IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 September 2023, accepted 27 September 2023, date of publication 2 October 2023,
date of current version 12 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3321100

RESEARCH ARTICLE

pAtbP-EnC: Identifying Anti-Tubercular Peptides Using Multi-Feature Representation and Genetic Algorithm-Based Deep Ensemble Model

SHAHID AKBAR[1], ALI RAZA[2], TAMARA AL SHLOUL[3], ASHFAQ AHMAD[2], AAMIR SAEED[4], YAZEED YASIN GHADI[5], ORKEN MAMYRBAYEV[6], AND ELSAYED TAG-ELDIN[7]

[1]Department of Computer Science, Abdul Wali Khan University Mardan, Mardan, Khyber Pakhtunkhwa 23200, Pakistan
[2]Department of Computer Science, MY University Islamabad, Islamabad 44000, Pakistan
[3]Department of General Education, Liwa College of Technology, Abu Dhabi, United Arab Emirates
[4]Department of Computer Science and IT, University of Engineering and Technology, Peshawar, Peshawar 25000, Pakistan
[5]Department of Computer Science, Al Ain University, Abu Dhabi, United Arab Emirates
[6]Institute of Information and Computational Technologies, 050010 Almaty, Kazakhstan
[7]Faculty of Engineering and Technology, Future University in Egypt, New Cairo 11835, Egypt

Corresponding authors: Elsayed Tag-Eldin (elsayed.tageldin@fue.edu.eg) and Shahid Akbar (shahid.akbar@awkum.edu.pk)

ABSTRACT Mycobacterium tuberculosis, a highly perilous pathogen in humans, serves as the causative agent of tuberculosis (TB), affecting nearly 33% of the global population. With the increasing prevalence of multidrug-resistant TB, there is a need for novel and efficacious alternative therapies. Peptide therapies have emerged as a favorable alternative due to their remarkable specificity in targeting cells without affecting healthy cells. However, the experimental identification methods of anti-tubercular peptides (AtbPs) are labor-intensive and costly. Therefore, accurate prediction of AtbPs has become challenging due to the large number of peptide samples. In this paper, we propose an ensemble learning model to enhance the prediction outcomes by addressing the limitations of individual learning models. We formulate the training samples by utilizing four distinct representation methods: AAindex, Composition/Transition/Distribution, Dipeptide Deviation from Expected Mean, and Enhanced Grouped Amino Acid Composition to numerically encode peptide samples. The feature vectors extracted from these methods are fused to develop a compact vector. We evaluate the prediction rates using three different classification models, employing both individual and heterogeneous vectors. Furthermore, we enhance the prediction and training capabilities of the proposed model by using the predicted labels of the individual classifiers for implementing an ensemble deep model via a genetic algorithm. Through evaluation of both the training datasets and independent datasets, our proposed ensemble learner achieves impressive accuracies of 97.80%, 95.13%, 93.91%, and 94.17%, using RD training, MD training, RD independent, and MD independent datasets, respectively. Our findings demonstrate that the proposed pAtbP-EnC model outperforms existing predictors by reporting approximately 11% higher training accuracy. We conclude that the pAtbP-EnC predictor will be a considerable tool in the field of pharmaceutical design and research academia. The used datasets and the source code are publicly available at https://github.com/Intelligent-models/pAtbP-EnC2023.

INDEX TERMS Anti-tubercular peptides, ensemble classification, genetic algorithm, hybrid representation, k-fold cross-validation test.

I. INTRODUCTION

Tuberculosis (TB) is a life-threatening condition initiated by an infection with the bacterium Mycobacterium

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Olague.

tuberculosis [1], [2]. It presents a substantial worldwide health risk, leading to widespread illness and mortality on a global scale. According to the World Health Organization (WHO), 10 million people were affected by tuberculosis (TB) in 2017, resulting in 1.6 million documented deaths. Despite advancements in treatment options, the prevalence of tuberculosis (TB) continues to escalate due to various factors, including inadequate medication utilization, low-quality pharmaceuticals, and premature discontinuation of therapy. These factors contribute to the development of drug-resistant strains for Mycobacterium TB, which are more difficult to manage [3]. Multidrug-resistant tuberculosis is a kind of drug-resistant tuberculosis in which bacteria resist primary anti-TB medications like rifampicin and isoniazid [4]. This critical global health challenge poses a substantial threat to global health security, demanding immediate attention [5]. Moreover, extensively drug-resistant tuberculosis (XDR-TB), a variant of multidrug-resistant tuberculosis (MDR-TB), exhibits non-responsiveness to secondary anti-TB medications, further complicating the scenario. Furthermore, the adverse consequences linked to current anti-tubercular drugs and the extended duration of treatment present significant hurdles. In light of these barriers, exploring innovative therapeutic options with inventive mechanisms of action against tuberculosis becomes imperative. Peptide-based interventions emerge as a promising avenue in this context. A notable attribute of peptides is their capacity to interact with various biological targets, including specific molecules within living organisms. This characteristic renders them appealing candidates for the development of effective anti-tubercular agents. Facilitating the expedited discovery and development of AtbPs assumes paramount importance. However, identifying and developing these peptides through empirical studies is time-intensive and financially burdensome. To tackle this challenge, there is an increasing interest in binding advanced computational methods and techniques to streamline and cost-effectively identify peptide candidates with anti-tubercular potential [6]. Integrating computational approaches into AtbPs, the discovery proves indispensable for expediting and streamlining their synthesis [7], [8].

Numerous intelligent learning frameworks have been developed to efficiently identify anti-TB peptides. For example, the AntiTBpred predictor utilized a concatenate vector approach [8]. The informative features were encoded from the amino acid sequences by amino acid composition, dicoupling composition (DPC), fragment-based composition, and binary structure information. To measure the predictive performance of their model, they employed an ensemble learning algorithm within their proposed framework. Similarly, Khatun et al. introduced the ''iAntiTB'' Predictor intending to identify anti-tuberculosis agents [9]. To represent the anti-TB samples, they employed four distinct encoding techniques, namely DPC, tripeptide composition (TPC), binary coding, and amino acid index characteristics. The resulting feature vectors were evaluated using Random Forest (RF), and Support Vector Machines (SVM). Furthermore, to improve the prediction outcomes, the resulting outcomes derived from SVM and RF were integrated through linear regression. A distinct framework named the AtbPpred model has been devised to predict anti-TB peptides [10]. The peptide samples within the training datasets were encoded utilizing nine diverse strategies. In the preliminary stage of AtbPpred, a two-level feature selection methodology was implemented to ascertain the most appropriate information derived from the encoding schemes. Following this, prediction models for each descriptor were established employing a comprehensive randomized tree classification model. To assess the final prediction performance, the prediction scores obtained from each model were combined and fed into an extremely randomized tree ERT. Similarly, Chen et al. have developed an iATP web server to identify anti-peptides TB [11]. To construct a robust computational model iATP used the combination of pseudo-g-gap DPC and support vector machines. For the numerical representation of amino acid sequences and sample preparation, a pseudo-g-gap peptide DPC has been used. To identify the most useful descriptors, the resulting feature vector is formed by selecting the optimal features using incremental feature selection. Finally, SVM was applied to evaluate the model's performance. Recently, to discriminate anti-TB peptides, the use of multiple descriptors and measure selection based on differentiation features has been proposed by Akbar et al. [12]. Their approach involved eliminating noisy and less informative features and leveraging a majority voting-based learning model to improve the predictive outcomes. However, despite these advancements, it is important to note that existing models still have certain limitations regarding accuracy. Moreover, the existing methods lacked important features with inadequate training and generalization abilities. Therefore, a dependable and efficient learning model is highly crucial to accurately discriminate AtbPs and non-AtbPs. Initially, we encoded the amino acid samples numerically using four distinct methodologies: amino acid index, dipeptide deviation from the expected mean, enhanced grouped amino acid composition, and composition/transition/distribution. Furthermore, multi-information vectors were combined to address the weaknesses of single encoding vectors. The performance of the extracted vectors was then evaluated using various machine learning techniques, including RF, ETC, and Deep neural Network (DNN). Additionally, an ensemble classifier via genetic algorithm was employed to improve the evaluation outcomes of the developed model. The overall structure of our suggested model is depicted in Figure 1. Moreover, to assess the model reliability two independent datasets were also utilized to tackle over fitting problems.

II. MATERIALS AND METHODS

A. DATASET

One of the primary challenges in machine learning is the development or selection of adequate benchmark datasets,

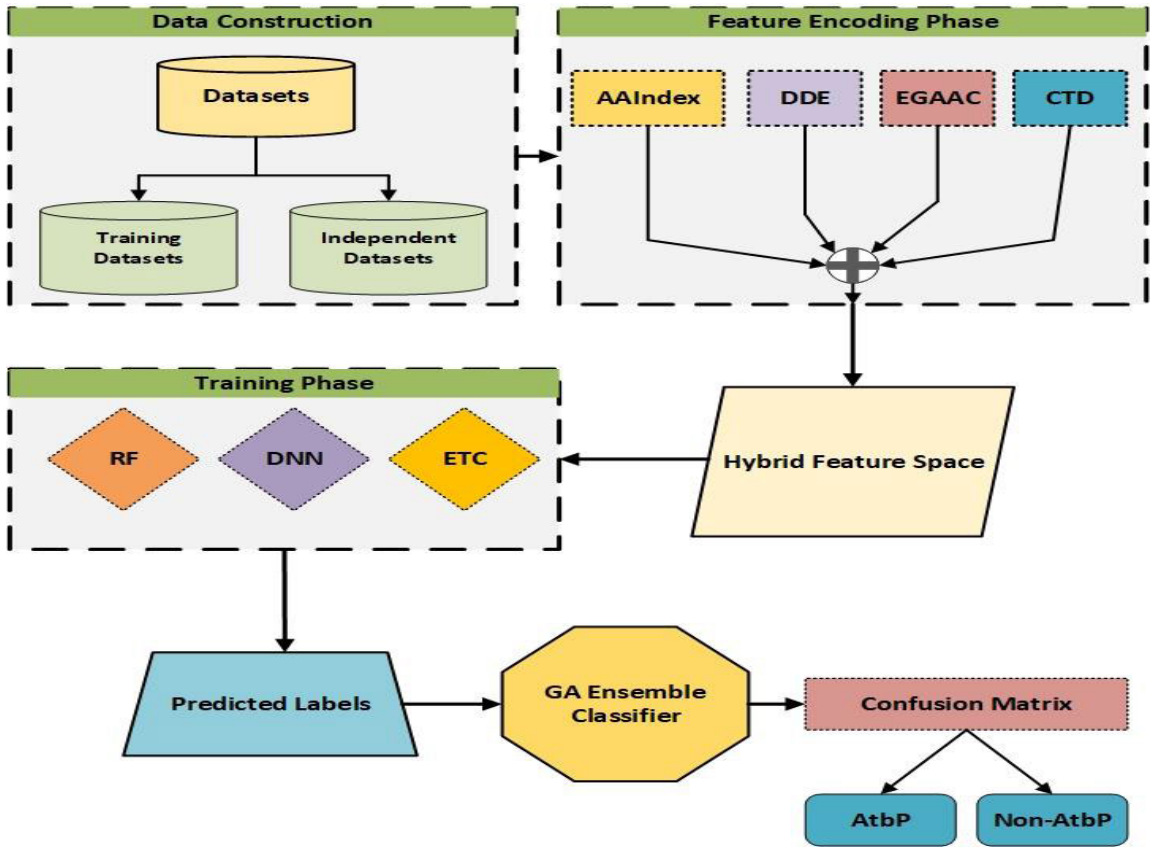


FIGURE 1. The framework of the proposed pAtbP-EnC model.

as it significantly impacts the training of computational models [13].

