1} & P_{b+2,2} & \ldots. & P_{b+2,20} \\ \vdots & \vdots & \vdots & \vdots \\ P_{b+N\left(\acute{\text{Å}}\right),1} & P_{b+N\left(\acute{\text{Å}}\right),2} & \ldots. & P_{b+N\left(\acute{\text{Å}}\right),20} \end{bmatrix}_{N\left(\acute{\text{Å}}\right)\times 20}, \tag{7}$$

where $\acute{\text{Å}}$ indicates the number of $S-PSSM$ and $N\left(\acute{\text{Å}}\right)$ rows in each $S-PSSM$, while $\lfloor * \rfloor$ operator shows the rounding down.

Pseudo position specific scoring matrix tri-slicing. PSSM computes the evolutionary features, however, avoids the correlation factors and sequence order patterns[30,31]. To cope with these limitations of PSSM, we extended Pseudo notion into TS-PSSM. Using Pseudo scheme, we calculated the sequence-order information from each slice and finally combined all three slices to make a super set[32]. The dimension of each slice (S-PSSM(ψ)) can be expressed by following equation:

$$S-PSSM(\psi)=\left[R^{A},R^{C},\ldots.,R^{\psi}\right]_{1\times 20}, \tag{8}$$

where R^{ψ} describes the corresponding residue type of 20 amino acids in a S-PSSM and ψ is the number of slice. Mathematically, TS-PSSM is calculated as:

$$TS-PSSM=[S-PSSM(1),S-PSSM(2),S-PSSM(3)]_{1\times 20}. \tag{9}$$

To calculate the PsePSSM (Pse) from each slice, the following formulation can be utilized:

$$Pse=[\overline{R}_1,\overline{R}_2,\ldots,\overline{R}_{20},\overline{R}_1^{\curlywedge},\overline{R}_2^{\curlywedge},\ldots,\overline{R}_{20}^{\curlywedge}]^{T}, \tag{10}$$

$$\overline{R}_n^{\curlywedge}=\frac{1}{L-\curlywedge}\sum_{m=1}^{L-\curlywedge}[\overline{R}_{m,n}-\overline{R}_{m+\curlywedge,n}]^2 (n=1,2,\ldots,20;\curlywedge<L), \tag{11}$$

where $\overline{R}_n^1$ and $\overline{R}_n^2$ are the first and second rank correlation factors, while $\curlywedge$ describes the correlation factor. TS-PSSM computes 60 features.

Position specific scoring matrix-segmentation-autocorrelation transformation (Sg-PSSM-ACT). Classifiers are unable to directly consider the correlation information of amino acids[33,34]. The correlation information is explored by encoding methods. We applied Sg-PSSM-ACT for consideration of correlation information. In Sg-PSSM-ACT, first, PSSM splits into three segments for extraction of the local region's patterns[35]. Second, autocorrelation transformation (ACT) is extended into each segmentation in order to discover the correlation information regarding the amino acids of evolutionary features[30]. The ACT from the first, second, and third segments are computed by ACT_1, ACT_2, and ACT_3 using the following three equations.

$$ACT_1{}_n^{lag}=\frac{1}{S_1-lag}\sum_{m=1}^{S_1-lag}\left(A_{m,n}-B_n\right)\left(A_{m+lag,n}-B_n\right), n=1,2,\ldots.,20, lag=1,2, \tag{12}$$

$$ACT_2{}_n^{lag}=\frac{1}{S_1-lag}\sum_{m=S_1+1}^{2S_1-lag}\left(A_{m,n}-C_n\right)\left(A_{m+lag,n}-C_n\right), n=1,2,\ldots.,20, lag=1,2, \tag{13}$$

$$ACT_3{}_n^{lag}=\frac{1}{S-S_1-lag}\sum_{m=2S_1+1}^{S-lag}\left(A_{m,n}-D_n\right)\left(A_{m+lag,n}-D_n\right), n=1,2,\ldots.,20, lag=1,2, \tag{14}$$

where B_n, C_n and D_n are the correlation factors between residues and *lag* represents the differences between amino acids. This method 60-dimensional feature vector.

