Archives of Computational Methods in Engineering (2023) 30:4033–4044
https://doi.org/10.1007/s11831-023-09933-w

REVIEW ARTICLE

Recent Advances in Machine Learning-Based Models for Prediction of Antiviral Peptides

Farman Ali[1] · Harish Kumar[2] · Wajdi Alghamdi[3] · Faris A. Kateb[3] · Fawaz Khaled Alarfaj[4]

Received: 5 January 2023 / Accepted: 19 April 2023 / Published online: 29 April 2023

Abstract
Viruses have killed and infected millions of people across the world. It causes several chronic diseases like COVID-19, HIV, and hepatitis. To cope with such diseases and virus infections, antiviral peptides (AVPs) have been applied in the design of drugs. Keeping in view the significant role in pharmaceutical industry and other research fields, identification of AVPs is highly indispensable. In this connection, experimental and computational methods were proposed to identify AVPs. However, more accurate predictors for boosting AVPs identification are highly desirable. This work presents a thorough study and reports the available predictors of AVPs. We explained applied datasets, feature representation approaches, classification algorithms, and evaluation parameters of performance. In this study, the limitations of the existing studies and the best methods were emphasized. Provided the pros and cons of the applied classifiers. The future insights demonstrate efficient feature encoding approaches, best feature optimization schemes, and effective classification techniques that can improve the performance of novel method for accurate prediction of AVPs.

1 Introduction

A virus is a microscopic agent, that comprises nucleic acid within a protein and can multiply itself in the living cells of a host. In humans, several painful conditions, terrible infections, and diseases such as hepatitis, AIDS, cancer, pneumonia or dehydration, measles, mumps, common cold, smallpox, and rabies are caused by viruses. Almost every perspective of human life including economic, morbidity, and mortality is affected by viral infections. Several chronic outbreaks caused by zoonotic viruses like COVID-19, Zika, and Ebola killed and critically infected millions of people across the world. Due to efficient replication, different transmission routes, high genetic variation, prevention of harmful activities of viruses is challenging problem. Therapeutic approaches were introduced to tackle viral diseases. However, due to emergence of novel viruses, available antiviral methods are limited. Recently, 90 antiviral drugs were developed for the treatment of 9 virus families including hepatitis B and C viruses, HIV, herpes virus, human papilloma virus, respiratory syncytial virus, varicella-zoster virus, and cytomegalo virus.

Great progress was achieved in medicines through vaccine production for treatment of viral infections like polio and small pox. However, the new vaccines are facing several challenges in the terms of high cost and time. Recently, peptide-based drugs have been proposed which have lower cost, possess good tolerability, are relatively safe, and are highly selective. Among these peptide-based drugs, antiviral peptides (AVPs) have great significance in the development of novel drugs. AVPs are a sub-class of antimicrobial peptides that also work as antimicrobial in addition to antiviral activities.

To deal with viral infections and diseases, a series of antiviral therapeutic activities like replication of viruses, prevention of virus's fusion, blocking virus's attachments, and interruption of viruses signal processing were used. For example, protegrin-1 is a cyclical cationic peptide possesses

✉ Farman Ali
farman335@yahoo.com

✉ Fawaz Khaled Alarfaj
falarfaj@kfu.edu.sa

[1] Sarhad University of Science and Information Technology Peshawar, Mardan Campus, Khyber Pakhtunkhwa, Pakistan

[2] Department of Computer Science, College of Computer Science, King Khalid University, Abha, Saudi Arabia

[3] Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

[4] Department of Management Information Systems, King Faisal University, Hufof, Saudi Arabia

antiviral property against dengue virus. P9 antiviral peptide shows antiviral activity against flu strains. The accurate identification of AVPs can help to further explore the AVPs activities and perform a great role in the development of novel drugs. Initially, the prediction of AVPs was carried out by experimental approaches however, these methods are slow, expensive, and laborious. With the development of advanced technologies, the discovery of novel peptide sequences is increasing rapidly in the databases. The experimental methods for such rapid explosive sequences are inefficient. To cover these limitations, machine learning-based methods are indispensable for reliable and fast prediction of antiviral peptide activities (Fig. 1).