In this work, we used the same AtbPs datasets, namely AntiTb RD and AntiTb MD which were previously employed by Usmani et al. [8]. These training datasets consisted of positive sequences (representing peptides with anti-tubercular activity) and negative sequences. Initially, the datasets comprised 492 sequences of peptides, out of which 246 were identified as anti-TB peptides (ATBPs). The samples were sourced from various databases, including Swiss-Prot [14] and databases specifically focused on antimicrobial peptides. To ensure the diverse representation of amino acids with different characteristics, these samples were randomly selected. Experimental confirmation revealed that all positive sequences in the datasets contain Anti-tubercular peptides (AtbPs) retrieved from the AntiTbpdb database [15]. The length of amino acids per sequence ranged from 5 to 61. To ensure data integrity, duplicate and homogeneous samples were eliminated using CD-HIT [16]. Subsequently, after the preprocessing, 80% of the total sequences were selected for training datasets. In this model, the training dataset comprised 398 samples, equally distributed between the two classes, specifically, 199 AtbPs and 199 non-AtbPs. To assess the generalization capacity of our proposed study and address overfitting concerns, the remaining 20% of sequences are applied as independent sets. The independent dataset comprised 94 samples, equally distributed into 47 AtbPs and 47 non-AtbPs. Moreover, while developing independent samples, none of the training samples were repeated.

B. FEATURE FORMULATION METHODS

1) AMINO ACID INDEX (AAI)

The physicochemical characteristics of amino acids perform an essential part in demonstrating biochemical reactions and have been widely utilized in the field of bioinformatics. To incorporate the diverse range of physicochemical properties associated with amino acids, we employed the AAIndex database, which aggregates numerous published indices [17]. Each physicochemical property is represented by a series of 20 numerical values, with each value corresponding to one of the 20 amino acids. The AAIndex database comprises a comprehensive collection of 544 distinct physicochemical properties. However, to ensure the reliability and comprehensiveness of our analysis, we excluded properties that had 'NA' (not available) values assigned to any of the amino acids. Consequently, we acquired a dataset of 531 physicochemical properties that were deemed suitable for subsequent analysis. Diverging from residue-based encoding methods that primarily focus on amino acid identity and evolutionary

information, we adopted an alternative approach. We constructed a vector representation by calculating the mean values across the 531 physicochemical properties. This vector representation allowed us to capture and incorporate a wide range of physicochemical characteristics associated with the amino acids. Furthermore, it also facilitated the representation of samples across different window sizes, enabling us to account for variations in peptide length and context [18]. The AAIndex descriptor has been successfully applied in various applications, including the prediction of protein ubiquitination sites, protein malonylation sites, and many more [19].

2) DIPEPTIDE DERIVATION FROM EXPECTED MEAN (DDE)

The Dipeptide Derivation from the expected mean is a sequential feature computing method that can be calculated using three different parameters such as theoretical variance (T_v), theoretical mean (T_m), and dipeptide composition (D_c). Mathematically, DDE can be computed using these parameters in the following manner.

$$D_c(r, s) = \frac{N_{rs}}{N-1} r, \quad s \in \{A, C, D, \ldots Y\}, \tag{1}$$

The dipeptide composition of a di-peptide 'rs', can be computed using $D_c(r,s)$, N_{rs} represents the count of dipeptides expressed by the amino acid categories r and s. The size of the peptide sequence can be denoted by 'N' [20]. The theoretical mean $T_m(r, s)$ is determined by:

$$T_m(r, s) = \frac{C_r}{C_N} \times \frac{C_s}{C_N}, \tag{2}$$

where C_r represents the count of codons, used to encode the initial amino acid, and C_s signifies the count of codons used for 2nd amino acid in a specific dipeptide rs. Additionally, the overall number of condons is represented by C_N, without the three termination codons [20]. The theoretical variance, $T_v(r, s)$, of the dipeptide 'rs' is determined by:

Lastly, $DDE(r, s)$ can be calculated as follows:

$$DDE(r, s) = \frac{D_c(r, s) - T_m(r, s)}{\sqrt{T_v(r, s)}} \tag{3}$$

The DDE attribute has been effectively utilized in the prediction of B-cell epitopes [21].

3) ENHANCED GROUPED AMINO ACID COMPOSITION (EGAAC)

EGAAC is an extension of the Group-based Amino Acid Composition (GAAC) method. EGAAC calculates the GAAC within sliding windows of a fixed size, continuously shifting from the N-terminal to the C-terminal of each amino acid sample. It is commonly used for protein samples with uniform lengths [22], [23].

The representation of EGAAC can be computed as follows:

$$f(g, win) = \frac{N(g, win)}{N(win)}, \quad g \in \{g1, g2, g3, g4, g5\},$$
$$win \in \{window1, window2, \ldots . window17\} \tag{4}$$

TABLE 1. Attribute classification.

	Charge	Hydrophobicity
C1	***±ve*** R,K	***Polar*** D, K, E, Q, N, R
C2	***Not +ve nor –ve*** A, C, Q, N H, I, L, G F, P, S, M, W, V, Y, T	***Not +ve nor –ve*** A, H, G, S, P, T, Y
C3	***–ve*** D, E	Hydrophobic C, I, F, V, M, W

In this approach, the count of amino acid samples belonging to group 'g' inside the sliding window '*win*' is represented by *N(g, win)*, while the '*win*' is denoted as *N(win)*. The EGAAC descriptor has been successfully employed in the prediction of lysine crotonylation sites [24].

4) COMPOSITION/TRANSITION/DISTRIBUTION DESCRIPTOR (CTD)

CTD analyzes the overall structure of a peptide sequence, including the arrangement of amino acids and the occurrences of two different adjacent amino acids. The major task of CTD is to arrange the sequence to calculate the structure, transition, and arrangement [25]. Based on their characteristics, amino acids are classified into three groups (group 1, group 2, and group 3) [26], also known as reduced amino acids [27], [28]. These classifications are related to charge and hydrophobicity are presented in Table 1. C/T/D, characterized by composition C_a, transition T_b, and distribution $D_{b,z}$, is defined.

$$C_a = \frac{N_a}{N} (\because a = 1, 2, 3 \ldots) \tag{5}$$

$$T_b = \frac{N_{b,c} + N_{c,b}}{N-1} (\because b = 1, 2, 3 \ldots, c \neq b) \tag{6}$$

$$D_{b,z} = \frac{N_{b,z}}{N} (\because b = 1, 2, 3 \ldots, z = 1, 0.15N \ldots, N) \tag{7}$$

In this context, N_a represents the category identifier, $N_{b,c}$ denotes the count of instances where class b and c are adjacent to each other, and $N_{b,z}$ indicates the count of amino acids that belong to the z-th position within the b-th class.

C. CLASSIFICATION MODELS

1) DEEP NEURAL NETWORK (DNN)

Deep learning algorithms are essential to train and classify complex and nonlinear functions represented by biological sequences [29], [30], [31]. In this study, to effectively train

TABLE 2. Hyper parameters of the DNN model.

Parameter	Optimal Value
Hidden Layers	4
Dropout	0.3
Activation Function	ReLu, Sigmoid
Learning Rate	0.001
Lasso Regularization	0.01
Batch size	32,64
Optimizer	Adam
Dense Hidden Layers	2
Weight initialization	Xavier

a model three distinct encoding techniques were utilized to transform the biological sequences into numerical structures. The DNN structure is comprised of an initial layer, four hidden layers, and an output layer [32], [33]. The initial layer takes the extracted features of different encoding methods and provides them to the hidden layers. Whereas, the hidden layers process the input neurons using different non-linear functions, enabling the network to learn the complex task effectively. Hidden layers are also an iterative procedure that continues from one hidden layer to the next until the final hidden layer is reached. The final layer consists of a single neuron that produces the output based on processed data using a sigmoid activation function. Which helps to map the output class labels to either 0 or 1. Consequently, the final output layer will predict either 1 for AtbP or 0 for non-AtbPs.

In deep learning literature, various linear and non-linear activation functions were applied to enable the neural network to effectively learn complex patterns. In our study, we focused on the Rectified Linear Unit (ReLU) activation function for evaluating the effectiveness of the proposed model. Additionally, for appropriate initialization of the weights to the neural network, the Xavier function was employed for stable and efficient learning. By appropriate scaling of the initial weights, the Xavier function helps to avoid the exploding or vanishing gradients issues during the training process [34]. Hence, for a smooth and reliable learning process, we used a learning rate of 0.001. To thoroughly assess the DNN model and to handle the overfitting issue, regularization techniques such as L1 regularization (LASSO), and a dropout value of 0.3 are applied [32]. To ensure model generalization, a diverse set of independent datasets are also employed. The optimized hyperparameters for the DNN model are presented in Table 2.

2) EXTRA TREE CLASSIFIER (ETC)

ETC is a kind of ensemble classification method that belongs to the family of RF algorithms [35]. That combines multiple decision trees to perform prediction tasks. ETC is considered more effective due to its computational efficiency and capability to handle high-dimensional data by creating an ensemble of decision trees. Each tree is trained through a random selection of the subset from the training samples. Throughout the training process, the decision trees are constructed by randomly choosing feature thresholds to partition the data at each node [36]. The final forecast of the ensemble is determined by averaging or voting on the prediction of individual trees. Compared to traditional decision trees and other ensemble methods like random forests, the ETC introduces additional randomness in the tree construction process [37], [38]. Which helps to reduce overfitting and improve generalization performance. Furthermore, to train faster than any other method the extra tree classification is capable of using parallel computing power.