Classification algorithms. To select an appropriate classifier for prediction of AFPs, we have used three classifiers namely RF, ERT, and LXGB. Among these classifiers, Light eXtreme Gradient Boosting (LXGB) has shown the best performance that has been elaborated in the following section.

Light eXtreme gradient boosting. Light GBM is implemented for model training and prediction. Light GBM was first introduced by Microsoft[15]. Compared with GBDTs, Decision Tree, and Random Forest, Light GBM has many advantages such as early stopping, bagging, regularization, multiple loss functions, parallel training, and sparse optimization[16]. Light GBM generates trees using leaf-wise strategy instead of level-wise which leads to a great decrease in loss[17]. The values of hyperparameters are provided in Table 1.

Hyper parameter	Value
Max depth	8
Alpha	1
Era	0.1
Lambda	1
No. of estimator	500

Table 1. Hyper parameters of the model.

Feature selection algorithm. Past research works reported that selection of best features by an effective algorithm enhances the performance of a model[24]. Feature selection (FS) techniques are mostly utilized for solving diverse biological problems in Bioinformatics research field[36–39]. FS removes the less informative and noisy patterns from the original feature set. FS cope with overfitting problem and can boost the model performance[40].

FS techniques are categorized into three classes: wrappers, filters, and embedded approaches[41]. Wrapper methods employ the classifiers to select the best features set. Filters examine the feature via information theoretic and correlation criteria. In embedded techniques, the classifiers first determine the important features by their coefficients and then select the best feature vector[42]. Extremely Randomized Tree-Recursive Feature Elimination (ERT-RFE) is embedded FS algorithm that evaluates the feature using ERT-based model and removes the less informative features recursively. Initially, the input features comprise a subset. In each turn, ERT model is constructed using the subset. The accuracy of the model is calculated and weight of each feature is computed due to its closeness to its target class. Based on weights, features are ranked and low-ranked features are eliminated from subset. When this process is completed, features with maximum accuracy are selected as final optimal feature set.

The feature selected from GAAC, DPC, Pse-PSSM-ACT, and PseTS-PSSM is 5, 88, 35 and 33, respectively. Finally, we attained 161 best feature set.

Assessment methods for model evaluation. After designing a novel method, its efficacy is analyzed by appropriate validation methods[21,22,43–47]. tenfold is mostly used for assessment a model performance[48]. We examined the prediction results by tenfold cross-validation while the generalization power was assessed by independent dataset. Onward, Acc (accuracy), Sn (sensitivity), F-measure, Sp (specificity), and MCC (Mathew's correlation coefficient) are employed as evaluation parameters. These indexes are expressed as:

$$Acc = (TP + TN)/(TP + FP + TN + FN) \tag{15}$$

$$Sn = TP/(TP + FN) \tag{16}$$

$$Sp = TN/(FP + TN) \tag{17}$$

$$MCC = (TN \times TP) - (FN \times FP)/\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)} \tag{18}$$

$$F - measure = 2 * (precision * recall/precision + recall) \tag{19}$$

$$Precision = TP/TP + FP \tag{20}$$

$$Recall = TP/TP + FN \tag{21}$$

Results and discussion

This section illustrates the results of our implemented feature extractors with diverse classifiers. The performance analysis is explained in the upcoming sections.

Results of classifiers using single feature encoder. The results of classifiers using each single feature set are reported in Table 2. RF achieved 69.83%, 76.16%, 86.50%, and 86.18% accuracies on GAAC, DPC, Sg-PSSM-ACT, and PseTS-PSSM, respectively. We can see that DPC features are informative which achieved good results. On Sg-PSSM-ACT, and PseTS-PSSM generated approximately same results however, these are better than GAAC and DPC. ERT yielded similar accuracies with RF using GAAC and DPC while enhanced the performance on Sg-PSSM-ACT and PseTS-PSSM.