1.1 Existing Methods for Prediction of Antiviral Peptides

Currently, 09 machine learning-based approaches were designed for prediction of antiviral peptides. These methods were designed from 2012 to till date. Each predictor tried to improve the prediction of AVPs. For example, Thakur et al. [1] extracted features by AAC and physicochemical properties and trained the model using SVM. Chang et al. [2] applied aggregation, secondary structure, and physicochemical properties with RF. Zare et al. [3] only used PseAAC approach for feature encoding and Adaboost as classifier. In AVCpred [4] method, authors implemented several features like electrostatic, topological, hydrophobic, binary fingerprints, geometric, constitutional with SVM while AntiVPP [5] approach utilized hydropathy index, molecular weight, Net charge, and number of hydrogen bond donors in corporation with RF. FIRM-AVP [6] predictor encoded the feature using AAC, DPC, PseAAC, and secondary structure. The features are ranked by applying mean decrease of gini index (MDGI) technique while the model training and prediction were performed by SVM. PandoraGAN [7] carried out the model development using GAN in combination of physicochemical properties. Onward, AI4AVP [8] also utilized GAN with AAC, PseAAC, AA index, DPC, and physicochemical properties. Akbar et al. [9] explored numerical patterns from primary sequences using PSSM and K-segmentation PSSM. They have also considered SHAP as feature selection. The classification and prediction were executed by genetic algorithm ensemble learning strategy. List of existing methods with applied algorithms is provided in Table 1.

1.2 Drawbacks of the Past Studies

Each predictor tried to boost identification of AVPs by applying diverse features and classifiers. Still, each predictor

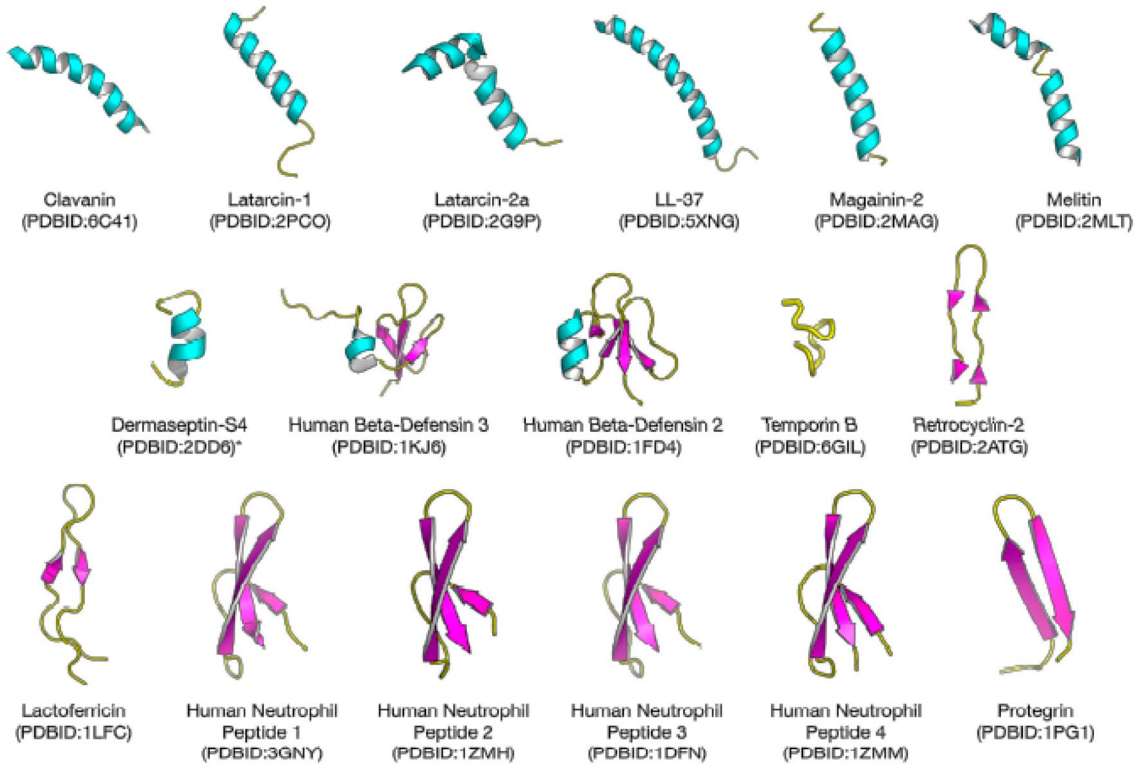

Fig. 1 Different kinds of antiviral peptides with their common name and number of Protein Databank identity number (PDBID) inside parenthesis