3) RANDOM FOREST (RF)

Random Forest (RF) is an ensemble classifier originally introduced by Breiman [39]. That has been effectively applied to various tasks, such as clustering, feature selection, regression, and classification [40], [41]. The RF algorithm consists of an ensemble of decision trees, collectively forming a forest. Each tree undergoes training using a subset of the training samples from the dataset. Subsequently, the RF model assigns a class label to a new sample by employing a majority voting scheme [42], [43]. One important advantage of RF lies in its ability to address the challenges of high variance or bias associated with individual trees, thereby ensuring that the overall performance of the model remains unaffected [44], [45]. This can be obtained by incorporation of a weighting scheme within RF, where lower weights are assigned to trees exhibiting higher error rates. Consequently, the overall predictive capability of the ensemble is enhanced, leading to improved performance. RF is particularly well-suited for handling large datasets, effectively managing missing data, and detecting outliers.

4) ENSEMBLE LEARNING

Ensemble classification is widely adopted by researchers in the field of computational bioinformatics due to its ability to achieve outstanding classification results and facilitate effective generalization [1], [46], [47], [48], [49], [50]. The primary objective of an ensemble classifier is to fuse the individual classifiers, thus creating a more reliable and intelligent learning model that enhances predictive outcomes while minimizing errors [51]. Moreover, incorporating an ensemble strategy in classical machine learning approaches proves particularly advantageous as it reduces variance arising from inconsistent prediction rates of conventional classifiers [52], [53]. Consequently, scientists have extensively employed ensemble learning methods across diverse domains over the last few years, encompassing topics such as neuro-peptides [54], protein subcellular localization [55], antiviral peptides [56], anti-cancer peptides [34], anti-fungal peptides [57], recombination spots [58], and malaria parasite [59]. In our

study, we utilized an ensemble learning method using an optimized genetic algorithm (GA) to assess the predictive outcomes of the composite features. GA is a heuristic approach that effectively tackles classification problems with a high degree of success [60]. GA involves a random strategy to select a subset of the whole population and then apply different genetic functions, i.e., crossover, selection, and mutation, to produce the best predictive results [12], [61]. Firstly, we calculated the predicted labels of the proposed encoding schemes using three learning algorithms: ERT, RF, and DNN. Subsequently, the predicted labels from each classifier were given to GA to construct an ensemble learning method, as elaborated below:

$$EnC = RF \oplus ET \oplus DNN \tag{8}$$

In equation (8), the symbol EnC signifies the collective learning model, while $\oplus$ representing the merging operator utilized to combine the predicted labels from individual classifiers. The Ensemble methodology, referred as EnC employs a diverse range of algorithms with distinct characteristics. Let us discuss in detail: Consider a classifier 'S' designed for analyzing a sequence of peptides denoted as 'P':

$$\{S_i, S_{ii}, S_{iii}\} \in \{C_a, C_b\} \tag{9}$$

In this context, the symbols S_i, S_{ii}, S_{iii} represent individual classifiers, while corresponding to the desired labels, namely AtbP and non-AtbPs.

$$Y_i = \sum_{j=1}^{3} \delta(S_i C_j), \quad (i = 1, 2) \tag{10}$$

$$\delta(S_i C_j) = \left\{ \begin{array}{ll} 1, & if\ S_i \in C_i \\ 0, & otherwise \end{array} \right\} \tag{11}$$

Lastly, the prediction rates of the ensemble learning algorithm, EnC, using GA, are evaluated as follows:

$$EnC_i = Max\{w_i y_i, \ldots..w_{iii} y_{iii}\} \tag{12}$$

In this scenario, EnC_i denotes the ensemble learning method based on a Genetic Algorithm. The term 'Max' indicates the highest classification rate achieved, while $w_i y_i, \ldots..w_{iii} y_{iii}$ representing the optimal weights assigned to each hypothesis learner.

D. PERFORMANCE EVALUATION PARAMETERS

To evaluate the performance of the constructed models, we utilized four widely employed prediction metrics usually applied in binary classification problems [62], [63], [64]: sensitivity, specificity, accuracy, and Matthews's correlation coefficient (MCC). The computation of these metrics was carried out according to the following formulas:

$$Accuracy = 1 - \frac{AtbP_{-}^{+} + AtbP_{+}^{-}}{AtbP^{+} + AtbP^{-}} \tag{13}$$

$$Sensitivity = 1 - \frac{AtbP_{-}^{+}}{AtbP^{+}} \tag{14}$$

$$Specificity = 1 - \frac{AtbP_{+}^{-}}{AtbP^{-}} \tag{15}$$

$$MCC = \frac{1 - \left(\frac{AtbP_{-}^{+} + AtbP_{+}^{-}}{AtbP^{+} + AtbP^{-}}\right)}{\sqrt{\left(1 + \frac{AtbP_{+}^{-} + AtbP_{-}^{+}}{AtbP^{+}}\right)\left(1 + \frac{AtbP_{-}^{+} + AtbP_{+}^{-}}{AtbP^{-}}\right)}} \tag{16}$$

where $AtbP^{+}$ represents the true-positive prediction of the AtbPs and $AtbP^{-}$ denotes the true-negative prediction of the non-AtbPs. Likewise, $AtbP_{+}^{-}$ demonstrates the errors of the model in terms of predicting the false-negative, the sequences that the model falsely predicted as true. Similarly, $AtbP_{-}^{+}$ shows the error of the model in terms of false negatives that are true.

III. RESULTS AND DISCUSSIONS

In this study, we outcomes of the predicted model are evaluated using a 10-fold Cross-validation (CV) test. Additionally, to generate reliable and accurate results, the mean value of the 10-fold CV was reported by repeating the Stratified loop strategy 30 times [57], [65]. Initially, the peptide sequences were formulated via DDE, EGAAC, AAindex, and CTD-based sequential and structured properties vectors. Then, a hybrid vector is formed by fusing the extracted features to cover the limitation of the individual feature vectors. The prediction analysis of the model was examined using the GA-based ensemble model. The predictive results of the training and independent datasets using the different classification algorithms are discussed in the subsections below.

A. PERFORMANCE OF pAtbP-ENC VIA MD TRAINING DATASET

The predictive results of the MD training samples by applying various classification models are presented in Table 4. As described in the II (A) section, the positive sequences of both RD and MD datasets are the same, however, only the negative sequences are different. At first, we trained our proposed model using the MD training sequences. In terms of single feature space, the ensemble genetic algorithm employed CTD, leading to an enhanced sensitivity of 94.85% and improved values of accuracy, specificity, and MCC, reaching 93.07%, 91.18%, and 0.86, respectively. Similarly, GA reported an accuracy of 92.32%, 91.92%, and 91.47% using EGAAC, DDE, and AAindex, respectively. Subsequently, we combined the extracted descriptors of the applied encoding methods to construct a heterogeneous vector. After reporting predictive rates for all the classifiers, Ensemble-GA

achieved remarkable outcomes, with an accuracy of 95.13%, sensitivity of 96.45%, specificity of 93.83%, and MCC of 0.90. The improved performance rates of the heterogeneous vector happen due to high training power and discriminative descriptors of AAIndex, DDE, and EGAAC. Consequently, the combined vector demonstrated superior outcomes as compared to the individual.

B. PERFORMANCE OF pAtbP-ENC VIA RD TRAINING DATASET

On the other hand, the RD training dataset is also utilized for purpose and to effectively assess their predictive potential. The predictive evaluation of the applied classifiers learners using RD-training samples is presented in Table 3. We applied several learning models including RF, ET, DNN, and GA-based ensemble classifier, to evaluate the predictive results of the extracted features. Similar to the MD training dataset, the proposed learning algorithms were used to evaluate the outcomes of RD training using an individual encoding scheme. Additionally, we combined all the descriptors from different encoding methods into a heterogeneous vector with a dimension size of 1056. The heterogeneous vector was also measured with the proposed classifiers. After thoroughly comparing the predictive outcomes of all the classifiers, it was found that the ensemble classifier surpasses than single classifiers. Especially, in individual DDE vectors having a dimension of 400D, ensemble-GA attains a higher sensitivity of 93.49% and demonstrates remarkable values of 91.84% accuracy, 90.19% specificity, and 0.83 MCC, respectively.

Conversely, our ensemble classifier by employing multi-perspective features performed outstanding well by achieving a prediction accuracy of 97.80%, along with a specificity of 96.80%, a sensitivity of 98.81%, and an MCC of 0.96. We choose an optimized ensemble classifier as a proposed training model due to its promising predictive results using RD and MD training samples. The proposed ensemble model combines the predicted labels of the applied single classifiers which was used to form an optimized using genetic algorithm and performs better than individual algorithms.

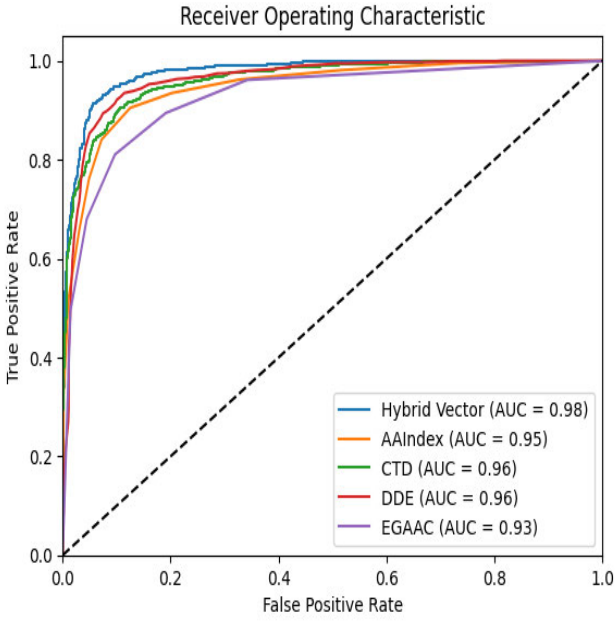


FIGURE 2. ROC analysis of RD training sample.