The best performance is secured by LXGB on all feature extractors. On GAAC, LXGB attained 1.83% and 1.33% higher accuracies than ERT and RF. Similarly, LXGB improved 7.64% and 7.84% accuracies more than ERT and RF with DPC. Compare with GAAC and DPC, all classifiers significantly boosted the performance over Sg-PSSM-ACT and attained accuracies of 86.50%, 87.83%, and 88.02% by LXGB, ERT, and RF, respectively. The better results of Sg-PSSM-ACT are due to several reasons such as PSSM explores the evolutionary profile, ACT considers correlation factors, and Sg computes the local patterns. The best performance is achieved by LXGB,

Classifier	Feature descriptor	Acc (%)	Sn (%)	Sp (%)	F-measure (%)	MCC
RF	GAAC	69.83	68.34	71.32	68.97	0.40
	DPC	76.16	67.31	84.99	72.96	0.53
	Sg-PSSM-ACT	86.50	82.66	90.33	93.70	0.72
	PseTS-PSSM	86.18	79.40	96.13	84.66	0.73
ERT	GAAC	69.33	63.35	74.97	67.41	0.38
	DPC	76.36	65.66	86.96	73.35	0.54
	Sg-PSSM-ACT	87.83	81.66	93.99	86.83	0.76
	PseTS-PSSM	90.50	91.03	89.98	90.55	0.81
LXGB	GAAC	71.16	71.33	71.01	71.03	0.43
	DPC	84.00	83.33	84.67	83.67	0.68
	Sg-PSSM-ACT	88.02	85.32	89.98	87.59	0.76
	PseTS-PSSM	92.50	90.01	95.00	92.23	0.84

Table 2. Results based on single feature descriptor.

Classifier	Feature descriptor	Acc (%)	Sn (%)	Sp (%)	F-measure (%)	MCC
RF	GAAC + DPC	78.83	73.66	84.00	77.49	0.57
	DPC + Sg-PSSM-ACT	83.49	76.67	90.31	81.87	0.67
	DPC + Sg-PSSM-ACT + PseTS-PSSM	90.33	85.32	95.43	89.65	0.81
	All feature set	92.50	88.00	97.00	92.14	0.85
ERT	GAAC + DPC	78.50	70.35	86.67	76.30	0.57
	DPC + Sg-PSSM-ACT	82.15	75.62	88.59	80.89	0.65
	DPC + Sg-PSSM-ACT + PseTS-PSSM	89.00	82.33	95.61	88.12	0.78
	All feature set	90.67	84.02	96.98	90.01	0.82
LXGB	GAAC + DPC	86.03	85.68	86.27	85.87	0.72
	DPC + Sg-PSSM-ACT	91.31	88.28	93.95	90.91	0.82
	DPC + Sg-PSSM-ACT + PseTS-PSSM	92.17	90.67	93.54	92.01	0.84
	All feature set	93.67	92.64	94.51	93.45	0.87

Table 3. Results based on hybrid features.

ERT, and RF employing PseTS-PSSM on all five evaluation indexes. LXGB, ERT, and RF generated 92.50%, 90.50%, and 86.18% accuracies which are the highest outcomes among all feature encoding approaches.

Performance of classifiers with heterogeneous features. Past studies have revealed that a combination of different features enriched the prediction models[36,49]. In this connection, we ensemble the features of different descriptors in various series combinations and summarized the results in Table 3. The accuracies showed by GAAC + DPC with RF, ERT, and LXGB are 78.83%, 78.50%, and 86.03%, respectively. We noted that integrated feature set achieved better prediction for AFP. Similarly, RF, ERT, and LXBG further boosted the performance with DPC + Sg-PSSM-ACT, which are 83.49%, 82.15%, and 91.31% in terms of accuracy. Onward, we analyzed the prediction results of classifiers over DPC + Sg-PSSM-ACT + PseTS-PSSM and "All feature set". It is observed from Table 2 that all classifiers using "All feature set" attained remarkable performance with all assessment indexes. The accuracies secured by RF, ERT, and LXGB are 92.50%, 90.67%, and 93.67%, respectively.

