Article

iMRSAPred: Improved Prediction of Anti-MRSA Peptides Using Physicochemical and Pairwise Contact-Energy Properties of Amino Acids

Muhammad Arif, Ge Fang, Huma Fida, Saleh Musleh, Dong-Jun Yu, and Tanvir Alam*

Cite This: ACS Omega 2024, 9, 2874−2883

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Methicillin-resistant *Staphylococcus aureus* (MRSA) is a growing concern for human lives worldwide. Anti-MRSA peptides act as potential antibiotic agents and play significant role to combat MRSA infection. Traditional laboratory-based methods for annotating Anti-MRSA peptides are although precise but quite challenging, costly, and time-consuming. Therefore, computational methods capable of identifying Anti-MRSA peptides accelerate the drug designing process for treating bacterial infections. In this study, we developed a novel sequence-based predictor "iM-RSAPred" for screening Anti-MRSA peptides by incorporating energy estimation and physiochemical and sequential information. We successfully resolved the skewed imbalance phenomena by using synthetic minority oversampling technique plus Tomek link (SMOTETomek) algorithm. Furthermore, the Shapley additive explanation method was leveraged to analyze the impact of top-ranked features in the prediction task. We evaluated multiple machine learning algorithms, i.e., CatBoost, Cascade Deep Forest, Kernel and Tree Boosting, support vector machine, and HistGBoost classifiers by 10-fold cross-validation and independent testing. The proposed iMRSAPred method significantly improved the overall performance in terms of accuracy and Matthew's correlation coefficient (MCC) by 5.45 and 0.083%, respectively, on the training data set. On the independent data set, iMRSAPred improved accuracy and MCC by 3.98 and 0.055%, respectively. We believe that the proposed method would be useful in large-scale Anti-MRSA peptide prediction and provide insights into other bioactive peptides.
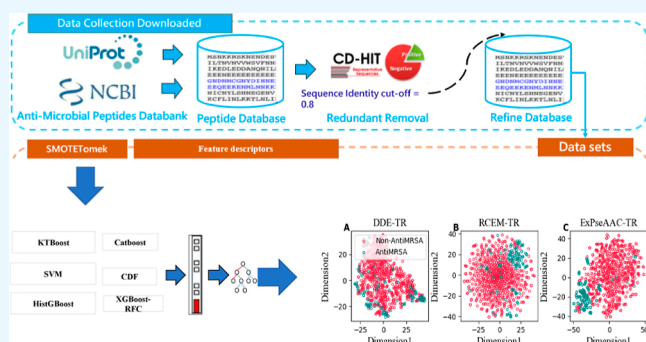
INTRODUCTION

The challenge of antibiotic resistance continues to pose a significant health threat on a global scale, prompting the World Health Organization (WHO) to call upon various research domains to address this complex issue. One of the most hazardous pathogens is Methicillin-resistant *Staphylococcus aureus* (MRSA), killing thousands of peoples both in the developed and developing countries every year.[1,2] These infections are fatal in numerous conditions, including bacteremia (15−60%) and staphylococcal pneumonia (30−40%).[3] The current clinically approved treatment for MRSA infection includes the use of antibiotics such as teicoplanin, vancomycin,[4,5] etc. Nevertheless, the effectiveness of these antibiotics on patients may be compromised due to the emergence of drug resistance in anti-MRSA medications. Thus, options to using other antibiotic drugs to treat MRSA are desperately needed.[6] Due to the continuing antibiotic resistance, antimicrobial peptides have gained attention as potential therapeutic options with quick and broad-spectrum antibacterial activity, including antibiotic-resistant germs such as MRSA.[7] Thus, owing to the biological applications as therapeutic agent to combat bacterial infections, the identification of Anti-MRSA peptides is crucial in developing new weapons as antibiotic drugs.

Over the past years, laboratory-based methods, i.e., mass spectrometry, fluorescence-based, microdilution-based, and rational design method, etc. have been devoted to screening and analyzing Anti-MRSA peptides. However, these bioassays are formidable, costly, and time-consuming, particularly for analyzing a large number of Anti-MRSA peptides. Therefore, computational methods more specifically machine learning (ML)- and deep learning-based methods are used for more accurate prediction of the Anti-MRSA peptide.
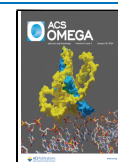
To the best of our knowledge, SCMRSA is the only ML-based predictor available in the literature for classifying Anti-MRSA and non-Anti-MRSA peptides.[8] This tool used scoring card methods with optimized dipeptide composition and
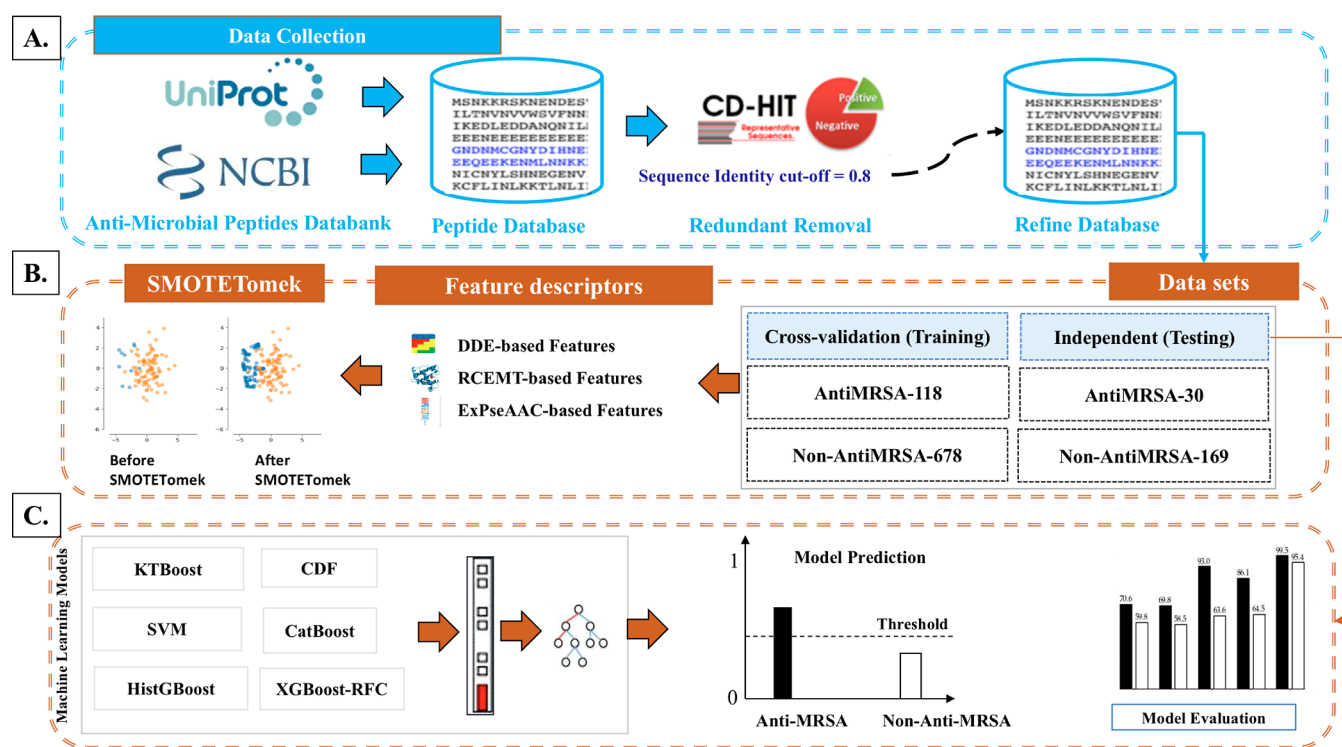
2874