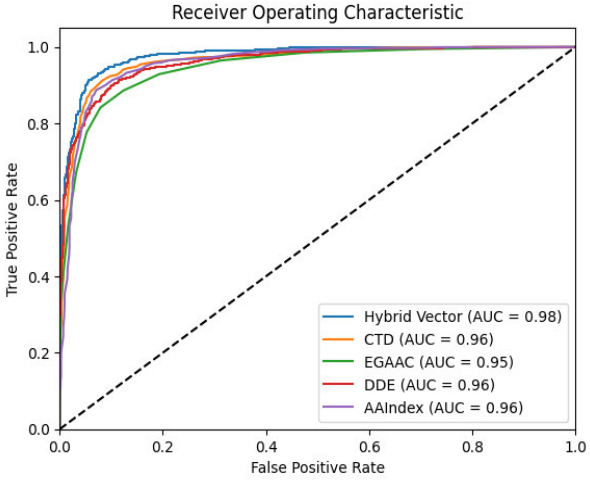


FIGURE 3. ROC analysis of MD training sample.

C. PERFORMANCE OF pAtbP-ENC ON INDEPENDENT DATASETS

To evaluate the effectiveness and generalization power of our proposed pAtbP-EnC model, we applied two different independent datasets as discussed in section II-A. The predictive assessment of MD-independent and RD-independent datasets using the proposed model is illustrated in Table 5. For instance, using the MD-independent dataset, we achieved an exceptional level of accuracy of 94.17% which shows the robustness and reliability of the model using unseen sequences. In addition, other performance metrics reported excellent results such as sensitivity, specificity, and MCC of 98.69%, 92.63%, and 0.88 respectively. Similarly, in the case of RD- RD-independent samples our ensemble learning model obtained better accuracy of 93.91%, a sensitivity of 95.60%, specificity of 92.24%, and MCC of 0.87, respectively. These outcomes show that our model can be effectively used to discriminate between positive and negative classes of unseen AtbP samples. Additionally, we calculated the area under the receiver operating characteristic curves (AuROCs) and generated ROC curves for the proposed method using all training and independent datasets, as illustrated in Figures 2–5. On the other hand, the precision-recall curves are calculated against the training and the independent dataset as shown in Figure 7(A-D). Hence, the evaluation of our model on the independent datasets confirms its robustness and capacity to discriminate peptide samples appropriately.

TABLE 3. Performance results of classification learners using RD training dataset.

Encoding Method	Classifier	ACC%	Sen%	Spe%	MCC
AAindex	RF	78.08	70.70	85.42	0.56
	ETC	79.58	76.76	82.41	0.59
	DNN	81.08	73.23	88.94	0.62
	GA	88.94	87.65	90.23	0.77
DDE	RF	80.84	72.22	89.44	0.60
	ETC	81.35	76.26	86.43	0.63
	DNN	82.85	77.27	88.44	0.66
	GA	91.84	93.49	90.19	0.83
EGAAC	RF	81.61	78.28	84.92	0.63
	ETC	82.37	79.29	85.42	0.64
	DNN	84.11	81.31	86.93	0.68
	GA	86.46	88.65	84.56	0.73
CTD	RF	77.81	76.26	79.39	0.55
	ETC	80.08	77.77	82.41	0.60
	DNN	82.12	77.27	86.93	0.64
	GA	91.42	90.48	92.37	0.82
Hybrid Features	RF	87.22	87.39	87.06	0.74
	ETC	87.14	86.96	87.31	0.74
	DNN	93.74	94.62	92.90	0.87
	GA	**97.80**	**98.81**	**96.80**	**0.96**

TABLE 4. Performance results of classification learner using MD training dataset.

Encoding Method	Classifier	Acc (%)	Sn (%)	Sp (%)	MCC
AAindex	RF	78.84	68.68	88.94	0.58
	ETC	79.09	71.12	86.93	0.58
	DNN	86.03	85.88	86.51	0.72
	GA	91.47	94.14	88.67	0.82
DDE	RF	77.57	75.25	79.89	0.55
	ETC	78.82	77.27	79.39	0.56
	DNN	89.75	90.81	88.71	0.79
	GA	91.92	94.38	89.33	0.83
EGAAC	RF	76.56	69.19	81.40	0.50
	ETC	78.10	75.75	80.40	0.56
	DNN	89.14	92.26	86.58	0.78
	GA	92.32	95.26	89.23	0.84
CTD	RF	75.08	71.71	78.39	0.50
	ETC	76.33	73.73	78.89	0.52
	DNN	89.74	91.07	88.77	0.79
	GA	93.07	94.85	91.18	0.86
Hybrid Features	RF	87.22	87.39	87.06	0.74
	ETC	89.45	89.62	89.28	0.78
	DNN	93.68	93.99	93.39	0.87
	GA	**95.13**	**96.45**	**93.83**	**0.90**

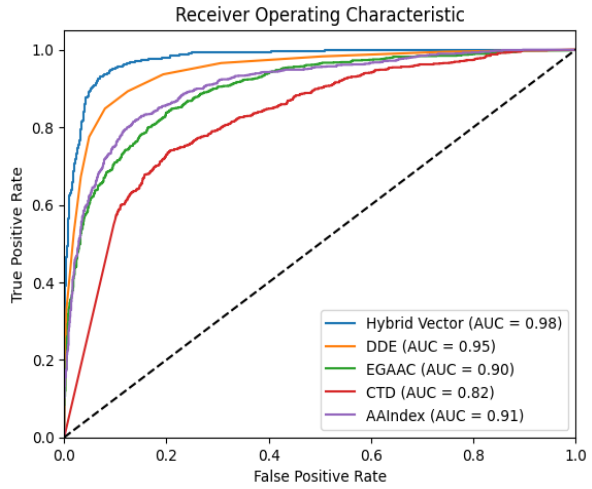


FIGURE 4. ROC analysis of RD independent dataset.

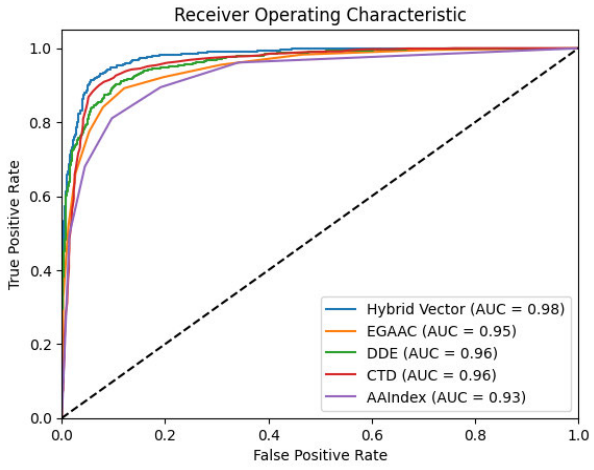


FIGURE 5. ROC analysis of MD independent dataset.

TABLE 5. Prediction analysis of heterogeneous feature vector using RD and MD independent datasets.

Dataset	Model	ACC%	Sen%	Spe%	MCC
RD-Independent	RF	85.13	80.30	89.94	0.70
	ETC	86.13	81.31	90.95	0.72
	DNN	90.66	89.33	93.77	0.81
	GA	**93.91**	**95.60**	**92.24**	**0.87**
MD-Independent	RF	74.09	76.30	72.07	0.48
	ETC	74.24	72.83	75.60	0.48
	DNN	87.59	90.88	84.89	0.75
	GA	**94.17**	**95.69**	**92.63**	**0.88**

D. MODEL INTERPRETATION

In this section, we performed the interpretation of the proposed model interpretation using the Shapley Additive Explanation Algorithm (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) analysis [66], [67], [68]. These methods investigate the contribution of the extracted features by visualizing the high contributory features from the whole feature set using machine learning

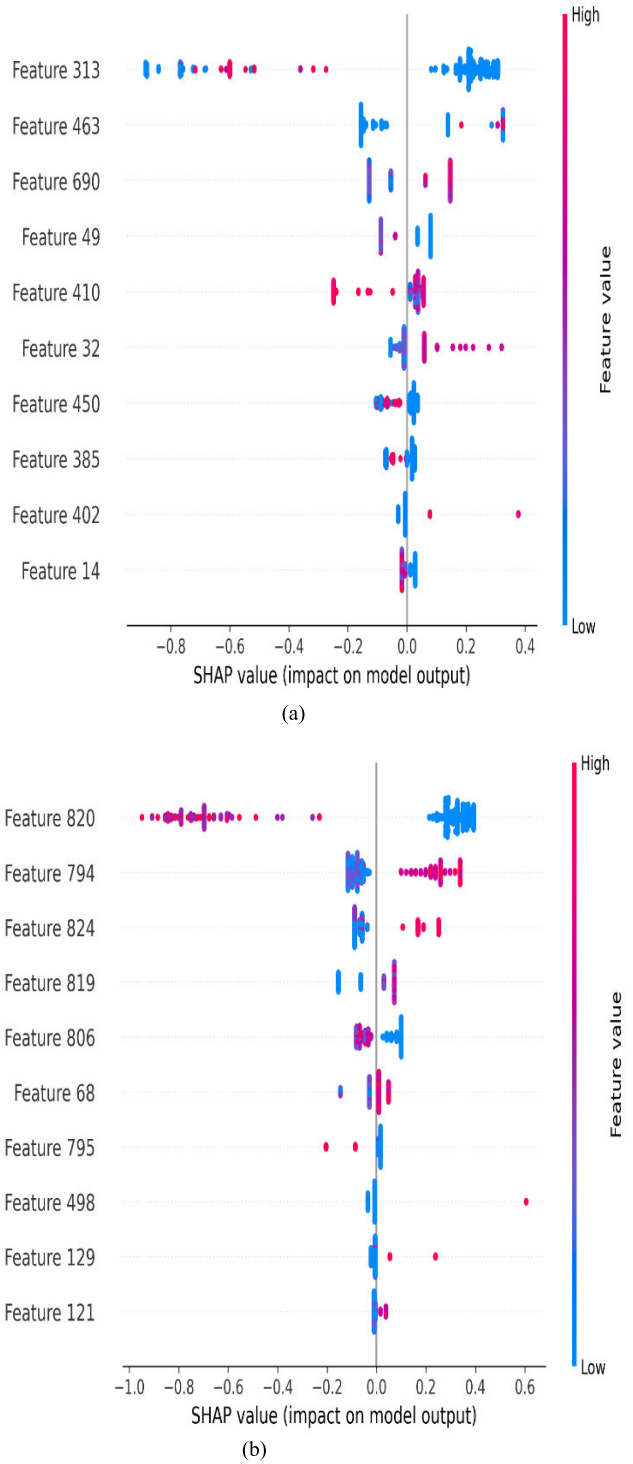


FIGURE 6. (a). SHAP analysis for contributory features using MD training dataset. (b). SHAP analysis for contributory features using RD training dataset.