Among all classifiers, LXGB obtained the highest results on the training dataset using tenfold. LXGB improved the Acc, Sn, F-measure, and MCC by 1.17%, 4.6%, 1.31%, and 0.02 than the second best classifier (i.e., RF) with the same feature encoder (i.e., All feature set). From all analyses, we can conclude that fused feature set of All feature set greatly contributed to the identification of AFPs.

Results analysis of classifiers on the best feature set. Best features selection is a key step in the design of a predictor[50]. Many researchers applied the feature selection techniques and boosted the predictor performance[24,51,52]. During the process of feature selection, discriminative features are selected that can significantly boost the model performance. This study uses ERT-RFE technique for selecting the optimized features. From Table 4, we can see that RF with the best feature set produced an accuracy of 90.67%, sensitivity of 85.33%, specificity of 96.12%, and MCC of 0.81. Similarly, ERT reduced performance than RF and secured an accuracy of 90.00%. The LXGB shows outstanding performance over the best feature set and attained 94.00% accuracy, 93.00% sensitivity, 95.00% specificity, and 0.88 MCC. LXGB improved 3.33% accuracy than RF and 4% higher accuracy than ERT-based model. The results reveal that the best features effectively explore the local region features and sequence order information. Moreover, LXGB showed better performance mostly with indi-

Predictor	Acc (%)	Sn (%)	Sp (%)	MCC
RF	90.67	85.33	96.12	0.81
ERT	90.00	83.01	98.99	0.80
LXGB	94.00	93.00	95.00	0.88

Table 4. Results based on the best feature set.

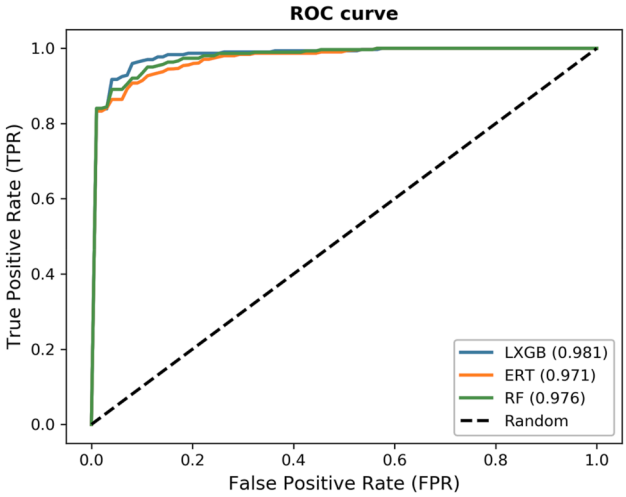

Figure 2. ROC curves of the classifiers.

Ratio of AFP:Non-AFP	Acc (%)	Sn (%)	Sp (%)	MCC	AuROC	AuPR
1:1	94.00	93.00	95.00	0.88	0.9820	0.9883
1:2	92.46	87.76	94.43	0.84	0.9631	0.9740
1:3	91.33	85.91	93.89	0.83	0.9565	0.9591

Table 5. Results with different ratios of AFP and non-AFP.

vidual, hybrid, and optimized feature sets than other classification algorithms. The ROC curves of classifiers are depicted in Fig. 2.

Ablation study using imbalanced dataset. We performed an ablation study to check the effectiveness of the proposed study using imbalanced dataset. We distributed the dataset into 1:1 (300:300), 1:2 (300:600), 1:3 (300:900) ratios of AFP and non-AFP and the performance of the model with each ratio is observed. The results of the proposed work with different ratios of AFP and non-AFP are reported in Table 5. The performance of the model with 1:1 ratio is promising and found the best prediction results. Increasing the samples of the non-AFP i.e., using 1:2 ratio, the model reduced the points of accuracy, specificity, MCC, AuROC, and AuPR specifically sensitivity. The imbalanced samples of both classes show that it will not only overall performance of the model but greatly affect the sensitivity. Onward, analyzing the performance of the model using 1:3 (300:900), the model further decreased performance on all evaluation parameters. These results illustrate that on balanced dataset a model can perform better and produce effective results.