Table 1 List of available methods for prediction of AVPs

Predictor	Feature encoder	Feature selection method	Classifier
Thakur et al. [1]	Amino acid composition (AAC), Physicochemical properties	–	Support vector machine (SVM)
Chang et al. [2]	Amino acid composition (AAC),Physicochemical properties	–	Random forest (RF)
Zare et al. [3]	Pseudo amino acid composition (PseAAC)	–	Adaboost
AVCpred [4]	Geometric, constitutional, electrostatic, topological, hydrophobic, binary fingerprints		Support vector machine
AntiVPP [5]	Net charge, number of hydrogen bond donors, molecular weight, and hydropathy index	–	Random forest
FIRM-AVP [6]	Amino acid composition, pseudo amino acid composition, Dipeptide composition (DPC), secondary structure	Mean decrease of Gini index	Support vector machine
PandoraGAN [7]	Physicochemical properties	–	Generative Adversarial Network (GAN)
AI4AVP [8]	Amino acid composition, pseudo amino acid composition, AA index, and dipeptide composition, physicochemical properties	–	Generative Adversarial Network
Akbar et al. [9]	Position specific scoring matrix (PSSM), K-segmentation PSSM	Shapley Additive explanation (SHAP)	Genetic algorithm (GA) ensemble learner

has some deficiencies that degrade the model performance. For instance, Thakur et al. [1] method used AAC and physicochemical properties for feature extraction which are unable to extract the discriminative patterns. Chang et al. [2] used physicochemical features, AAC, secondary structure information, and aggregation. However, in the databanks, structural features have not existed for all proteins. Zare et al. [3] only utilized PseAAC for extraction of local information. However, an individual feature encoder is unable to explore informative features. AVCpred [4] extracted features by geometric, constitutional, electrostatic, topological, hydrophobic, binary fingerprints which are secondary features. AntiVPP [5] used number of hydrogen bond donors, molecular weight, hydropathy index, and Net charge. The best features can improve the performance of a model using feature optimization scheme. FIRM-AVP [6] adopted AAC, PseAAC, DPC, and secondary features. Feature vector of these encoding approaches generate high dimensional space that can affect a model performance. The authors also implemented mean decrease of gini index (MDGI) feature selection approach.

Onward, PandoraGAN [7] applied physicochemical properties and AI4AVP [8] used AAC, PseAAC, AA index, DPC, and physicochemical properties. Both methods performed the model training using deep learning framework named Generative Adversarial Network (GAN). However, deep learning models can generate accurate results on large datasets. In another work, Akbar et al. [9] used evolutionary profile however, one feature descriptor can't explore the important patterns.

This approach also implemented SHAP for selection of best features.

The online web server can enrich significance of a model and a user can determine class-label for a sequence. However, most predictors have not implemented web servers. More importantly, a brief review can highlight the merits of the existing predictors. This study illustrates the employed datasets, feature representation methods, and classification algorithms. Onward, we discussed performance of all existing predictors with their merits and demerits, and demonstrated the best methods for accurate identification of AVPs. This study can provide fruitful direction to scientists and researchers to select the best AVPs predictor. Furthermore, we explore more efficient feature descriptors, and classifiers to design predictors with high precision. The phases of model development are shown in Fig. 2.

Fig. 2 Steps of a model development

2 Materials and Methods

2.1 Construction of Datasets

Application of a reliable dataset improve the quality of a predictor. In AVPpred method, Thakur et al. [1] constructed two training datasets and two validation datasets. These datasets were collected from 48 research articles in which 91% were retrieved from natural source and rest are synthetic source. After elimination of identical peptides, final training set-1 contains 544 AVPs and 407 non-AVPs peptides. Validation set-1 comprises 60 AVPs and 45 non-AVPs sequences of peptides. Second training set-2 consists of 544 AVPs and 544 non-AVPs. Similarly, validation set-2 contains 60 AVPs and 60 non-AVPs. Chang et al. [2] also used the same datasets in their work.

In another work, Zare et al. method applied a novel dataset which is downloaded from antiviral peptides database [3]. Initially, dataset was 614 AVPs and 452 non-AVPs. The similar sequences were removed using CD-HIT tool with cut-off value 90%. The final dataset-3 comprises 342 AVPs and 312 non-AVPs.