Figure 1. Workflow of the proposed iMRSAPred method. (A) Collection of data set and refinement of sequences, (B) feature extraction from the data set and handling imbalance data distribution using SMOTETomek, and (C) ML model development and evaluation.

amino acid (AA) properties to achieve 92.70% accuracy. We believe that this level of accuracy can be improved as the data imbalance problem hampered the prediction results of SCMRSA with a high error rate. Second, the energy estimation and biochemical properties contained in Anti-MRSA peptides were not considered. The aforementioned problems motivated us to construct a novel method iMRSAPred for characterizing and predicting Anti-MRSA peptides with higher accuracy. We extracted the energy estimation-, sequential- and physicochemical-based properties of AAs by considering pairwise residue contact-energy matrix transformation (RCEMT),[9] dipeptide deviation from expected mean (DDE) and extended form of pseudoamino-acid composition (ExPseAAC), respectively. The imbalanced data set issue in the training data set was tackled by using synthetic minority oversampling technique plus Tomek link (SMOTETomek) algorithm.[10] We deployed several state-of-the-art ML classifiers such as Cascade Deep Forest (CDF), combined Kernel and Tree Boosting (KTBoost), CatBoost, histogram-gradient Boost (HistG-Boost), support vector machine (SVM), and combined extreme gradient-boost random forest (XGBoost-RFC). Among these classifiers, CDF and CatBoost achieved the best results using proposed features both on 10-fold cross-validation (CV) and independent testing (IND). The schematic workflow of the proposed iMRSAPred method is illustrated in Figure 1.

In short, the contribution of our work can be summarized as follows:

(a) We captured the physicochemical-based and interaction energy estimation-based local and global properties of AAs from given peptide sequence using ExPseAAC, RCEMT, and DDE descriptors.

(b) We employed the SMOTETomek algorithm as an effective solution to overcome the challenges of imbalanced data sets in this particular problem.

(c) We proposed CatBoost and CDF as the best classifiers for predicting Anti-MRSA peptides with outstanding performance both on training and testing data sets obtaining improved accuracy compared to existing state-of-the-art tool for the same purpose.

(d) We investigated relative importance of the proposed features using Shapley additive explanation (SHAP) and t-SNE algorithms. This provides insights on the impact of features as well as the interpretability of the proposed model.

■ MATERIALS AND METHODS

Benchmark Data Sets. The collection of valid data set is the key to developing an efficient computational model.[11−13] For this purpose, we considered the same benchmark data set in the paper[8] for fair comparison. The benchmark data set contains experimentally verified peptides (including 444 Anti-MRSA and 9898 Non-Anti-MRSA), which were originally retrieved from antimicrobial peptide database (APD3).[14] The collected peptide sequences were split into two subsets at an 8:2 ratio for training (CV) and evaluation (independent) of the proposed iMRSAPred method. We provided both training and testing data set instances in Table 1.

Feature Encoding Schemes. Feature encoding schemes are challenging task used to convert a biological sequence into fixed length numerical feature.[15] In this research, the energy estimation, sequence, and physiochemical-based properties were considered for encoding Anti-MARSA peptides. The details of each feature descriptor are explained below.

Extraction from Pairwise Contact Energy Matrix. The pairwise energy-derived properties of AAs provide deep

Table 1. Dataset Summary

data set	total sequence	(Pos, Neg)[a]
AMRSA$_{train}$	796	(118, 768)
AMRSA$_{test}$	199	(30,169)

[a]Neg and Pos represent the total number of Non-Anti-MRSA and Anti-MRSA peptides, respectively.

insights to understand the peptide structure and function.[16] Peptides' structural stability relies on extensive interactions among internal residues.[9] These interactions can be estimated using an energy function, typically derived from known structures, to assess the energy contribution of these residue interactions.[17] However, in the case of peptides, or unstructured proteins with unknown conformation, the energy function is unable to calculate the cumulative energies due to the lack of the defined structure. As a result, this energy function is not applicable to peptides, or unstructured proteins lacking a specific structural arrangement.[18] Motivated by this, we utilized the derived predicted energy estimation-based properties, i.e., pairwise contact-energy matrix (RCEM)[9] provided in Table S2, to extract significant information that is inherently associated with interactions among AA residues and intrinsically disordered regions. The RCEM is a matrix with dimensions of 20 × 20, where each row and column corresponds to one of the 20 standard AAs.[9] The RCEMT can be represented in the matrix form as

$$RCEMT = \begin{bmatrix} s_{1,1} & s_{1,1} & s_{1,1} \cdots & s_{1,20} \\ s_{2,1} & s_{2,2} & s_{2,3} \cdots & s_{2,20} \\ s_{L,1} & s_{L,2} & s_{L,3} \cdots & s_{L,20} \end{bmatrix} \quad (1)$$

where, within each group, the sum of the RCEM values in each column is calculated and forms 400 dimension features. The readers are referred for the further details to a study by Mishra et al.[19]

Extended Pseudo Amino Acid Composition (ExPseAAC). TExPseAAC is widely used feature encoding descriptor proposed by Chou and Cai,[20] for formulating biological proteins/peptide sequences. Unlike, the simple alignment-free amino-acid composition method ExPseAAC considers both the compositional and correlation physicochemical characteristics of peptides.[21] Motivated from our previous study, we extended the concept of PseAAC for encoding Anti-MRSA peptides by using new biochemical properties of AAs, namely, irreplaceability, hydrophobicity, rigidity, hydrophilicity, and flexibility.[21] We listed the values of these physicochemical properties for 20 AAs in Table S1. A peptide sequence is represented as an array of short length (5−30) AA residues typically denoted as

$$A_1, A_2, A_3, A_4, A_5, ..., A_L \quad (2)$$

The correlation factors can be defined as

$$\begin{cases} \delta_1 = \frac{1}{L-1} \sum_{i=1}^{L-1} \delta(A_i, A_{i+1}), \\ \delta_2 = \frac{1}{L-2} \sum_{i=2}^{L-2} \delta(A_i, A_{i+2}), \\ \delta_3 = \frac{1}{L-3} \sum_{L-3}^{L-3} \delta(A_i, A_{i+3}), L - \lambda \\ \vdots \\ \delta_\lambda = \frac{1}{L-\lambda} \sum_{i=\lambda}^{L-\lambda} \delta(A_i, A_{i+\lambda}) \end{cases} \quad (3)$$

In eq 3, δ_1 corresponds to the first-rank of correlation factor and represents the consecutive AAs sequence order information, δ_2 corresponds to the second-rank factor and represents the second-order correlation of the entire second consecutive AAs, and so forth. Consequently, we can define the correlation factor as

$$\delta(A_i, A_j) = \left\{ \frac{1}{2}[(H_1(A_j) - H_1(A_i))^2 + (H_2(A_j) - H_2(A_i))^2] \right\} \quad (4)$$

where $H_1(A_i)$ and $H_2(A_i)$ present the derived biochemical value of AAs A_i.