The second ablation study is performed by applying a feature selection/reduction approach named Extremely Randomized Tree-Recursive Feature Elimination (ERT-RFE) to individual feature vector of GAAC, DPC, Pse-PSSM-ACT, and PseTS-PSSM. The features selected from GAAC, DPC, Pse-PSSM-ACT, and PseTS-PSSM are 5, 88, 35 and 33, respectively. All classifiers on each feature vector improved the performance. For instance, the accuracy of RF with GAAC before feature reduction is 69.83% and after applying the ERT-RFE is 70.11%. ERT and LXGB also enhanced the results on GAAC. Similarly, with reduced feature vector of DPC, Sg-PSSM-ACT, and PseTS-PSSM, all classifiers raised the accuracies.

Past studies have revealed that hybrid features enrich the predictor performance. In this connection, we ensemble the features of GAAC, DPC, Pse-PSSM-ACT, and PseTS-PSSM descriptors and make one super set of 161-dimension. The results are recorded in Table 4. The accuracy obtained by RF on "reduced all feature set" is 92.67%, while it is 92.50% accuracy before applying feature reduction. In the same manner, ERT and LXGB have

Predictor	Acc (%)	Sn (%)	Sp (%)	MCC
AFP-Pred	83.38	84.67	82.32	0.66
AFP-PseAAC	89.69	88.89	91.00	0.80
CryoProtect	89.50	89.54	89.50	0.79
AFP-LSE	90.30	86.70	93.90	0.80
PoGB-pred	89.38	73.17	90.01	0.37
AFP-LXGB	94.00	93.00	95.00	0.88

Table 6. Comparison with existing predictors on the training set.

Predictor	Acc (%)	Sn (%)	Sp (%)	MCC
AFP-Pred	77.34	91.16	77.04	0.23
AFP-PseAAC	84.75	85.08	84.74	0.27
CryoProtect	88.28	87.27	88.30	0.31
AFP-SRC	85.40	86.10	84.70	0.28
AFP- LXGB	92.37	79.56	92.63	0.35

Table 7. Comparison with existing predictors on the testing set.

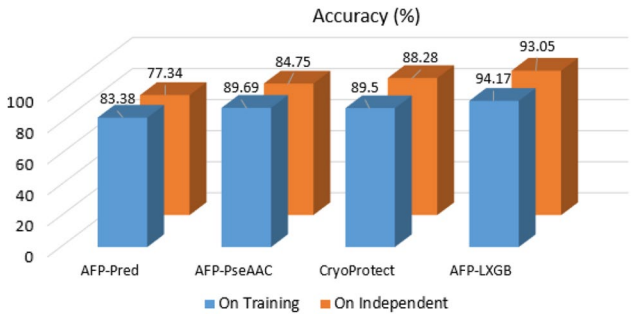

Figure 3. Accuracy comparison with the existing predictor.

also improved the results. It is concluded that reducing the feature vectors by an appropriate feature selection approach and then combining it all specifically raised the performance of a model.

Comparative analysis with past work. We performed the comparison of the proposed system (AFP-LXGB) with the existing predictors like CryoProtect[12], AFP-PseAAC[10], AFP-Pred[7], AFP-LSE[14], and PoGB-pred[16] on both training and testing datasets and summarized the outcomes in Tables 6 and 7. From Table 6, it is observed that our predictor yielded an accuracy of 94.00%, sensitivity of 93.00%, specificity of 95.00%, and MCC of 0.88, which are 3.70%, 6.30%, 1.1%, and 0.08 higher than the best method i.e., AFP-LSE. The proposed protocol also boosted the Acc, Sn, Sp, and MCC by 4.48%, 4.15%, 3.96%, and 0.08 are higher than the second-best method i.e., AFP-PseAAC. In the same fashion, our predictor surpassed other previous approaches on all four evaluation indexes. The efficacy of a novel model can be assessed by its high generalization ability. In this connection, we carried out the experiments on the independent dataset and it is noted in Table 7 that AFP-LXGB outperformed the previous methods in the literature.

On the testing dataset, AFP-LXGB achieved 4.09% (Acc), 4.33% (Sp), and 0.04 (MCC) higher than Cryo-Protect. Similarly, this work also enhanced the Acc, Sp, and MCC by 7.62%, 7.89%, and 0.08, respectively than second best predictor (AFP-PseAAC). The comparison has also been indicated in Fig. 3.