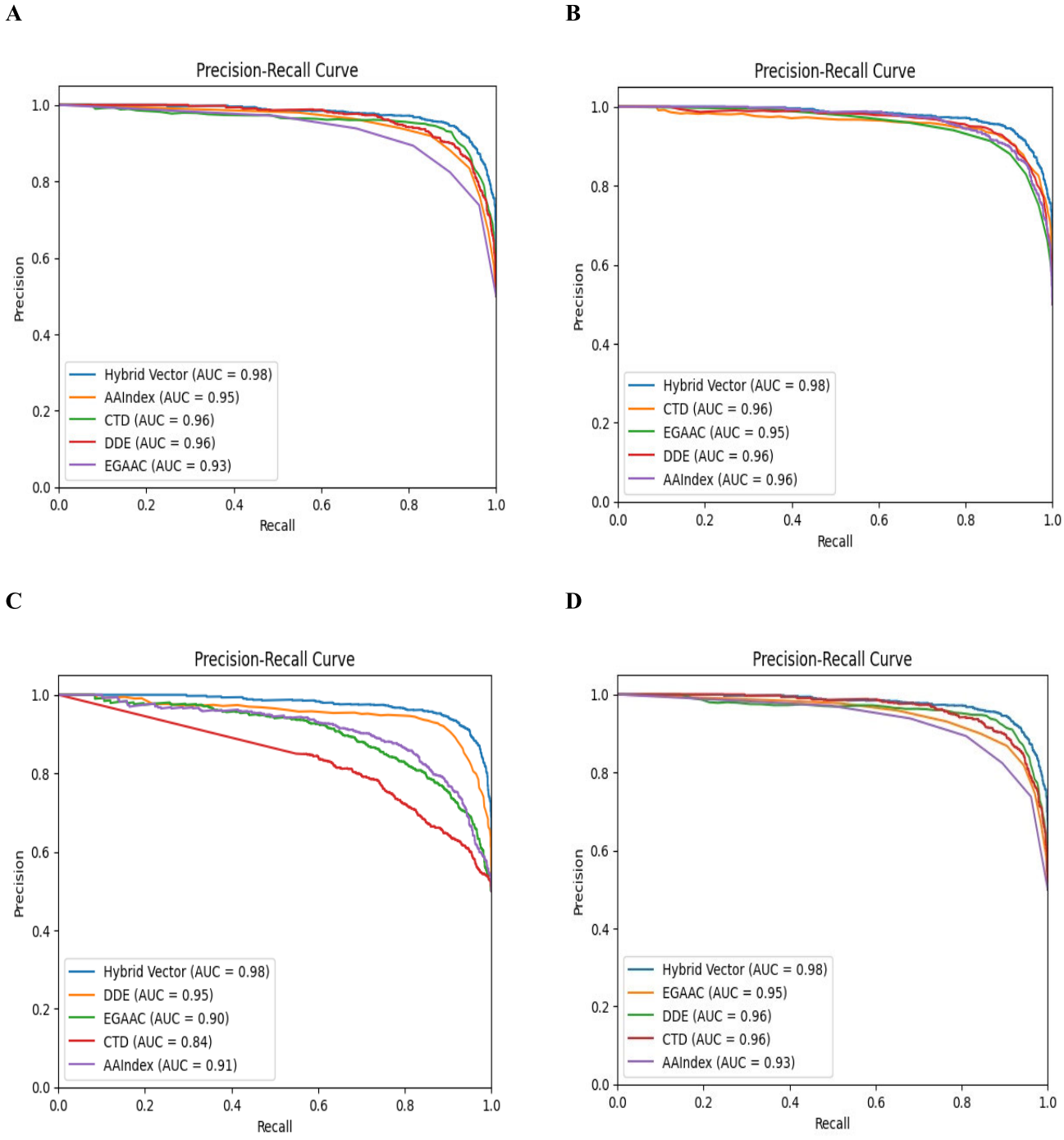
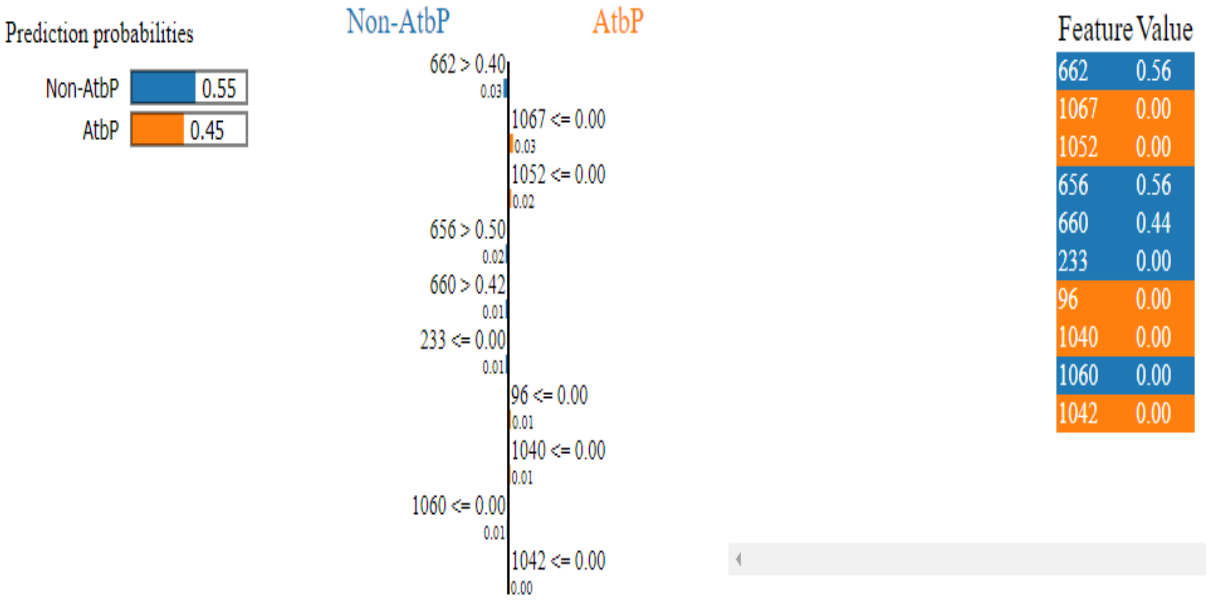
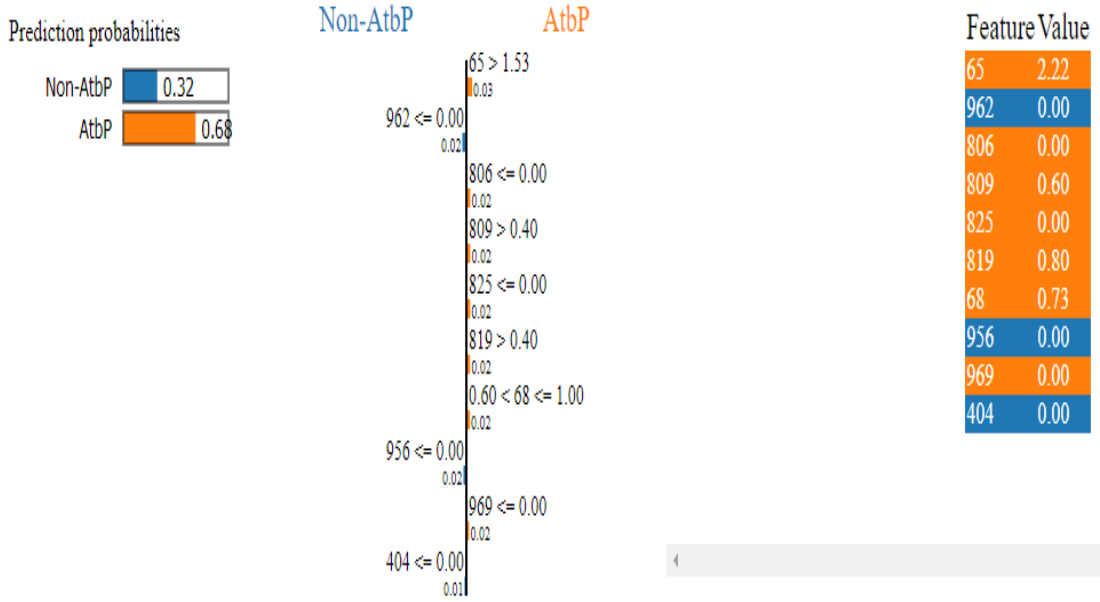


FIGURE 7. Precision-Recall analysis for (A) RD Training Dataset (B) MD training dataset (C) RD independent dataset (D) MD independent dataset.

algorithms. SHAP is a global interpolation scheme to measure the contribution of each feature via aggregating its shapely values [69]. The SHAP analysis of 10 contributory features using MD training and RD training datasets are illustrated in Figures 6(A), and 6(B). Each row in the figure represents the SHAP value distribution for a specific feature. Data point colors indicate feature values, with red denoting higher values and blue lower ones. Colored dots signify each feature's influence on the model output. Whereas, SHAP > 0 indicates a positive prognosis (AtbPs), while SHAP < 0 suggests a negative prognosis (non-AtbPs). Our findings emphasize the significance of the Hybrid feature, highlighting the importance of predicting the targeted classes.



(a)



(b)

FIGURE 8. **(a). LIME-based instance analysis using an MD-independent dataset. (b). LIME-based instance analysis using RD independent dataset.**

Additionally, LIME analysis is also applied to represent the significance of the model per instance [70]. LIME simplifies models through feature matrix permutations. A pivotal aspect of LIME is the construction of a similarity matrix, measuring distances between query sequences and perturbed sequences. In this work, we applied the LIME analysis on the test samples as shown in Figures 8(A), and 8(B) for MD-independent and RD-independent datasets. LIME analysis discriminates the input instance based on its correlation with AtbPs (orange color) and non-AtbPs (blue color).