$$\begin{cases} H_1(i) = \frac{H_1^i - \sum_{i=1}^{20} H_1^i / 20}{\sqrt{\frac{\sum_{i=1}^{20}(H_1^i - \sum_{i=1}^{20} H_1^i / 20)^2}{20}}}, \\ H_2(i) = \frac{H_2^i - \sum_{i=1}^{20} H_2^i / 20}{\sqrt{\frac{\sum_{i=1}^{20}(H_2^i - \sum_{i=1}^{20} H_2^i / 20)^2}{20}}} \end{cases} \quad (5)$$

Given an index j, the primary AA residues of the peptide can be formulated into a $P_{20+\lambda}$ feature space

$$P^i = [P_1, P_1, P_1, ..., P_{20}, P_{20+1}, ... P_{20+\lambda}] \quad (6)$$

where

$$P_u = \begin{cases} \frac{f_u}{\sum_{m=1}^{20} f_m + \sum_{i=1}^{\lambda} \delta_i^j}, & (1 \leq u \leq 20) \\ \frac{\delta_{u-20}^j}{\sum_{m=1}^{20} f_m + \sum_{i=1}^{\lambda} \delta_i^j}, & (20+1 \leq u \leq 20+\lambda) \end{cases} \quad (7)$$

where f_m denotes the frequency of 20 AAs in peptide and δ_i^j is the i-tier sequence correlation factor. The first 20 elements denote the effect of the AAC, and the elements from $20+1$ to $20+\lambda$ denote the effect of sequence order.

Generally, ExPseAAC can be formulated as

$$ExPseAAC = [f_1, ..., f_{20}, f_{20+1}, ..., f_{20+\lambda}]^T, (\lambda < N) \quad (8)$$

where the first 20 attributes denote the frequency information on 20 natural AAs in the peptide sequence and the 21st feature vector, i.e., f_{20+1} denotes the additional correlation factor related to first tier sequence, the 22nd factor to the second tier,

and so on.[22] In this study, after experimental analysis, we kept the value for encoding Anti-MRSA peptides. Thus, the resultant feature space is $(20 + 2 \times 5 = 30)$ dimensions.

Dipeptide Deviation from Expected Mean. The DDE is an effective protein feature representation method proposed by Saravanan et al.,[23] for linear B-Cell Epitope prediction. DDE considers the consecutive pairs (local sequence information) of AA in peptides and generates 400-dimension feature vector. These dipeptides have an associated properties that influence the protein's function and structure. The working principle of the DDE descriptor relies on three parameters: DPC, theoretical mean (T_m), and theoretical variance (T_v).[24] To compute DPC, the following mathematical expression can be used

$$DPC(a, b) = \frac{M_{ab}}{L - 1}, \; ab \in \{AA, AC, AD, ... YY\} \tag{9}$$

where M_{ab} is the number of dipeptides denoted by AA types a and b and L is the length of peptide sequence. $T_m(a, b)$, the theoretical mean, is formulated as follows

$$T_m(a, b)) = \frac{C_a}{C_L} \times \frac{C_b}{C_L} \tag{10}$$

where in the given peptide dipeptide "ab", C_a, and C_b denotes the number of codons coding for the first and second residue and C_L is the total number of all possible codons except three stop codons. $T_v(a, b)$, theoretical variance is given as follows

$$T_v(a, b) = T_m(a, b)\left(\frac{1 - T_m(a, b)}{L - 1}\right) \tag{11}$$

Finally, using eqs 9, 10, and 11, DDE (a, b) can be mathematically expressed as follows[25]

$$DDE(a, b) = \frac{DPC(a, b) - T_m(a, b)}{\sqrt{T_v(a, b)}} \tag{12}$$

Learning from Imbalanced Data. In ML and bioinformatics, one of the inevitable challenging tasks is handling imbalance class distribution.[26−28] The performance of classical ML models, especially SVM, Decision Tree, AdaBoost, K-Nearest Neighbor, etc., detrimentally is affected due to ignoring the minority class and exhibits a bias toward the majority class.[29] Sampling methods can broadly be divided into two main groups: oversampling and under-sampling techniques.[30] Synthetic minority oversampling (SMOTE)[31] considers the minority class while in contrast random under-sampling considers the majority class to equalize the class distribution.[32] Thus, to take the advantages of both imbalance techniques, in this research we utilized SMOTETomek.[33] SMOTETomek is a hybrid sampling technique that combines the oversampling (SMOTE) and undersampling (Tomek Links) method and has widely been acknowledged in many domains such as software defect prediction,[32] medical data (diabetes),[34] for balancing the skewed data. In other words, the key concept of using this algorithm is to combine SMOTE method as data sampling and Tomek link as data cleaning method proposed by Tomek[35] to address the issue of imbalance data set. The pseudo code of the SMOTETomek algorithm is presented in below steps: SMOTETomek Algorithm:

1. Identify the minority class samples and the majority class samples in the imbalanced data set.

2. Apply the SMOTE algorithm to oversample the minority class

$$x_{synth} = x + (x_{neighbor} - x) \times random\_number$$

where random_number is a random value between 0 and 1.

(a) Select a minority class sample, denoted as x.

(b) Determine the k nearest neighbors (NN) of x from the minority class, denoted as $NN(x)$.

(c) Randomly select a neighbor, denoted as $x_{neighbor}$, from $NN(x)$.

(d) Generate a synthetic sample, denoted as x_{synth}, by interpolating between x and $x_{neighbor}$.

(e) Repeat steps 2b-2d for each minority class sample to generate the desired number of synthetic samples.

3. Use the Tomek Links technique to identify and remove potentially noisy samples:

(a) Construct a distance matrix between all samples in the data set.

(b) Identify the Tomek Links, which are pairs of samples from different classes that are each other's nearest neighbors.

(c) Remove the samples involved in the Tomek Links. This step removes samples that are potentially misclassified or overlapping.

4. Repeat steps 2 and 3 until the desired class balance or desired number of iterations is reached.

Classification Algorithms. Classification is a type of supervised learning used to make predictions on categorical instances. In this research, we implemented six ML algorithms for predicting Anti-MRSA peptides: KTBoost,[36] SVM,[37] CatBoost,[38] Hist-GBoost,[39] CDF,[40] and XGBoost-RFC.[41] The implementation of all these classifiers was based on the Scikit-learn,[42] gcforest,[40] and KTBoost[43] packages.