During the development of AVCpred [4] method, authors used experimentally validated dataset that contains 416 hepatitis B virus (HBV), 1383 human immunodeficiency virus (HIV), 803 hepatitis C virus (HCV), 473 human herpesvirus (HHV), and 26 general AVPs.

AntiVPP [5], FIRM-AVP [6], PandoraGAN [7], AI4AVP [8], and Akbar et al. [9] methods implemented the datasets constructed by Thakur et al. [1]. In this study, we referred training set-1, validation set-1, training set-2, validation set-2, and dataset-3 as training dataset-1, validation dataset-1, training dataset-2, validation dataset-2, and training dataset-3, respectively.

2.2 Features Representation Methods

Machine learning algorithms are unable to directly deal with primary sequences datasets of peptides during the model development. In this connection, features are extracted by descriptors into numerical format. The existing predictors for identification of AVPs applied different feature encoding methods which are explained in the following subsections.

2.2.1 Amino Acid Composition

A peptide is a combination of two or more residues among 20 amino acids that possessed properties of that sequence. AAC calculates the frequency of each residue and extracts the global patterns. AAC is calculated using the equation below.

$$AAC = (R_1, R_2, \ldots .R_{20}) \tag{1}$$

where $R_i (i = 1, 2, ..., 20)$ shows 20 residues frequencies in a protein sequence.

AAC is broadly applied for solving diverse biological issues such as prediction of membrane protein types [10], identification of antiviral peptides [9], identification of antifreeze proteins [11], web server for protein/peptide feature extraction [12]. AAC features are used by Thakur et al. [1], Chang et al. [2], FIRM-AVP [6], and AI4AVP [8] in their models for AVPs prediction.

2.2.2 Physicochemical Properties

Each peptide contains two or more amino acids that are connected by peptide bonds [13]. Due to these peptide bonds, AVPs possess different physicochemical properties that are highly associated with AVPs activities. These physicochemical properties include residue composition, size, overall charge, secondary structure, amphiphilic character, and hydrophobicity. Keeping in view the crucial role of these properties, these are used in the construction of predictive models such as prediction of DNA-binding proteins [14], allergenic proteins [15], and essential proteins [16].

In the literature, Thakur et al. [1], Chang et al. [2], FIRM-AVP [6], PandoraGAN [7], and AI4AVP [8] approaches designed their models by incorporating these properties.

2.2.3 Pseudo Amino Acid Composition

AAC is a conventional approach that explores the compositional information regarding amino acids. AAC only computes the frequencies of residues but avoids the sequential order and correlation among amino acids [17]. To cover these limitations, PseAAC was introduced by Chou's [18]. The PseAAC can be formulated using the following equation.

$$PseAAC = (R_1, R_2, \ldots .R_{20,} R_{20+1,\ldots.} R_{20+\lambda,}) \tag{2}$$

where $R_1, R_2, \ldots .R_{20}$ are 20 amino acids and $R_{20+1,\ldots.} R_{20+\lambda})$ are the correlation factors of amino acids [19].

Due to the effectiveness of PseAAC, it was implemented by many researchers like sub-cellular location of proteins [20], conotoxin super-family and family classification [21], protease types [22], human papillama viruses [23], submitochondria localization [24], apoptosis protein sub-cellular localization [25], protein quaternary structure [26, 27], and bacterial secreted proteins [28] identification.

To predict AVPs, Zare et al. [3], FIRM-AVP [6], and AI4AVP [8] approach adopted PseAAC as feature descriptor.

2.2.4 Dipeptide Composition

DPC is an efficient algorithm that computes the frequencies of two consecutive amino acids along a sequence [29]. In addition to global patterns, DPC can extract the correlation factors of residues [30]. Therefore, it can provide discriminative motifs regarding a sequence. This approach contributed well to the design of several predictors such as prediction of extracellular matrix proteins [31], antifungal proteins [32], and DNA-binding proteins prediction [14]. This method is also adopted for prediction of Antifreeze proteins [33]. DPC generates a vector of 400 dimension that can be calculated by following equation.

$$DPC = \frac{N(a)}{L} \tag{3}$$

where N is the fraction of dipeptide a and L is the number of dipeptides.

FIRM-AVP [6] and AI4AVP [8] methods used DPC for identification of AVPs.

2.2.5 Position Specific Scoring Matrix

Like physicochemical properties, evolutionary features are the significant properties of AVPs. Evolutionary features are derived using position specific scoring matrix [34]. PSSM is computed by PSI-BLAST tool with three iterations and 0.01 as cut-off value [35]. A PSSM is represented by L × 20 matrix. L is the number of rows and 20 indicates columns.