Conclusion

In the current study, we established a novel predictor, called AFP-LXGB for antifreeze protein identification. It is a challenging job to explore the discriminative features of diverse and complex nature of AFP. To cope with this issue, we discovered the dominant information by PseTS-PSSM, Sg-PSSM-ACT, GAAC, and DPC. Further, we concatenated these feature vectors and applied ERT-RFE feature selection approach. The models are trained models with RF, ERT, and LXGB. After analyzing the performance of all models, it is concluded that AFP-LXGB has shown the best performance compared with the previous. The supreme achievement of the current study is due to several reasons such as effective feature coding approaches and appropriate classification algorithms.

In the future, we will apply more effective feature descriptors, feature selection approaches, and classifiers to further improve the performance of a predictor.

Data availability

The datasets used in this study has provided in the supplementary file and codes are provided at the link https://github.com/Farman335/AFP-LXGB.

Received: 16 May 2022; Accepted: 16 November 2022
Published online: 30 November 2022

References

1. Kim, S.-K. *Marine Proteins and Peptides: Biological Activities and Applications* (Wiley, 2013).
2. Griffith, M. *et al.* Antifreeze proteins in winter rye. *Physiol. Plant.* **100**, 327–332 (1997).
3. Davies, P. L. & Hew, C. L. Biochemistry of fish antifreeze proteins. *FASEB J.* **4**, 2460–2468 (1990).
4. Feeney, R. E. & Yeh, Y. Antifreeze proteins: Current status and possible food uses. *Trends Food Sci. Technol.* **9**, 102–106 (1998).
5. Breton, G., Danyluk, J., ois Ouellet, F. & Sarhan, F. Biotechnological applications of plant freezing associated proteins. *Biotechnol. Annu. Rev.* **6**, 59–101 (2000).
6. Urrutia, M. E., Duman, J. G. & Knight, C. A. Plant thermal hysteresis proteins. *Biochimica et Biophysica Acta (BBA) Protein Struct. Mol. Enzymol.* **1121**, 199–206 (1992).
7. Kandaswamy, K. K. *et al.* AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. *J. Theor. Biol.* **270**, 56–62 (2011).
8. Zhao, X., Ma, Z. & Yin, M. Using support vector machine and evolutionary profiles to predict antifreeze protein sequences. *Int. J. Mol. Sci.* **13**, 2196–2207 (2012).
9. Yu, C.-S. & Lu, C.-H. Identification of antifreeze proteins and their functional residues by support vector machine and genetic algorithms based on n-peptide compositions. *PLoS ONE* **6**, e20445 (2011).
10. Mondal, S. & Pai, P. P. Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. *J. Theor. Biol.* **356**, 30–35 (2014).
11. He, X. *et al.* TargetFreeze: Identifying antifreeze proteins via a combination of weights using sequence evolutionary information and pseudo amino acid composition. *J. Membr. Biol.* **248**, 1005–1014 (2015).
12. Pratiwi, R. *et al.* CryoProtect: A web server for classifying antifreeze proteins from nonantifreeze proteins. *J. Chem.* **2017**, 1–15 (2017).
13. Khan, S., Naseem, I., Togneri, R. & Bennamoun, M. Rafp-pred: Robust prediction of antifreeze proteins using localized analysis of n-peptide compositions. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **15**, 244–250 (2016).
14. Usman, M., Khan, S. & Lee, J.-A. Afp-lse: Antifreeze proteins prediction using latent space encoding of composition of k-spaced amino acid pairs. *Sci. Rep.* **10**, 1–13 (2020).
15. Usman, M., Khan, S., Park, S. & Wahab, A. AFP-SRC: Identification of antifreeze proteins using sparse representation classifier. *Neural Comput. Appl.* https://doi.org/10.1007/s00521-021-06558-7 (2021).
16. Alim, A., Rafay, A. & Naseem, I. PoGB-pred: Prediction of antifreeze proteins sequences using amino acid composition with feature selection followed by a sequential-based ensemble approach. *Curr. Bioinform.* **16**, 446–456 (2021).