Performance Evaluation Metrics. The performance predictions of machine-learning and deep-learning models can be measured by different metrics. We use the commonly used indices, i.e., sensitivity (Sn), specificity (Sp), Matthew's correlation coefficient (MCC), and accuracy (Acc) for computing the overall performance of the proposed iMRSAPred predictor. These measures can be expressed by mathematical notation as follows

$$Sn = \frac{t_p}{t_p + f_n} \tag{13}$$

$$Sp = \frac{t_n}{t_n + f_p} \tag{14}$$

$$Acc = \frac{(t_p + t_n)}{(t_p + t_n + f_p + f_n)} \tag{15}$$

$$MCC = \frac{(t_p \times t_n - f_p \times f_n)}{\sqrt{(t_p + f_p)(t_p + f_n)(t_n + f_p)(t_n + f_n)}} \tag{16}$$

In the above eqs 13−16, t_p denotes correct positive prediction, t_n denotes correct negative prediction, f_n denotes the incorrect negative prediction, and f_p denotes the incorrect prediction of positive samples, respectively. In addition, for model robustness we used area under the receiver operating

Table 2. Performance of Different Features Using ML Classifiers over Both 10-Fold CV and Independent Tests without SMOTETomek[a]

features descriptor	classifier	10-fold CV test					independent test				
		Acc (%)	Sn (%)	Sp (%)	MCC	AUC	Acc (%)	Sn (%)	Sp (%)	MCC	AUC
DDE	KTBoost	88.57	34.09	98.08	0.464	0.876	90.45	56.66	94.44	0.594	0.917
	Hist-GBoost	90.08	48.63	97.34	0.552	0.865	93.46	66.66	98.22	0.726	0.964
	SVM	96.93	30.93	96.76	0.372	0.898	90.95	46.66	98.81	0.598	0.957
	XGBoost-RFC	86.68	17.07	98.82	0.293	0.882	89.94	40.00	98.81	0.543	0.929
	CatBoost	88.57	26.43	99.41	0.434	0.897	88.94	46.66	97.63	0.552	0.949
	CDF	91.95	55.98	98.23	0.646	0.940	94.97	70.00	99.40	0.791	0.986
RCEMT	KTBoost	92.96	66.81	97.49	0.703	0.944	94.47	76.66	97.63	0.776	0.987
	Hist-GBoost	94.46	73.78	98.08	0.768	0.951	96.48	83.33	98.81	0.858	0.988
	SVM	87.43	28.18	97.78	0.364	0.804	87.93	33.33	97.63	0.433	0.818
	XGBoost-RFC	92.21	66.21	96.75	0.676	0.924	92.46	73.33	95.85	0.701	0.977
	CatBoost	94.21	69.39	98.52	0.753	0.961	97.48	86.66	99.40	0.899	0.992
	CDF	94.09	76.28	97.19	0.765	0.965	95.47	80.00	98.22	0.817	0.991
ExPseAAC	KTBoost	94.59	71.13	98.67	0.772	0.951	95.47	76.66	98.81	0.814	0.986
	Hist-GBoost	94.84	71.13	98.96	0.780	0.960	94.97	76.66	98.22	0.795	0.987
	SVM	92.95	57.42	99.11	0.693	0.936	94.97	66.66	99.00	0.793	0.988
	XGBoost-RFC	93.84	58.48	98.99	0.735	0.953	93.96	63.33	99.40	0.746	0.990
	CatBoost	94.09	72.04	99.11	0.793	0.967	95.97	76.66	99.40	0.835	0.993
	CDF	94.21	77.12	97.19	0.768	0.963	97.48	90.00	98.81	0.900	0.994

[a]Best results are highlighted in bold.

Table 3. Performance of Different Features Using ML Classifiers over Both 10-Fold CV and Independent Test with SMOTETomek[a]

features descriptor	classifier	10-fold CV test					independent test				
		Acc (%)	Sn (%)	Sp (%)	MCC	AUC	Acc (%)	Sn (%)	Sp (%)	MCC	AUC
DDE	KTBoost	96.02	95.43	96.60	0.924	0.990	89.94	66.66	94.08	0.607	0.936
	Hist-GBoost	96.24	95.14	97.33	0.930	0.993	90.95	53.33	97.63	0.606	0.949
	SVM	92.10	84.21	100.00	0.853	0.997	90.04	36.66	100.00	0.574	0.986
	XGBoost-RFC	95.43	93.67	97.19	0.941	0.990	89.94	63.33	94.67	0.596	0.913
	CatBoost	96.90	96.02	97.78	0.941	0.995	93.46	73.33	97.04	0.735	0.956
	CDF	97.71	98.38	97.05	0.955	0.997	95.97	76.66	99.40	0.835	0.980
RCEMT	KTBoost	97.82	97.49	95.15	0.936	0.996	96.48	93.33	97.04	0.869	0.991
	Hist-GBoost	97.63	98.68	96.60	0.949	0.997	94.47	80.00	97.04	0.781	0.991
	SVM	96.30	97.49	95.11	0.927	0.992	95.97	90.00	97.04	0.847	0.991
	XGBoost-RFC	92.39	94.10	90.70	0.848	0.981	90.95	80.00	92.89	0.667	0.966
	CatBoost	97.41	98.97	95.86	0.950	0.997	96.98	93.33	97.63	0.886	0.994
	CDF	96.01	97.05	94.98	0.921	0.993	93.96	90.00	94.67	0.787	0.987
ExPseAAC	KTBoost	97.05	98.23	95.87	0.941	0.997	95.47	90.00	96.44	0.831	0.991
	Hist-GBoost	96.67	99.11	96.22	0.953	0.999	93.97	86.66	97.63	0.842	0.989
	SVM	90.30	80.61	100	0.822	0.999	95.47	83.33	97.63	0.821	0.982
	XGBoost-RFC	93.84	58.48	100	0.735	0.953	93.96	63.33	97.89	0.746	0.990
	CatBoost	98.15	99.41	96.90	0.963	0.999	97.48	93.33	98.22	0.903	0.996
	CDF	96.09	96.46	95.71	0.923	0.995	95.97	93.33	96.44	0.853	0.992

[a]Best results are highlighted in bold.

characteristic (ROC) curve (AUC) values as an independent evaluation metric.

Model Assessment and Evaluation. CV is the widely used performance evaluation method of the ML and DL models.[44] CV provides precise and accurate estimation of the prediction system by splitting the whole set of data into the training and testing part: the training part can be used to build/develop the model and the testing part to assess the generalization capability of the trained model. Thus, k-fold deem to be the simplest CV technique in developing the computational model.[45] The data set, in this strategy, is divided into k-folds or fixed-sized subsets. The predictive model is then trained on $k − 1$ of the subsets and tested on the rest of the

subset. The process is k times iteratively repeated, with each subset serving as the testing set exactly once. The average efficacy of the developed predictor is then calculated by summing the across all k iterations. In our study, we used the 10-fold CV method for designing the proposed iMRSAPred. Furthermore, we also performed independent tests to better estimate the generalization efficacy of the proposed Anti-MRSA protocol on unseen peptides.