PSSM boosted the predictive results of many predictors like identification of cancerlectins peptides [36], growth hormone-binding proteins [37], membrane protein types [38], druggable proteins [39], antioxidant proteins [40], antitubercular peptides [41], antifreeze proteins [11], prediction of hormone-binding proteins [42] and antifungal peptides [32].

It is reported by past work that local regions of evolutionary profile contain crucial patterns. To capture these local patterns, each PSSM was split into k-segmentations. This approach was used by Akbar et al. [9]. In this method, each PSSM is decomposed into three segments, computed the local regions, and finally combined to make one super set. KS-PSSM improved the performance of the predictor and predicted antiviral peptides more accurately.

2.3 Feature Selection Approaches

Sometimes feature set contains less informative and noisy features that degrade a model performance. To cover this limitation, feature selection approaches are applied to feature set for selection of the best feature set. Using these approaches, only informative patterns are selected that boost a predictor performance. In this connection, only two methods i.e., Shapley Additive exPlanation (SHAP) [9] and Mean Decrease of Gini Index (MDGI) [6] are used for prediction of antiviral peptides. The working strategy of these algorithms is listed below.

2.3.1 Shapley Additive exPlanation

To select the best feature, Akbar et al. [9] used SHAP algorithm. SHAP examines the participation of each feature using aggregations of best shapley instances. It interprets the performance of classifier and addresses the model deficiency. Akbar et al. used eXtreme Gradient Boosting classifier to perform classification and prediction tasks[43].

SHAP approach keeps the important features which leads to promising performance and avoids the less information features [44]. After evaluating each feature, authors selected 35 high-ranked features for model development. This algorithm was also implemented in other research problems including identification of hypoxaemia during surgery [45] and dimensionality reduction [46].

2.3.2 Mean Decrease of Gini Index

MDGI was implemented by FIRM-AVP [6] for enhancing the prediction performance of the model. The feature set was obtained 649. To reduce the size and improve model performance, first authors determined the correlation between features using Pearson's correlation scheme. Features with greater correlation value than threshold value was eliminated.

Further important features were selected by MDGI using the RF model. With MDGI algorithm, each feature is measured with respect to homogeneity to leaves and nodes of RF model. The feature is considered more important if MDGI is closer to 0. After completion of feature selection process, 169 were considered the best set. This best set was provided to the classifier and achieved progressive performance.

2.4 Model Training and Prediction

After feature extraction and best feature selection, the next phase is the application of appropriate classifier. A classifier is used to train the model and predict the unlabel sample. Considering this, different existing methods of AVPs utilized different classifiers which are explained in the following sections.

2.4.1 Support Vector Machine

SVM is considered a promising classifier for both regression and classification problems [47–49]. It was first proposed by Vapnik [50]. SVM transforms the dataset samples

into high-dimensional feature space into classes. A separator called hyperplane is drawn between classes and margin lines parallel to hyperplane [51]. These marginal lines are called support vectors. SVM uses four kernel functions for transformation of data namely sigmoid, linear, polynomial, and radial basis function (RBF) [52, 53]. Linear function is used for linear separable problems while other functions are applied to solve non-linear issues [54].

SVM implements grid search approach to find the best values for C and γ parameters to improve the model prediction. Due to promising performance of SVM, it was used for solving many research challenging tasks like protein remote homology detection [55], protein structure prediction [56], identification of antifreeze protein [57], identification of DNA-binding proteins [14], protein fold recognition, and prediction of promoter [58]. SVM was applied by Thakur et al. [1], AVCpred [4], and FIRM-AVP [6] for prediction of antiviral peptides.

2.4.2 Random Forest

RF is an ensemble classifier that was established by Breiman [59, 60]. It can be utilized for clustering, feature selection, regression, and classification tasks [61, 62]. RF consists of many trees like a forest. Each tree is trained by training sample of dataset. After training phase, a class label is assigned to new sample using majority voting scheme.

Keeping in view the majority voting benefits, the high variance or bias of a single tree can't affect the overall performance of a model [63]. Onward, RF uses the weighting scheme that assigns a low weight if a tree has high error rate and boosts the tree performance[43]. RF is mostly favorable for large datasets, handling efficiently missing data, and detecting of outlier issues [64]. For identification of AVPs, Chang et al. [2] and AntiVPP [5] predictors used RF in their models.