E. PERFORMANCE COMPARISON OF THE pAtbP-ENC MODEL WITH EXISTING METHODS

The performance analysis of the pAtbP-EnC model is compared with currently available models as displayed in Table 6, and Figures 9-11. Which provides an overview of the existing computational methods applied for the prediction of AtbPs. As mentioned in the literature, the existing studies relied heavily on feature representation approaches based on typical machine learners. Such as the Antitbpred model [8] developed a novel approach, which integrates strategies

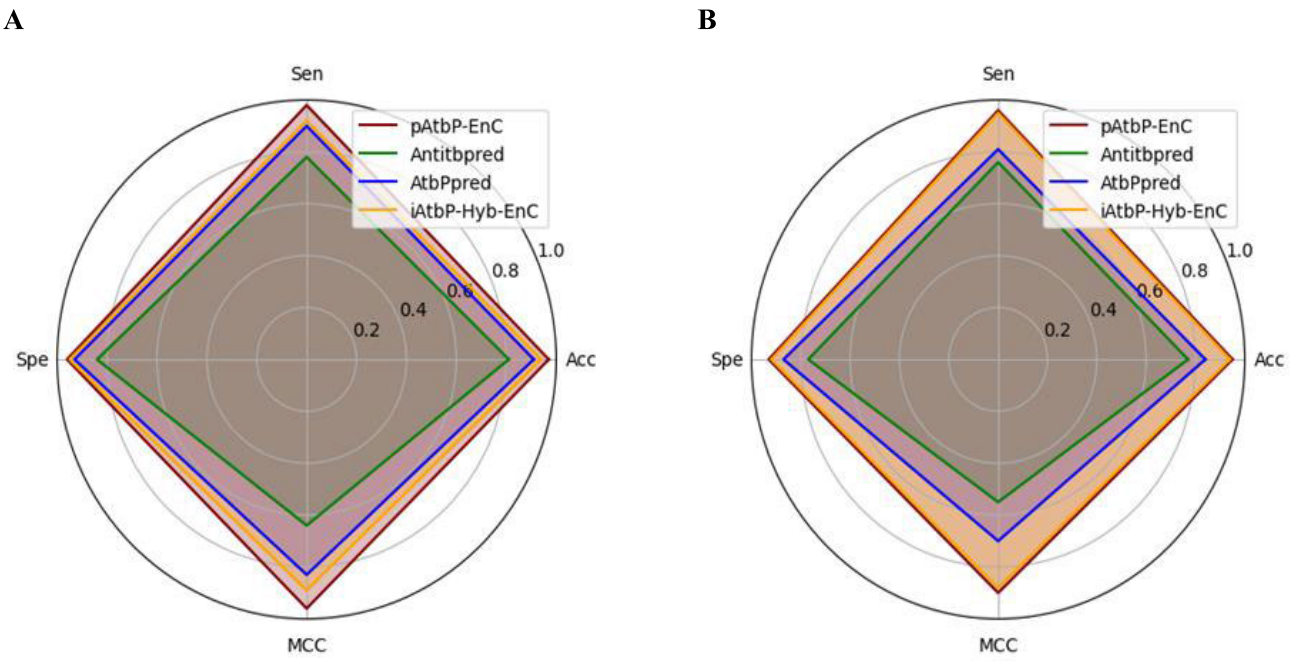


FIGURE 9. Comparison analysis of pAtbP-EnC with existing predictors (A) RD training dataset (B) MD training dataset.

TABLE 6. Comparison of proposed model with existing methods.

Dataset	Model	ACC%	Sen%	Spe%	MCC
RD-Training	Antitbpred [8]	81.70	78.70	84.60	0.64
	AtbPpred [10]	91.70	90.50	93.01	0.83
	iAtbP-Hyb-EnC [12]	94.47	92.96	95.97	0.89
	pAtbP-EnC	**97.80**	**98.81**	**96.80**	**0.96**
MD-Training	Antitbpred [8]	77.01	76.80	77.30	0.55
	AtbPpred [10]	84.90	81.90	87.90	0.70
	iAtbP-Hyb-EnC [12]	94.22	95.60	92.81	0.88
	pAtbP-EnC	**95.13**	**96.45**	**93.83**	**0.90**
RD-Independent	Antitbpred [8]	78.50	73.30	83.80	0.57
	AtbPpred [10]	85.10	80.90	89.40	0.70
	iAtbP-Hyb-EnC [12]	92.55	93.04	91.87	0.85
	pAtbP-EnC	**93.91**	**95.60**	**92.24**	**0.87**
MD-Independent	Antitbpred [8]	75.90	75.01	76.70	0.52
	AtbPpred [10]	89.40	83.01	95.70	0.79
	iAtbP-Hyb-EnC [12]	92.68	95.24	92.96	0.89
	pAtbP-EnC	**94.17**	**95.69**	**92.63**	**0.88**

such as sequential peptide approaches and a binary pattern strategy. After training the RD dataset, the Antitbpred model achieved an accuracy rate of 81.70%. Subsequently, Antitbpred on RD independent samples reported an accuracy of 78.50%. Similarly, the AtbPpred applied sequential feature representation for training RD and MD samples [10]. A 91.70% training accuracy was achieved via the MD training dataset. Additionally, the AtbPpred predictor demonstrated a prediction accuracy of 85.10% on RD training samples and an accuracy of 89.40% via MD-independent samples. Recently, Akbar proposed an ensemble classification model called iAtbP-Hyb-EnC [12], which achieved prediction accuracies of 94.47% and 92.22% on RD and MD training datasets, respectively. On the testing datasets,

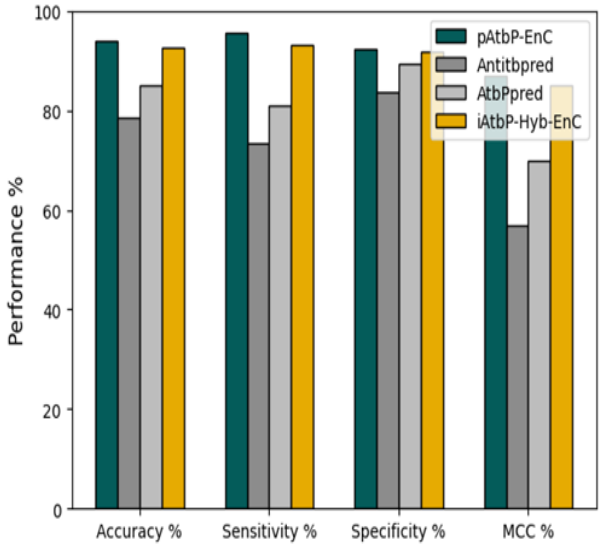


FIGURE 10. Comparison analysis of pAtbP-EnC with existing predictors using RD independent dataset.

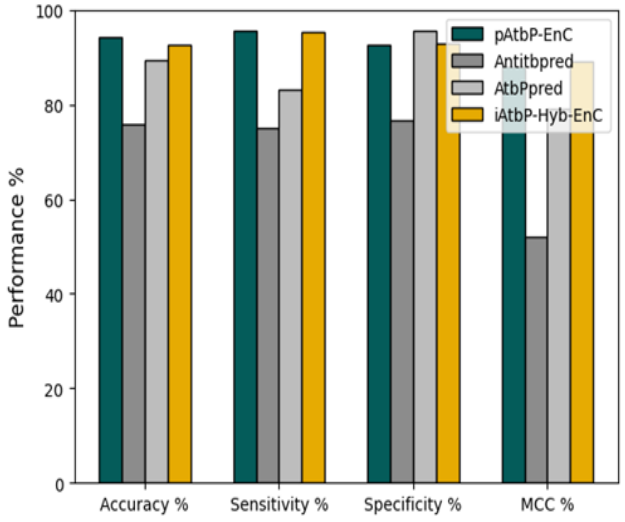


FIGURE 11. Comparison analysis of pAtbP-EnC with existing predictors using MD independent dataset.

iAtbP-Hyb-EnC achieved accuracies of 92.55% and 92.68% on both independent datasets. In contrast, our pAtbP-EnC model using a multi-varied feature vector achieved the highest training accuracy of 97.80% using RD training samples, with a sensitivity of 98.81%, specificity of 96.80%, and MCC of 0.96. It was observed that the pAtbP model outperformed all the computational models i.e., Antitbpred predictor, AtbPpred model, and iAtbP-Hyb-EnC via both RD and MD training sequences. Furthermore, while measuring the MD and RD independent samples, the pAtbP-EnC model using all predictive metrics, significantly performed than the AtbPpred, Antitbpred, and iAtbP-Hyb-EnC models as shown in Figure 10 and Figure 11.

IV. CONCLUSION

This study presents a reliable and precise computational method for effectively discriminating Antitubercular peptides. The proposed comprises several phases to ensure accurate predictions. Firstly, we formulate the peptide samples using four diverse nature formulation techniques: AAindex, DDE, CTD, and EGAAC. These methods enable transforming the peptide samples into numerical information that is highly mandatory for the training machine learning model. Hence, various individual feature vectors and a hybrid vector are generated. The hybrid vector thoroughly represents the numerous information of all encoding methods used for the prediction of AtbP sequences. Finally, we develop an ensemble classification algorithm by concatenating the predicted outcomes of the applied classifiers using an optimized genetic algorithm. Which leads to overcoming the limitations of the individual classifiers. Especially, an ensemble learner by incorporating hybrid features performed remarkably well in all evaluation metrics using training and independent datasets. Our pAtbP-EnC model will serve as a valuable tool in the field of drug discovery and advancing academic research in this domain.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] A.-L. Bañuls, A. Sanou, N. T. Van Anh, and S. Godreuil, "Mycobacterium tuberculosis: Ecology and evolution of a human bacterium," *J. Med. Microbiol.*, vol. 64, no. 11, pp. 1261–1269, Nov. 2015.

[2] N. Mandal, P. K. Anand, S. Gautam, S. Das, and T. Hussain, "Diagnosis and treatment of paediatric tuberculosis: An insight review," *Crit. Rev. Microbiol.*, vol. 43, no. 4, pp. 466–480, 2017.

[3] A. Khusro, C. Aarti, and P. Agastian, "Anti-tubercular peptides: A quest of future therapeutic weapon to combat tuberculosis," *Asian Pacific J. Tropical Med.*, vol. 9, no. 11, pp. 1023–1034, Nov. 2016.

[4] L. Pinto and D. Menzies, "Treatment of drug-resistant tuberculosis," *Infection Drug Resistance*, vol. 4, pp. 129–135, 2011.

[5] K. Kaushik, A. Bhardwaj, S. Bharany, N. Alsharabi, A. U. Rehman, E. T. Eldin, and N. A. Ghamry, "A machine learning-based framework for the prediction of cervical cancer risk in women," *Sustainability*, vol. 14, no. 19, p. 11947, Sep. 2022.

[6] A. Khusro, C. Aarti, A. Barbabosa-Pliego, and A. Z. M. Salem, "Neoteric advancement in TB drugs and an overview on the anti-tubercular role of peptides through computational approaches," *Microbial Pathogenesis*, vol. 114, pp. 80–89, Jan. 2018.