■ RESULTS AND DISCUSSION

Classifiers Performance Using Different Feature Encoding Schemes without Applying SMOTETomek.
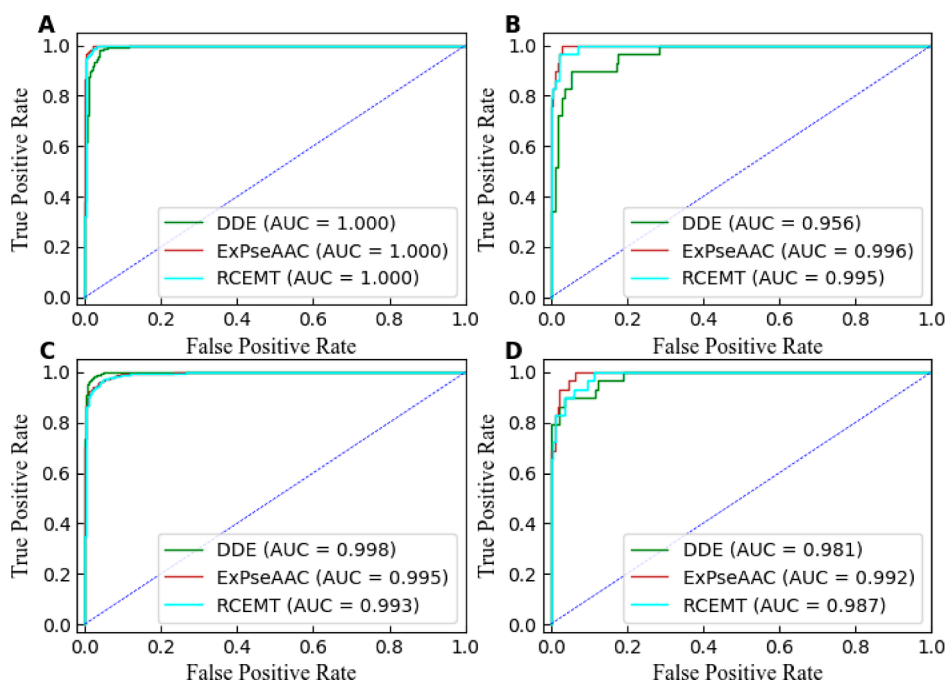
Figure 2. ROC curves of the proposed iMRSAPreda (A,B) and iMRSAPredb (C,D) for training and testing data sets.
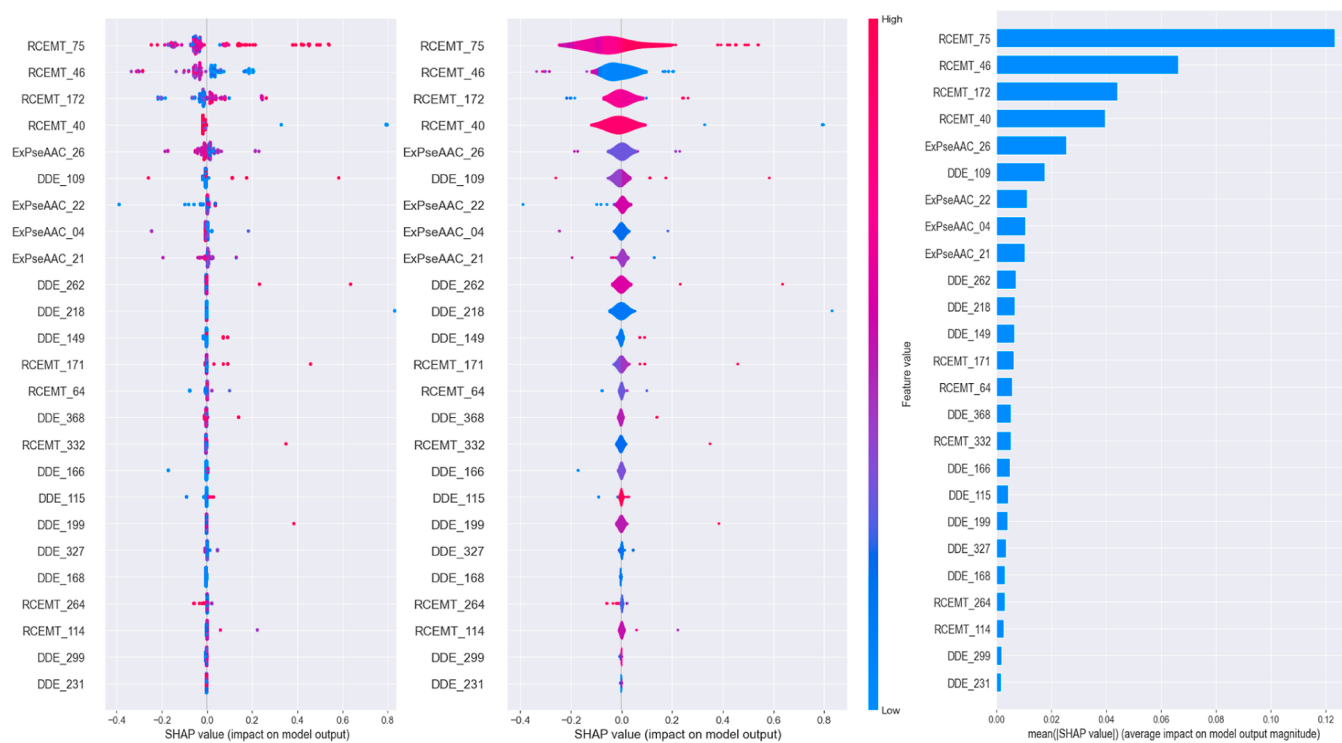

Figure 3. Feature analysis and contribution of the top ranked attributes using the SHAP method.

In this section, we analyze the predictive performance of six ML classifiers, namely, KTBoost, HistGBoost, SVM, XGBoost-RFC, CatBoost, and CDF algorithms, using three effective feature encoding schemes, i.e., DDE, RCEMT, and ExPseAAC feature vectors in Anti-MRSA prediction. The ML algorithms were evaluated on 10-fold CV and independent tests without applying the SMOTETomek method. The classifiers performance on imbalance data along with the evaluation indexes Acc, Sn, AUC, Sp, and MCC are reported in Table 2. It can be seen from Table 2 that in the case of DDE feature vector, CDF