17. Miyata, R., Moriwaki, Y., Terada, T. & Shimizu, K. Prediction and analysis of antifreeze proteins. *Heliyon* **7**, e07953 (2021).
18. Sonnhammer, E. L., Eddy, S. R. & Durbin, R. Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405–420 (1997).
19. Li, W. & Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
20. Barukab, O., Ali, F. & Khan, S. A. DBP-GAPred: An intelligent method for prediction of DNA-binding proteins types by enhanced evolutionary profile features with ensemble learning. *J. Bioinform. Comput. Biol.* **19**, 2150018 (2021).
21. Ali, F. & Hayat, M. Classification of membrane protein types using voting feature interval in combination with Chou's Pseudo amino acid composition. *J. Theor. Biol.* **384**, 78–83 (2015).
22. Ali, F. & Hayat, M. Machine learning approaches for discrimination of extracellular matrix proteins using hybrid feature space. *J. Theor. Biol.* **403**, 30–37 (2016).
23. Ali, F., Ahmed, S., Swati, Z. N. K. & Akbar, S. DP-BINDER: Machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information. *J. Comput.-Aided Mol. Des.* **33**, 645–658 (2019).
24. Ali, F. *et al.* DBPPred-PDSD: Machine learning approach for prediction of DNA-binding proteins using discrete wavelet transform and optimized integrated features space. *Chemom. Intell. Lab. Syst.* **182**, 21–30 (2018).
25. Kabir, M. *et al.* Prediction of membrane protein types by exploring local discriminative information from evolutionary profiles. *Anal. Biochem.* **564**, 123–132 (2019).
26. Ali, F. *et al.* SDBP-Pred: Prediction of single-stranded and double-stranded DNA-binding proteins by extending consensus sequence and K-segmentation strategies into PSSM. *Anal. Biochem.* **589**, 113494 (2020).
27. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
28. Dehzangi, A., Paliwal, K., Lyons, J., Sharma, A. & Sattar, A. Exploring potential discriminatory information embedded in pssm to enhance protein structural class prediction accuracy. In *IAPR International Conference on Pattern Recognition in Bioinformatics* 208–219 (Springer Berlin Heidelberg, 2013).
29. Shen, C., Ding, Y., Tang, J., Song, J. & Guo, F. Identification of DNA–protein binding sites through multi-scale local average blocks on sequence information. *Molecules* **22**, 2079 (2017).
30. Akbar, S., Hayat, M., Kabir, M. & Iqbal, M. iAFP-gap-SMOTE: An efficient feature extraction scheme gapped dipeptide composition is coupled with an oversampling technique for identification of antifreeze proteins. *Lett. Org. Chem.* **16**, 294–302 (2019).
31. Akbar, S. *et al.* iHBP-DeepPSSM: Identifying hormone binding proteins using PsePSSM based evolutionary features and deep learning approach. *Chemom. Intell. Lab. Syst.* **204**, 104103 (2020).
32. Akbar, S. *et al.* iAtbP-Hyb-EnC: Prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model. *Comput. Biol. Med.* **137**, 104778 (2021).
33. Ahmad, A. Akbar, S. Hayat, M. Ali, F. & Sohail, M. Identification of antioxidant proteins using a discriminative intelligent model of k-space amino acid pairs based descriptors incorporating with ensemble feature selection. *Biocybern. Biomed. Eng.* **42**, 727–735 (2020).
34. Ahmad, A. *et al.* Deep-AntiFP: Prediction of antifungal peptides using distanct multi-informative features incorporating with deep neural networks. *Chemom. Intell. Lab. Syst.* **208**, 104214 (2021).
35. Barukab, O., Ali, F., Alghamdi, W., Bassam, Y. & Khan, S. A. DBP-CNN: Deep learning-based prediction of DNA-binding proteins by coupling discrete cosine transform with two-dimensional convolutional neural network. *Expert Syst. Appl.* **197**, 116729 (2022).