[7] T. Teng, J. Liu, and H. Wei, "Anti-mycobacterial peptides: From human to phage," *Cellular Physiol. Biochem.*, vol. 35, no. 2, pp. 452–466, 2015.

[8] S. S. Usmani, S. Bhalla, and G. P. S. Raghava, "Prediction of antitubercular peptides from sequence information using ensemble classifier and hybrid features," *Frontiers Pharmacol.*, vol. 9, p. 954, Aug. 2018.

[9] S. Khatun, M. Hasan, and H. Kurata, "Efficient computational model for identification of antitubercular peptides by integrating amino acid patterns and properties," *FEBS Lett.*, vol. 593, no. 21, pp. 3029–3039, Nov. 2019.

[10] B. Manavalan, S. Basith, T. H. Shin, L. Wei, and G. Lee, "AtbPpred: A robust sequence-based prediction of anti-tubercular peptides using extremely randomized trees," *Comput. Struct. Biotechnol. J.*, vol. 17, pp. 972–981, 2019.

[11] W. Chen, P. Feng, and F. Nie, "iATP: A sequence based method for identifying anti-tubercular peptides," *Medicinal Chem.*, vol. 16, no. 5, pp. 620–625, 2020.

[12] S. Akbar, A. Ahmad, M. Hayat, A. U. Rehman, S. Khan, and F. Ali, "iAtbP-Hyb-EnC: Prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model," *Comput. Biol. Med.*, vol. 137, Oct. 2021, Art. no. 104778.

[13] A. Sundas, S. Badotra, S. Bharany, A. Almogren, E. M. Tag-ElDin, and A. U. Rehman, "HealthGuard: An intelligent healthcare system security framework based on machine learning," *Sustainability*, vol. 14, no. 19, p. 11934, Sep. 2022.

[14] The UniProt Consortium, "UniProt: The universal protein knowledgebase," *Nucleic Acids Res.*, vol. 45, pp. D158–D169, Jan. 2017.

[15] S. S. Usmani, R. Kumar, V. Kumar, S. Singh, and G. P. S. Raghava, "AntiTbPdb: A knowledgebase of anti-tubercular peptides," *Database*, vol. 2018, Jan. 2018, Art. no. bay025.

[16] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: Accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012.

[17] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, "AAindex: Amino acid index database, progress report 2008," *Nucleic Acids Res.*, vol. 36, pp. D202–D205, Dec. 2007.

[18] C.-W. Tung and S.-Y. Ho, "Computational identification of ubiquitylation sites from protein sequences," *BMC Bioinf.*, vol. 9, no. 1, pp. 1–15, Dec. 2008.

[19] Z. Chen, N. He, Y. Huang, W. T. Qin, X. Liu, and L. Li, "Integration of a deep learning classifier with a random forest approach for predicting malonylation sites," *Genomics, Proteomics Bioinf.*, vol. 16, no. 6, pp. 451–459, Dec. 2018.

[20] G. White and W. Seffens, "Using a neural network to backtranslate amino acid sequences," *Electron. J. Biotechnol.*, vol. 1, no. 2, pp. 196–201, Dec. 1998.

[21] V. Saravanan and N. Gautham, "Harnessing computational biology for exact linear B-cell epitope prediction: A novel amino acid composition-based feature descriptor," *OMICS, J. Integrative Biol.*, vol. 19, no. 10, pp. 648–658, Oct. 2015.

[22] Z. Chen, P. Zhao, F. Li, T. T. Marquez-Lago, A. Leier, J. Revote, Y. Zhu, D. R. Powell, T. Akutsu, G. I. Webb, K.-C. Chou, A. I. Smith, R. J. Daly, J. Li, and J. Song, "iLearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data," *Briefings Bioinf.*, vol. 21, no. 3, pp. 1047–1057, May 2020.

[23] Z. Chen, P. Zhao, F. Li, A. Leier, T. T. Marquez-Lago, Y. Wang, G. I. Webb, A. I. Smith, R. J. Daly, K.-C. Chou, and J. Song, "iFeature: A Python package and web server for features extraction and selection from protein and peptide sequences," *Bioinformatics*, vol. 34, no. 14, pp. 2499–2502, Jul. 2018.

[24] Y. Zhao, N. He, Z. Chen, and L. Li, "Identification of protein lysine crotonylation sites by a deep learning framework with convolutional neural networks," *IEEE Access*, vol. 8, pp. 14244–14252, 2020.

[25] I. Dubchak, I. Muchnik, S. R. Holbrook, and S. H. Kim, "Prediction of protein folding class using global description of amino acid sequence," *Proc. Nat. Acad. Sci. USA*, vol. 92, no. 19, pp. 8700–8704, Sep. 1995.

[26] Y. Zuo, Y. Li, Y. Chen, G. Li, Z. Yan, and L. Yang, "PseKRAAC: A flexible web server for generating pseudo K-tuple reduced amino acids composition," *Bioinformatics*, vol. 33, no. 1, pp. 122–124, Jan. 2017.

[27] L. Zheng, D. Liu, W. Yang, L. Yang, and Y. Zuo, "RaacLogo: A new sequence logo generator by using reduced amino acid clusters," *Briefings Bioinf.*, vol. 22, no. 3, May 2021, Art. no. bbaa096.

[28] L. Zheng, S. Huang, N. Mu, H. Zhang, J. Zhang, Y. Chang, L. Yang, and Y. Zuo, "RAACBook: A web server of reduced amino acid alphabet for sequence-dependent inference by using Chou's five-step rule," *Database*, vol. 2019, Jan. 2019, Art. no. baz131.

[29] S. Khan, M. Khan, N. Iqbal, T. Hussain, S. A. Khan, and K.-C. Chou, "A two-level computation model based on deep learning algorithm for identification of piRNA and their functions via Chou's 5-steps rule," *Int. J. Peptide Res. Therapeutics*, vol. 26, no. 2, pp. 795–809, Jun. 2020.

[30] A. Ahmad, S. Akbar, S. Khan, M. Hayat, F. Ali, A. Ahmed, and M. Tahir, "Deep-AntiFP: Prediction of antifungal peptides using distanct multi-informative features incorporating with deep neural networks," *Chemometric Intell. Lab. Syst.*, vol. 208, Jan. 2021, Art. no. 104214.

[31] S. Akbar, M. Hayat, M. Tahir, S. Khan, and F. K. Alarfaj, "CACP-DeepGram: Classification of anticancer peptides via deep neural network and skip-gram-based word embedding model," *Artif. Intell. Med.*, vol. 131, Sep. 2022, Art. no. 102349.

[32] S. Khan, M. Khan, N. Iqbal, S. A. Khan, and K.-C. Chou, "Prediction of piRNAs and their function based on discriminative intelligent model using hybrid features into Chou's PseKNC," *Chemometric Intell. Lab. Syst.*, vol. 203, Aug. 2020, Art. no. 104056.

[33] S. Akbar, S. Khan, F. Ali, M. Hayat, M. Qasim, and S. Gul, "iHBP-DeepPSSM: Identifying hormone binding proteins using PsePSSM based evolutionary features and deep learning approach," *Chemometric Intell. Lab. Syst.*, vol. 204, Sep. 2020, Art. no. 104103.

[34] S. Akbar, M. Hayat, M. Iqbal, and M. A. Jan, "IACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space," *Artif. Intell. Med.*, vol. 79, pp. 62–70, Jun. 2017.

[35] L. Hussain, S. A. Qureshi, A. Aldweesh, J. U. R. Pirzada, F. M. Butt, E. T. Eldin, M. Ali, A. Algarni, and M. A. Nadim, "Automated breast cancer detection by reconstruction independent component analysis (RICA) based hybrid features using machine learning paradigms," *Connection Sci.*, vol. 34, no. 1, pp. 2784–2806, Dec. 2022.

[36] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006.

[37] A. Zafari, R. Zurita-Milla, and E. Izquierdo-Verdiguier, "Land cover classification using extremely randomized trees: A kernel perspective," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 10, pp. 1702–1706, Oct. 2020.

[38] J. Sharma, C. Giri, O.-C. Granmo, and M. Goodwin, "Multi-layer intrusion detection system with ExtraTrees feature selection, extreme learning machine ensemble, and softmax aggregation," *EURASIP J. Inf. Secur.*, vol. 2019, no. 1, pp. 1–16, Dec. 2019.

[39] L. Breiman, "Random forests," Dept. Statist., Univ. California, Berkeley, CA, USA, Tech. Rep., 2001.

[40] A. Shahzad, B. Zafar, N. Ali, U. Jamil, A. J. Alghadhban, M. Assam, N. A. Ghamry, and E. T. Eldin, "COVID-19 vaccines related user's response categorization using machine learning techniques," *Computation*, vol. 10, no. 8, p. 141, Aug. 2022.

[41] F. Ali, S. Akbar, A. Ghulam, Z. A. Maher, A. Unar, and D. B. Talpur, "AFP-CMBPred: Computational identification of antifreeze proteins by extending consensus sequences into multi-blocks evolutionary information," *Comput. Biol. Med.*, vol. 139, Dec. 2021, Art. no. 105006.

[42] Y. Zhou, Z. Ahmad, Z. Almaspoor, F. Khan, E. Tag-Eldin, Z. Iqbal, and M. El-Morshedy, "On the implementation of a new version of the Weibull distribution and machine learning approach to model the COVID-19 data," *Math. Biosci. Eng.*, vol. 20, no. 1, pp. 337–364, 2022.

[43] S. Akbar, M. Hayat, M. Kabir, and M. Iqbal, "iAFP-gap-SMOTE: An efficient feature extraction scheme gapped dipeptide composition is coupled with an oversampling technique for identification of antifreeze proteins," *Lett. Organic Chem.*, vol. 16, no. 4, pp. 294–302, Mar. 2019.

[44] S. Akbar, M. Hayat, M. Tahir, and K. T. Chong, "cACP-2LFS: Classification of anticancer peptides using sequential discriminative model of KSAAP and two-level feature selection approach," *IEEE Access*, vol. 8, pp. 131939–131948, 2020.

[45] F. Ali, S. Ahmed, Z. N. K. Swati, and S. Akbar, "DP-BINDER: Machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information," *J. Comput.-Aided Mol. Des.*, vol. 33, no. 7, pp. 645–658, Jul. 2019.

[46] Q. Dai, Z. Wang, J. Song, X. Duan, M. Guo, and Z. Tian, "A stacked ensemble learning framework with heterogeneous feature combinations for predicting ncRNA-protein interaction," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2020, pp. 67–71.