classifier achieved the highest performance in terms of *Acc = 91.95* and *94.47%* and *MCC = 0.646* and 0.791 are on training and testing data sets, respectively. In case of RCEM descriptor, Catboost classifier attained the better overall outcomes compared to other ML algorithms. Similarly, using the ExPseAAC encoding method in conjunction with CDF Classifier improved the *Acc 1.51%* and *MCC 0.060* compared to CatBoost Classifier on testing data. Thus, the aforementioned investigation reveals several observations: first, the PCP and energy estimation-based attributes in combination with
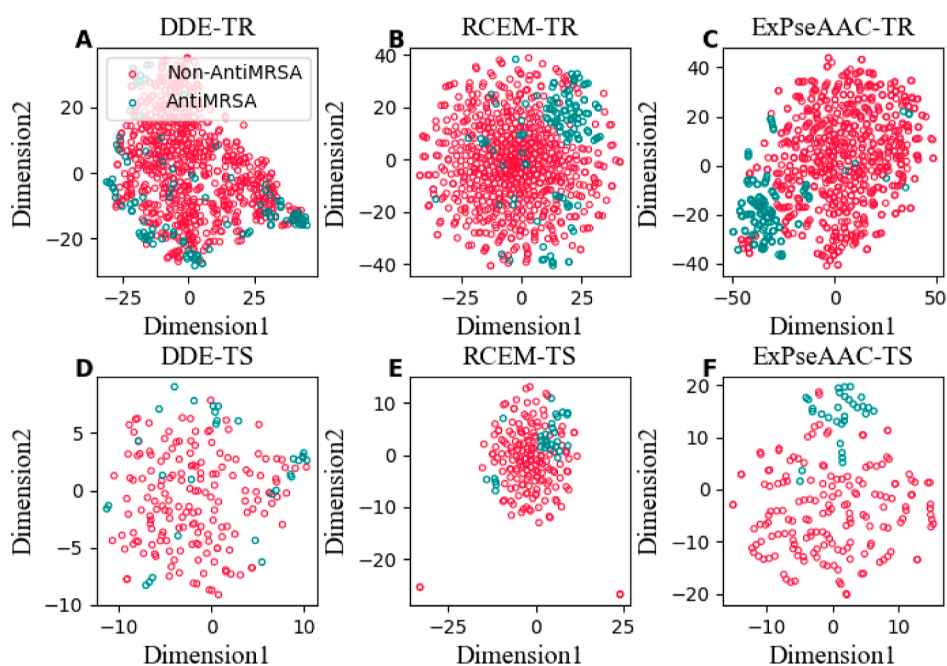
Figure 4. t-SNE visual depict of Anti-MRSA (green) and Non-Anti-MRSA (red) peptides for the training data set (A−C) and independent data sets (D−F) in a two-dimensional feature-vector: DDE-TR (A), RCEMT-TR (B), ExPseAAC-TR (C), DDE-TS (D), RCEMT-TS (E), and ExPseAAC-TS (F).
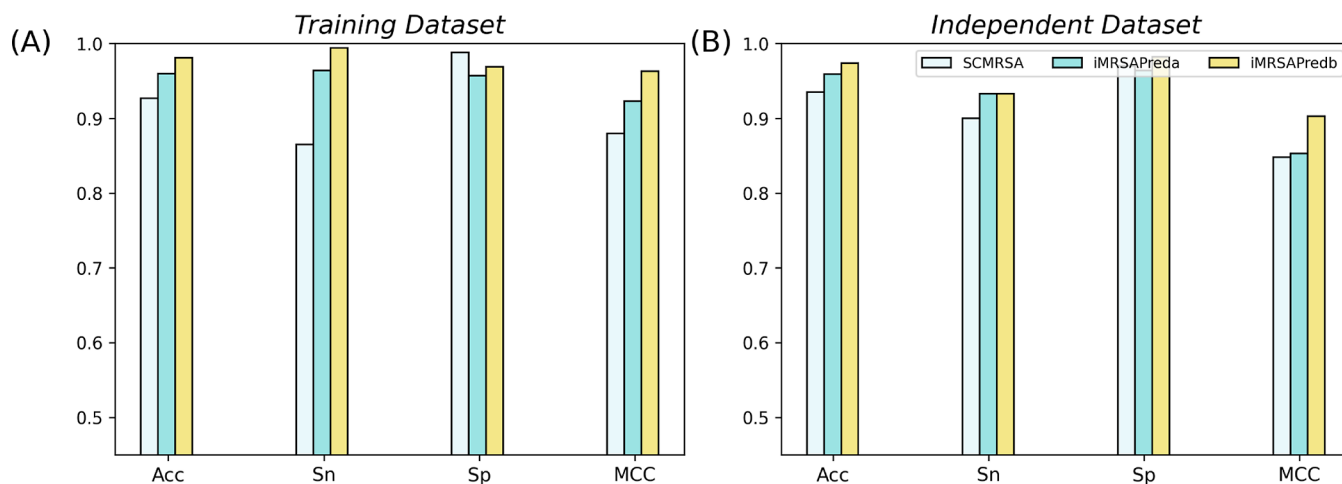

Figure 5. Performance comparison of our proposed methods with the SCMRSA tool on training (A,B) independent data set.

different ML classifiers generate better results over the training and independent data sets. This demonstrate that RCEMT and ExPseAAC feature-space effectively contribute in discriminating Anti-MRSA and Non-Anti-MRSA peptides. Second, due to skewed data, the learning models predict the inconsistent and bias results, i.e., specifically balance Acc and MCC on the independent test. To solve this problem, we motivated to apply SMOTETomek algorithm to achieve more stable and high Anti-MRSA predictions.

Classifiers Performance Using Various Feature Encoding Schemes after Applying SMOTETomek. In the present subsection, we examine the classification performance of Anti-MRSA peptides by applying the SMOTETomek method. In Table 3, we report the success rates of different ML algorithms against three feature representation methods. The anticipated prediction score shows that the ML classifiers particularly KTBoost, CDF, and CatBoost models enhanced

the average performance in terms of all evaluation indicators on training and testing samples. As can be seen from ROC curve in Figure 2, the CatBoost model is outperformer using the ExPseAAC encoding scheme. The highest obtained *Acc* is *98.15* and *97.48%* on training and independent test, respectively. The second best performer is the CDF model which obtained relatively lower prediction rates, i.e., *2.06% Acc* and *MCC* of 0.004. Interestingly, KTBoost, Hist-GBoost, and XGBoost-RFC produced impressive results on training data but performed poorly on the blind test (independent data set). Consequently, the observed evidence indicates that the ML models on balanced data consistently predict the unbiased outcomes. The ROC curve for the best models was created for both the training and independent sets. The results, as depicted in Figure 2, indicate that the ExPseAAC feature representation method achieved the highest *AUC* values of

Table 4. Performance Comparison against the Existing Method for Benchmark Data Set[a]

predictors	training data set				independent data set			
	Acc (%)	Sn (%)	Sp (%)	MCC	Acc (%)	Sn (%)	Sp (%)	MCC
SCMRSA[8]	92.70	86.50	98.80	0.880	93.50	90.00	97.00	0.848
iMRSAPred[a] (proposed)	96.09	96.46	95.71	0.923	95.97	93.33	96.44	0.853
iMRSAPred[b] (proposed)	98.15	99.41	96.90	0.963	97.48	93.33	98.22	0.903

[a]Best results are highlighted in bold.

0.996 and 0.992 using CatBoost and CDF models on independent data set.

Feature Ranking and Contribution Analysis. The prediction power of a model can be analyzed by examining the contribution of each feature vector.[46] To do so, in this research, we considered the well-known algorithm named SHAP[47] to interpret the prediction of developed iMRSApred model. The SHAP method, assigned each extracted attribute a SHAP value in the descending order, indicating the impact of each feature-space on the classification of each sample.[48]

Figure 3 shows the top 25 high-ranked discriminative properties extracted from three descriptors, i.e., RCEMT, ExPseAAC, and DDE. In the context of corresponding feature-space, the color scatterplot represents the influence of specific feature. Thus, overall energy estimation-based (RCEMT_75, RCEMT_46, RCEMT_172, and RCEMT_40), physicochemical (ExPseAAC_26, ExPseAAC_22, and ExPseAAC_21), and sequential-based properties (DDE_262, DDE_218, and DDE_149) contributed well in predicting accurate Anti-MRSA peptides.

In order to further explain the contribution of engineered features, we used two dimension scatters plot t-SNE,[49] as shown in Figure 4A−F. The red dots denote the Non-AntiMARSA, and green dots denote the Anti-MRSA peptide samples.

Comparison against Existing Methods. For evaluation purposes, we compare the prediction performance of our proposed methods with the existing SCMRSA tool[8] for identifying Anti-MRSA peptides. Figure 5 illustrates the comparative scores of developed Anti-MRSA predictors on training (A) and testing (B) data sets, respectively.