36. Ali, F., Ahmed, S., Swati, Z. N. K. & Akbar, S. DP-BINDER: Machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information. *J. Comput. Aided Mol. Des.* **33**, 645–658 (2019).
37. Ahmad, A., Akbar, S., Tahir, M., Hayat, M. & Ali, F. iAFPs-EnC-GA: Identifying antifungal peptides using sequential and evolutionary descriptors based multi-information fusion and ensemble learning approach. *Chemom. Intell. Lab. Syst.* **222**, 104516 (2022).
38. Ali, F. *et al.* Deep-PCL: A deep learning model for prediction of cancerlectins and non cancerlectins using optimized integrated features. *Chemom. Intell. Lab. Syst.* **221**, 104484 (2022).
39. Sikander, R., Ghulam, A. & Ali, F. XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set. *Sci. Rep.* **12**, 1–9 (2022).
40. Ali, F. *et al.* Target-DBPPred: An intelligent model for prediction of DNA-binding proteins using discrete wavelet transform based compression and light eXtreme gradient boosting. *Comput. Biol. Med.* **145**, 105533 (2022).
41. Ali, F. *et al.* Deep-GHBP: Improving prediction of growth hormone-binding proteins using deep learning model. *Biomed. Signal Process. Control* **78**, 103856 (2022).
42. Yan, K. & Zhang, D. Feature selection and analysis on correlated gas sensor data with recursive feature elimination. *Sens. Actuators B Chem.* **212**, 353–363 (2015).
43. Khan, Z. U., Ali, F., Khan, I. A., Hussain, Y. & Pi, D. iRSpot-SPI: Deep learning-based recombination spots prediction by incorporating secondary sequence information coupled with physio-chemical properties via Chou's 5-step rule and pseudo components. *Chemom. Intell. Lab. Syst.* **189**, 169–180 (2019).
44. Swati, Z. N. K. *et al.* Brain tumor classification for MR images using transfer learning and fine-tuning. *Comput. Med. Imaging Graph.* **75**, 34–46 (2019).
45. Khan, Z. U., Ali, F., Ahmad, I., Hayat, M. & Pi, D. iPredCNC: Computational prediction model for cancerlectins and non-cancerlectins using novel cascade features subset selection. *Chemom. Intell. Lab. Syst.* **195**, 103876 (2019).
46. Arif, M. *et al.* TargetCPP: Accurate prediction of cell-penetrating peptides from optimized multi-scale features using gradient boost decision tree. *J. Comput.-Aided Mol. Des.* **34**(8), 841–856 (2020).
47. Ahmed, S. *et al.* An integrated feature selection algorithm for cancer classification using gene expression data. *Comb. Chem. High Throughput Screen.* **21**, 631–645 (2018).
48. Ullah, M., Iltaf, A., Hou, Q., Ali, F. & Liu, C. A foreground extraction approach using convolutional neural network with graph cut. In *2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)* 40–44 (IEEE, 2018).
49. Ali, F. *et al.* AFP-CMBPred: Computational identification of antifreeze proteins by extending consensus sequences into multi-blocks evolutionary information. *Comput. Biol. Med.* **139**, 105006 (2021).
50. Ghulam, A. *et al.* Accurate prediction of immunoglobulin proteins using machine learning model. *Inform. Med. Unlocked* **29**, 100885 (2022).
51. Khan, Z. U. *et al.* piEnPred: A bi-layered discriminative model for enhancers and their subtypes via novel cascade multi-level subset feature selection algorithm. *Front. Comp. Sci.* **15**, 1–11 (2021).
52. Ghulam, A. *et al.* ACP-2DCNN: Deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network. *Chemom. Intell. Lab. Syst.* **226**, 104589 (2022).

Author contributions

A.K. performed the experiments, J.U. wrote the manuscript, F.A. reviewed the manuscript, A.A.: Performed new experiments as suggested reviewers. O.A.: Performed new experiments as suggested reviewers. A.B.: Summarized new experiments results. A.D.: Incorporated the new text in the manuscript, contributed in experimental work, and revised the whole manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-24501-1.

Correspondence and requests for materials should be addressed to F.A. or A.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.