2.4.3 Generative Adversarial Network

GAN is one of the most popular networks of deep learning [65]. GAN is a generative framework where a model is trained by adversarial process. GAN comprises two models which are trained at the same time i.e., discriminative model and generative model [66]. Discriminative model estimates the probability of a sample in the training set and generative model captures the distribution of data. During training phase, generative model looks for maximization of probability of discriminative model [67].

Generative model produces images from random noise and seeks to produce realistic images [43]. To complete this process, random noise is provided to generator that yields fake images. These images are provided to discriminative model that differentiates between real and fake images [68].

GAN has many real applications in different fields of life. For instance, improving cybersecurity [69], predictors in healthcare [70], stock market prediction [71], producing animation models[72], editing photographs, and image translation are popular applications [73, 74].

2.4.4 Adaboost

Adaboost stands for Adaptive Boosting which was constructed by Schapire and Freund [75]. Adaboost concatenates many weak classifiers to construct one strong classifier. Because a single classifier cannot generate accurate results. Therefore, grouping multiple weak classifiers learn from each other wrong classification of samples which made a strong one.

Weak classifiers are decision trees having a single split, called decision stumps. Decision stumps are not fully grown that have one node and two leaves. Adaboost assigns more weights to high error rate classifier while putting low weights to a less error rate classifier. Zare et al. [3] used adaboost in the design of the proposed model for identifying antiviral peptides.

Adaboost has several advantages like less prone to overfitting, boosting the accuracies of weak classifiers, and being usable for image and text classification instead of binary classification problems [76].

2.4.5 Ensemble Learning

Ensemble learning is the combination of several algorithms that construct one optimal classification algorithm. This strategy was used by researchers due its effective generalization power and high prediction rate. Ensemble learning is considered more fruitful as it decreases bias and variance of a model [77]. Considering these merits, many scientists applied it in the development of models including nucleosome positioning [78], anticancer peptides [79], antifreeze proteins [80], enhancers types [81].

Akbar et al. [9] used K-Nearest Neighbor (KNN), SVM, Extremely Randomized Tree (ERT), and eXtreme Gradient Boosting (XGB) classifiers are provided to genetic algorithm ensemble learning approach. The Ensemble learning significantly improved the model performance and achieved the highest success rate for identification of antiviral peptides.

The pros and cons of each classifier are reported in Table 2.

2.5 Model Validation Methods

After construction of a novel model, its effectiveness is validated. For this purpose, tenfold test is widely used by existing methods [14, 37, 82]. In tenfold test, dataset is decomposed into 10-folds. ninefold are used for model training

Table 2 Pros and cons of the applied learning models

Classifier	Pros	Cons
RF	RF has the ability to learn complex decision boundary. It reduces variance and boosts the model performance. Required minimum data preprocessing. Model training is speed up by creation of parallel trees	RF has high computational cost. It required more resources. RF is not suitable for regression problems. Changes are difficult in RF model
SVM	SVM generates good predictive results by transforming data into high dimensional space. It draws hyperplane with maximum margin using kernel functions. Over-fitting issue is dealt with by L2 Regularization approach	SVM is time-consuming and becomes complex in case of large dataset. It reserves maximum memory. SVM scalability is hard
GAN	GAN produces looks like original data. It can produce similar versions of audio, text, and video. It develops better modeling of data distribution	Generation of prediction results using text or speech is complex. GAN model training is hard and unstable. During learning process of model, it can miss patterns
Adaboost	Adaboost implementation is easy. It progressively improves model performance by concatenating weak learners. Adaboost can easily cope with overfitting issues	Adaboost is sensitive to noise data. It is not prone to outliers. Adaboost is slower than other boosting classifiers like XGBoost
Ensemble learning	Ensemble learning generates lower bias and variance. It creates deeper understanding of the data. Ensemble strategy is used for accuracy improvement	Ensemble is difficult in terms of interpretation. It is costly to create and train. Ensemble learning models stuck in local optima