[47] X. Chen, Z. Zhou, and Y. Zhao, "ELLPMDA: Ensemble learning and link prediction for miRNA-disease association prediction," *RNA Biol.*, vol. 15, no. 6, pp. 807–818, 2018.

[48] H. Su, Y. Yu, Q. Du, and P. Du, "Ensemble learning for hyperspectral image classification using tangent collaborative representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3778–3790, Jun. 2020.

[49] O. Barukab, F. Ali, and S. A. Khan, "DBP-GAPred: An intelligent method for prediction of DNA-binding proteins types by enhanced evolutionary profile features with ensemble learning," *J. Bioinf. Comput. Biol.*, vol. 19, no. 4, Aug. 2021, Art. no. 2150018.

[50] F. Althoey, M. N. Akhter, Z. S. Nagra, H. H. Awan, F. Alanazi, M. A. Khan, M. F. Javed, S. M. Eldin, and Y. O. Özkılıç, "Prediction models for Marshall mix parameters using bio-inspired genetic programming and deep machine learning approaches: A comparative study," *Case Stud. Construct. Mater.*, vol. 18, Jul. 2023, Art. no. e01774.

[51] M. Hayat, A. Khan, and M. Yeasin, "Prediction of membrane proteins using split amino acid and ensemble classification," *Amino Acids*, vol. 42, no. 6, pp. 2447–2460, Jun. 2012.

[52] L. Rokach, "Ensemble-based classifiers," *Artif. Intell. Rev.*, vol. 33, nos. 1–2, pp. 1–39, Feb. 2010.

[53] B. Liu, S. Wang, R. Long, and K.-C. Chou, "iRSpot-EL: Identify recombination spots with an ensemble learning approach," *Bioinformatics*, vol. 33, no. 1, pp. 35–41, Jan. 2017.

[54] S. Akbar, H. G. Mohamed, H. Ali, A. Saeed, A. Ahmed, S. Gul, A. Ahmad, F. Ali, Y. Y. Ghadi, and M. Assam, "Identifying neuropeptides via evolutionary and sequential based multi-perspective descriptors by incorporation with ensemble classification strategy," *IEEE Access*, vol. 11, pp. 49024–49034, 2023.

[55] W. Wattanapornprom, C. Thammarongtham, A. Hongsthong, and S. Lertampaiporn, "Ensemble of multiple classifiers for multilabel classification of plant protein subcellular localization," *Life*, vol. 11, no. 4, p. 293, Mar. 2021.

[56] S. Akbar, F. Ali, M. Hayat, A. Ahmad, S. Khan, and S. Gul, "Prediction of antiviral peptides using transform evolutionary & SHAP analysis based descriptors by incorporation with ensemble learning strategy," *Chemometric Intell. Lab. Syst.*, vol. 230, Nov. 2022, Art. no. 104682.

[57] A. Ahmad, S. Akbar, M. Tahir, M. Hayat, and F. Ali, "iAFPs-EnC-GA: Identifying antifungal peptides using sequential and evolutionary descriptors based multi-information fusion and ensemble learning approach," *Chemometric Intell. Lab. Syst.*, vol. 222, Mar. 2022, Art. no. 104516.

[58] M. Kabir and M. Hayat, "iRSpot-GAEnsC: Identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples," *Mol. Genet. Genomics*, vol. 291, no. 1, pp. 285–296, Feb. 2016.

[59] M. Hayat, M. Tahir, F. K. Alarfaj, R. Alturki, and F. Gazzawe, "NLP-BCH-ens: NLP-based intelligent computational model for discrimination of malaria parasite," *Comput. Biol. Med.*, vol. 149, Oct. 2022, Art. no. 105962.

[60] B. Chowdhury and G. Garai, "A review on multiple sequence alignment from the perspective of genetic algorithm," *Genomics*, vol. 109, nos. 5–6, pp. 419–431, Oct. 2017.

[61] J. Lin, H. Chen, S. Li, Y. Liu, X. Li, and B. Yu, "Accurate prediction of potential druggable proteins based on genetic algorithm and bagging-SVM ensemble classifier," *Artif. Intell. Med.*, vol. 98, pp. 35–47, Jul. 2019.

[62] A. Raza, W. Alam, S. Khan, M. Tahir, and K. T. Chong, "iPro-TCN: Prediction of DNA promoters recognition and their strength using temporal convolutional network," *IEEE Access*, vol. 11, pp. 66113–66121, 2023.

[63] A. Razaa, A. Ahmada, Z. Iqbala, Q. Yasina, H. Javeda, A. Shaha, and S. Chaudharya, "iAFP-ET: A robust approach for accurate identification of antifungal peptides using extra tree classifier and multi-view fusion," Tech. Rep.

[64] W.-R. Qiu, X. Xiao, W.-Z. Lin, and K.-C. Chou, "iMethyl-PseAAC: Identification of protein methylation sites via a pseudo amino acid composition approach," *BioMed Res. Int.*, vol. 2014, pp. 1–12, May 2014.

[65] S. Akbar and M. Hayat, "iMethyl-STTNC: Identification of N^6-methyladenosine sites by extending the idea of SAAC into Chou's PseAAC to formulate RNA sequences," *J. Theor. Biol.*, vol. 455, pp. 205–211, Oct. 2018.

[66] S. Ahmad, P. Charoenkwan, J. M. W. Quinn, M. A. Moni, M. M. Hasan, P. Lio, and W. Shoombuatong, "SCORPION is a stacking-based ensemble learning framework for accurate prediction of phage virion proteins," *Sci. Rep.*, vol. 12, no. 1, p. 4106, Mar. 2022.

[67] P. Charoenkwan, S. Ahmed, C. Nantasenamat, J. M. W. Quinn, M. A. Moni, P. Lio, and W. Shoombuatong, "AMYPred-FRL is a novel approach for accurate prediction of amyloid proteins by using feature representation learning," *Sci. Rep.*, vol. 12, no. 1, p. 7697, May 2022.

[68] S. Akbar, H. Ali, A. Ahmad, M. R. Sarker, A. Saeed, E. Salwana, S. Gul, A. Khan, and F. Ali, "Prediction of amyloid proteins using embedded evolutionary & ensemble feature selection based descriptors with extreme gradient boosting model," *IEEE Access*, vol. 11, pp. 39024–39036, 2023.

[69] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020.

[70] D. Garreau and U. Luxburg, "Explaining the explainer: A first theoretical analysis of LIME," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 1287–1296.

SHAHID AKBAR received the bachelor's degree in computer science and information technology from the Islamic University of Technology, Bangladesh, in 2011, and the M.S. and Ph.D. degrees in computer science from Abdul Wali Khan University Mardan (AWKUM), Pakistan, in 2015 and 2021, respectively. Currently, he is a Lecturer with the Department of Computer Science, AWKUM. He is also a Researcher with UESTC. He has published more than 25 papers in renowned journals in the field of computer engineering. He has more than ten years of research experience. His research interests include bioinformatics, digital image processing biomedical engineering, machine learning, and deep learning. Additionally, he serves as a reviewer for more than 20 well-known journals. He is also an Academic Editor of *PLOS One* journal.

ALI RAZA received the bachelor's degree in computer science from the University of Peshawar, Pakistan, in 2013, and the M.S. degree in computer science from City University Peshawar (CUSIT), Pakistan, in 2018. He is currently pursuing the Ph.D. degree with Qurtuba University, Peshawar, Pakistan. He is also a Lecturer with the Department of Computer Science, MY University, Islamabad. His research interests include bioinformatics, machine learning, and deep learning.

TAMARA AL SHLOUL is currently an Assistant Professor in humanities with the Liwa College of Technology. She has vast experience in teaching education and humanities courses, along with experience in school supervision, thinking skills, and higher education improvement ability. Her research interests include teacher socialization and professional development.

ASHFAQ AHMAD received the M.S. and Ph.D. degrees in computer science from Abdul Wali Khan University Mardan (AWKUM), Pakistan, in 2016 and 2023, respectively. Currently, he is an Assistant Professor with the Department of Computer Science, Muslim Youth University, Islamabad. His research interests include machine learning, deep learning, and bioinformatics.

AAMIR SAEED received the Ph.D. degree in wireless communication from Aalborg University, Denmark. He is currently an Assistant Professor with the Department of Computer Science and IT, University of Engineering and Technology, Peshawar. His research interests include big data structures (LSM and bloom filters), micro-services architecture, and the IoT with security in focus.

YAZEED YASIN GHADI received the Ph.D. degree in electrical and computer engineering from The University of Queensland. His dissertation on developing novel hybrid plasmonic photonic on chip biochemical sensors received the Sigma Xi Best Ph.D. thesis award. He was a Postdoctoral Researcher with The University of Queensland, before joining Al Ain University. His research interests include bioinformatics, digital image processing biomedical engineering, machine learning, and deep learning. He has published more than 80 peer-reviewed journals and conference papers and he holds three pending patents. His current research interests include developing novel electro-acousto-optic neural interfaces for large-scale high-resolution electrophysiology and distributed optogenetic stimulation. He was a recipient of several awards.

ORKEN MAMYRBAYEV received the B.S. and M.S. degrees in information systems from Abai University, Almaty, Kazakhstan, and the Ph.D. degree in information systems from Kazakh National Technical University named after K. I. Satbayev. He was an Associate Professor with the Institute of Information and Computational Technologies, Kazakhstan. He has been a Senior Researcher with the Laboratory of Computer Engineering of Intelligent Systems, Institute of Information and Computational Technologies. He is currently the Deputy General Director and the Head of the Laboratory of Computer Engineering of Intelligent Systems, Institute of Information, Kazakhstan. He is also a member of the Dissertation Council "Information Systems," L. N. Gumilyov Eurasian National University in the specialties computer sciences and information systems. He is the author of five books, more than 130 articles, and more than 20 inventions and copyright certificates for an intellectual property object in software. His main research interests include machine learning, deep learning, and speech technologies.

ELSAYED TAG-ELDIN was the Dean of the Faculty of Engineering, Cairo University, where he achieved many unique signs of progress in both academia and research on the impact of emerging technologies in electrical engineering. He is currently with Future University in Egypt, on leave from Cairo University, where he has nearly 30 years of service with the Faculty of Engineering. He was a PI of several nationally and internationally funded projects. He has many publications in highly refereed international journals and specialized conferences on the applications of artificial intelligence in the protection of electrical power networks. In addition, he is in the editorial boards of several *International Journal of Power and Energy Systems*.