The comparison outcomes between our developed computational methods for Anti-MRSA activity prediction and the SCMRSA predictor are noted in Table 4. We denoted the best success rates for the respective evaluation indicators *Acc*, *Sn*, *Sp*, and *MCC* with a bold face in Table 4. We can observe that iMRSAPredb is the best performer in terms of all performance measures both on training and independent data sets. iMRSAPredb improved the balance *Acc* by 5.45% and *MCC* by 8.3% on training data set and *Acc* of 3.98% and *MCC* of 5.5% on independent test compared with SCMRSA. However, our proposed methods are relatively lower than the existing model in terms of Sp which are not great powerful. The second best predictor outperformed the existing tool on training and testing data set by *Acc* of 3.39 and 2.47%, respectively. Thus, the comparative discussion indicates the capability of the iMRSAPredb protocol to accurately discriminate Anti-MRSA peptides.

CONCLUSIONS

Owning to the biological applications as a therapeutic agent to combat bacterial infections, the identification of Anti-MRSA peptides is crucial in developing new weapons as antibiotic drugs. In this study, we developed iMRSAPred, a novel ML predictor for targeting Anti-MRSA peptides. The proposed model outperformed the existing state-of-the-art SCMARSA predictor and achieved well-balanced results in terms of all performance metrics. We extracted the biological features from AA residues considering their physiochemical-, energy estimation-, and sequence-based descriptors. Finally, we applied the SMOTETomek algorithm to achieve better results compared with the existing method in the literature. Our work has some limitations that need to be highlighted. We tested our model on only one data set. Based on the availability of other data sets, we will extend this work for further improvement. In future, we will build a publicly accessible web server for recognizing large-scale therapeutic peptides having Anti-MRSA activity and other activities, i.e., anticancer activity, antimicrobial activity, antiviral, antibacterial activity, antifungal, antihypertensive, cell-penetration activity, etc.

ASSOCIATED CONTENT

Data Availability Statement

Data set and source code are publicly available at GitHub: https://github.com/Muhammad-Arif-NUST/iMRSAPred.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c08303.

Derived values of five physicochemical properties of 20 AAs and RCEM properties for iMRSAPred (PDF)

AUTHOR INFORMATION

Corresponding Author

Tanvir Alam − *College of Science and Engineering, Hamad Bin Khalifa University, Doha 34110, Qatar;* orcid.org/0000-0001-7033-3693; Phone: +974 4454 2227; Email: talam@hbku.edu.qa

Authors

Muhammad Arif − *College of Science and Engineering, Hamad Bin Khalifa University, Doha 34110, Qatar*

Ge Fang − *State Key Laboratory for Organic Electronics and Information Displays, Institute of Advanced Materials (IAM), Nanjing 210023, P. R. China; Center for Research Innovation and Biomedical Informatics, Faculty of Medical Technology, Mahidol University, Bankok 10700, Thailand*

Huma Fida − *Department of Microbiology, Abdul Wali Khan University, Mardan 23200 KPK, Pakistan*

Saleh Musleh − *College of Science and Engineering, Hamad Bin Khalifa University, Doha 34110, Qatar*

Dong-Jun Yu − *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210023, China;* orcid.org/0000-0002-6786-8053

Complete contact information is available at:
https://pubs.acs.org/10.1021/acsomega.3c08303

REFERENCES

(1) Songnaka, N.; Lertcanawanichakul, M.; Hutapea, A. M.; Krobthong, S.; Yingchutrakul, Y.; Atipairin, A. Purification and Characterization of Novel Anti-MRSA Peptides Produced by Brevibacillus sp. SPR-20. *Molecules* **2022**, *27*, 8452.

(2) Junnila, J.; Hirvioja, T.; Rintala, E.; Auranen, K.; Rantakokko-Jalava, K.; Silvola, J.; Lindholm, L.; Gröndahl-Yli-Hannuksela, K.; Marttila, H.; Vuopio, J. Changing epidemiology of methicillin-resistant Staphylococcus aureus in a low endemicity area—new challenges for MRSA control. *Eur. J. Clin. Microbiol. Infect. Dis.* **2020**, *39*, 2299−2307.

(3) De la Calle, C.; Morata, L.; Cobos-Trigueros, N.; Martinez, J.; Cardozo, C.; Mensa, J.; Soriano, A. Staphylococcus aureus bacteremic pneumonia. *Eur. J. Clin. Microbiol. Infect. Dis.* **2016**, *35*, 497−502.

(4) Stogios, P. J.; Savchenko, A. Molecular mechanisms of vancomycin resistance. *Protein Sci.* **2020**, *29*, 654−669.

(5) Ahmed, M. O.; Baptiste, K. E. Vancomycin-resistant enterococci: a review of antimicrobial resistance mechanisms and perspectives of human and animal health. *Microb. Drug Resist.* **2018**, *24*, 590−606.

(6) Masimen, M. A. A.; Harun, N. A.; Maulidiani, M.; Ismail, W. I. W. Overcoming methicillin-resistance Staphylococcus aureus (MRSA) using antimicrobial peptides-silver nanoparticles. *Antibiotics* **2022**, *11*, 951.

(7) Zhu, Y.; Hao, W.; Wang, X.; Ouyang, J.; Deng, X.; Yu, H.; Wang, Y. Antimicrobial peptides, conventional antibiotics, and their synergistic utility for the treatment of drug-resistant infections. *Med. Res. Rev.* **2022**, *42*, 1377−1422.

(8) Charoenkwan, P.; Kanthawong, S.; Schaduangrat, N.; Li', P.; Moni, M. A.; Shoombuatong, W. SCMRSA: a New Approach for Identifying and Analyzing Anti-MRSA Peptides Using Estimated Propensity Scores of Dipeptides. *ACS Omega* **2022**, *7*, 32653−32664.

(9) Mishra, A.; Pokhrel, P.; Hoque, M. T. StackDPPred: a stacking based prediction of DNA-binding protein from sequence. *Bioinformatics* **2019**, *35*, 433−441.

(10) Wang, Z.; Wu, C.; Zheng, K.; Niu, X.; Wang, X. SMOTETomek-based resampling for personality recognition. *IEEE Access* **2019**, *7*, 129678−129689.

(11) Ahmed, S.; Arif, M.; Kabir, M.; Khan, K.; Khan, Y. D. PredAoDP: Accurate identification of antioxidant proteins by fusing different descriptors based on evolutionary information with support vector machine. *Chemom. Intell. Lab. Syst.* **2022**, *228*, 104623.

(12) Ge, F.; Hu, J.; Zhu, Y.-H.; Arif, M.; Yu, D.-J. TargetMM: Accurate Missense Mutation Prediction by Utilizing Local and Global Sequence Information with Classifier Ensemble. *Comb. Chem. High Throughput Screening* **2021**, *25*, 38−52.

(13) Ge, F.; Muhammad, A.; Yu, D.-J. DeepnsSNPs: Accurate prediction of non-synonymous single-nucleotide polymorphisms by combining multi-scale convolutional neural network and residue environment information. *Chemom. Intell. Lab. Syst.* **2021**, *215*, 104326.

(14) Wang, G.; Li, X.; Wang, Z. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.* **2016**, *44*, D1087−D1093.

(15) Ahmed, S.; Kabir, M.; Arif, M.; Ali, Z.; Khan Swati, Z. N. Prediction of human phosphorylated proteins by extracting multi-perspective discriminative features from the evolutionary profile and physicochemical properties through LFDA. *Chemom. Intell. Lab. Syst.* **2020**, *203*, 104066.

(16) Dosztanyi, Z.; Csizmok, V.; Tompa, P.; Simon, I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* **2005**, *347*, 827−839.

(17) Hoque, M. T.; Yang, Y.; Mishra, A.; Zhou, Y. sDFIRE: Sequence-specific statistical energy function for protein structure prediction by decoy selections. *J. Comput. Chem.* **2016**, *37*, 1119−1124.

(18) Fu, X.; Cai, L.; Zeng, X.; Zou, Q. StackCPPred: a stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency. *Bioinformatics* **2020**, *36*, 3028−3034.

(19) Mishra, A.; Khanal, R.; Kabir, W. U.; Hoque, T. AIRBP: accurate identification of RNA-binding proteins using machine learning techniques. *Artif. Intell. Med.* **2021**, *113*, 102034.

(20) Chou, K.-C.; Cai, Y.-D. Using GO-PseAA predictor to identify membrane proteins and their types. *Biochem. Biophys. Res. Commun.* **2005**, *327*, 845−847.

(21) Arif, M.; Ahmed, S.; Ge, F.; Kabir, M.; Khan, Y. D.; Yu, D.-J.; Thafar, M. StackACPred: prediction of anticancer peptides by integrating optimized multiple feature descriptors with stacked ensemble approach. *Chemom. Intell. Lab. Syst.* **2022**, *220*, 104458.

(22) Hayat, M.; Tahir, M.; Alarfaj, F. K.; Alturki, R.; Gazzawe, F. NLP-BCH-Ens: NLP-based intelligent computational model for discrimination of malaria parasite. *Comput. Biol. Med.* **2022**, *149*, 105962.

(23) Saravanan, V.; Gautham, N. Harnessing computational biology for exact linear B-cell epitope prediction: a novel amino acid composition-based feature descriptor. *OMICS: J. Integr. Biol.* **2015**, *19*, 648−658.

(24) Manavalan, B.; Lee, J. FRTpred: A novel approach for accurate prediction of protein folding rate and type. *Comput. Biol. Med.* **2022**, *149*, 105911.

(25) Chen, Z.; Zhao, P.; Li, F.; Marquez-Lago, T. T.; Leier, A.; Revote, J.; Zhu, Y.; Powell, D. R.; Akutsu, T.; Webb, G. I.; et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Briefings Bioinf.* **2020**, *21*, 1047−1057.

(26) Japkowicz, N. *Learning from imbalanced data sets: a comparison of various strategies*; AAAI workshop, 2000, pp 10−15.

(27) Wan, S.; Duan, Y.; Zou, Q. HPSLPred: an ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source. *Proteomics* **2017**, *17*, 1700262.

(28) Song, L.; Li, D.; Zeng, X.; Wu, Y.; Guo, L.; Zou, Q. nDNA-prot: identification of DNA-binding proteins based on unbalanced classification. *BMC Bioinf.* **2014**, *15*, 298−310.

(29) Arif, M.; Ali, F.; Ahmad, S.; Kabir, M.; Ali, Z.; Hayat, M. Pred-BVP-Unb: Fast prediction of bacteriophage Virion proteins using un-biased multi-perspective properties with recursive feature elimination. *Genomics* **2020**, *112*, 1565−1574.

(30) Khuat, T. T.; Le, M. H. Evaluation of sampling-based ensembles of classifiers on imbalanced data for software defect prediction problems. *SN Comput. Sci.* **2020**, *1*, 108.

(31) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321−357.

(32) Khleel, N. A. A.; Nehéz, K. A novel approach for software defect prediction using CNN and GRU based on SMOTE Tomek method. *J. Intell. Inf. Syst.* **2023**, *60*, 673−707.

(33) Ning, Q.; Zhao, X.; Ma, Z. A novel method for Identification of Glutarylation sites combining Borderline-SMOTE with Tomek links technique in imbalanced data. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2022**, *19*, 2632−2641.

(34) Zeng, M.; Zou, B.; Wei, F.; Liu, X.; Wang, L. Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data *IEEE International Conference of Online Analysis and Computing Science*; ICOACS, 2016, pp 225−228.

(35) Tomek, I. *Two modifications of CNN*; IEEE, 1976.

(36) Sigrist, F. KTBoost: Combined kernel and tree boosting. *Neural Process. Lett.* **2021**, *53*, 1147−1160.

(37) Schuldt, C.; Laptev, I.; Caputo, B. *Proceedings of the 17th International Conference on Pattern Recognition*; ICPR, 2004, pp 32−36.Recognizing human actions: a local SVM approach

(38) Musleh, S.; Islam, M. T.; Qureshi, R.; Alajez, N.; Alam, T. MSLP: mRNA subcellular localization predictor based on machine learning techniques. *BMC Bioinf.* **2023**, *24*, 109−123.

(39) Arif, M.; Ahmad, S.; Ali, F.; Fang, G.; Li, M.; Yu, D.-J. TargetCPP: accurate prediction of cell-penetrating peptides from optimized multi-scale features using gradient boost decision tree. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 841−856.

(40) Zhou, Z.-H.; Feng, J. Deep forest. *Natl. Sci. Rev.* **2019**, *6*, 74−86.

(41) Zhong, J.; Sun, Y.; Peng, W.; Xie, M.; Yang, J.; Tang, X. XGBFEMF: an XGBoost-based framework for essential protein prediction. *IEEE Trans. NanoBiosci.* **2018**, *17*, 243−250.

(42) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(43) Khattak, A.; Zhang, J.; Chan, P.-W.; Chen, F. Turbulence along the Runway Glide Path: The Invisible Hazard Assessment Based on a Wind Tunnel Study and Interpretable TPE-Optimized KTBoost Approach. *Atmosphere* **2023**, *14*, 920.

(44) Hu, J.; Zeng, W.-W.; Jia, N.-X.; Arif, M.; Yu, D.-J.; Zhang, G.-J. Improving DNA-Binding Protein Prediction Using Three-Part Sequence-Order Feature Extraction and a Deep Neural Network Algorithm. *J. Chem. Inf. Model.* **2023**, *63*, 1044−1057.

(45) Ahmed, S.; Kabir, M.; Ali, Z.; Arif, M.; Ali, F.; Yu, D.-J. An integrated feature selection algorithm for cancer classification using gene expression data. *Comb. Chem. High Throughput Screening* **2019**, *21*, 631−645.

(46) Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2020**, *2*, 573−584.

(47) Cai, L.; Wang, L.; Fu, X.; Xia, C.; Zeng, X.; Zou, Q. ITP-Pred: an interpretable method for predicting, therapeutic peptides with fused features low-dimension representation. *Briefings Bioinf.* **2021**, *22*, bbaa367.

(48) Wang, Y.; Xie, Y.; Luo, Y.; Jia, P.; Wei, J.; Zhang, J.; Yan, W.; Huang, J. iASMP: An interpretable in silico predictive tool focusing on species-specific antimicrobial peptides. *J. Pept. Sci.* **2023**, *29*, No. e3490.

(49) Ge, R.; Xia, Y.; Jiang, M.; Jia, G.; Jing, X.; Li, Y.; Cai, Y. HybAVPnet: a novel hybrid network architecture for antiviral peptides identification. 2022, bioRxiv, 2022-06. https://www.biorxiv.org/content/10.1101/2022.06.10.495721v1 (accessed Jun13, 2022).