Table 3 Comparison of previous models using training dataset-1

Predictor	Acc (%)	Sn (%)	Sp (%)	MCC
AVPpred	85.00	82.20	88.20	0.70
Chang et al	85.10	86.60	83.00	0.70
AntiVPP	-	-	-	-
FIRM-AVP	92.40	93.30	91.10	0.84
Meta-iAVP	88.20	89.20	86.90	0.76
Akbar et al	97.33	92.36	98.85	0.89

and onefold is assigned for model validation. This process continues up to ten times so that each fold is used as testing. The mean of all folds is considered the final value [83]. Further, the model is validated by four evaluation parameters like accuracy (Acc), sensitivity (Sn), specificity (Sp), and Mathew correlation coefficient (MCC) [60]. These parameters can be formulated as

$$\begin{cases} Sn = 1 - \frac{AV_{-}^{+}}{AV^{+}} \\ Sp = 1 - \frac{AV_{+}^{-}}{AV^{-}} \\ Acc = 1 - \frac{AV_{-}^{+} + AV_{+}^{-}}{AV^{+} + AV^{-}} \\ MCC = \frac{1 - \left(\frac{AV_{-}^{+} + AV_{+}^{-}}{AV^{+} + AV^{-}} \right)}{\sqrt{\left(1 + \frac{AV_{+}^{-} - AV^{+}}{AV^{+}} \right) \left(1 + \frac{AV_{-}^{+} - AV_{+}^{-}}{AV^{-}} \right)}} \end{cases} \tag{4}$$

AV^{+} indicates positive (antiviral) samples, AV^{-} represents negative (non-antiviral) samples. AV_{+}^{-} are the negative samples predicted mistakenly as positive and AV_{-}^{+} are positive samples that are incorrectly predicted as negative.

3 Results and Discussion

In this section, we analyzed the performance of existing methods and elaborate the best method for prediction of AVPs. We performed a comparison of all methods and point out the best predictor in the literature. The performance of all existing studies dataset-wise is discussed in the following sections.

3.1 Comparison of Existing Studies on Training Dataset-1 (544 + 407)

AVPpred, Change et al., AntiVPP, FIRM-AVP, Meta-iAVP, and Akbar et al. used training dataset-1. The results of Acc, Sn, Sp, and MCC are listed in Table 3. These methods deployed different feature descriptors and classifiers in their models. AVPpred achieved an accuracy of 85.00%, sensitivity of 82.20%, specificity of 88.20%, and MCC of 0.70. The same accuracy and MCC were secured by Chang et al.

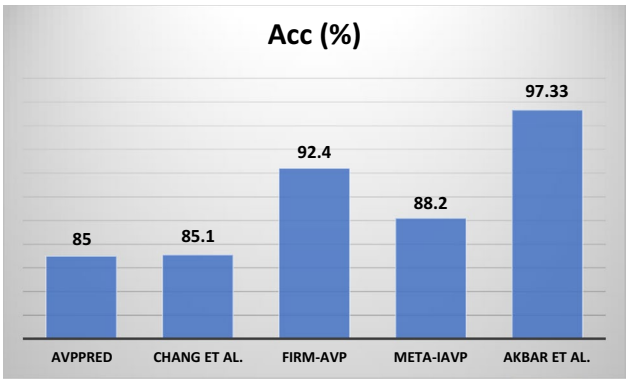

Fig. 3 Accuracy comparison of predictors on training dataset-1

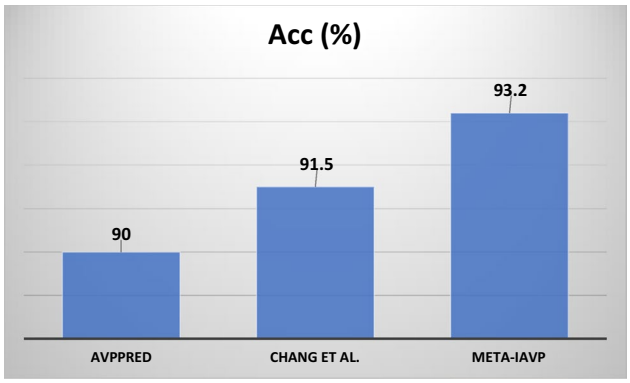

Fig. 4 Accuracy comparison on predictors on training dataset-2

Table 4 Comparison of previous models using training dataset-2

Predictor	Acc (%)	Sn (%)	Sp (%)	MCC
AVPpred	90.00	89.70	90.30	0.80
Chang et al	91.50	89.00	94.10	0.83
AntiVPP	–	–	–	–
FIRM-AVP	–	–	-	–
Meta-iAVP	93.20	89.00	97.40	0.87
Akbar et al	–	–	–	–

Table 5 Comparison of previous models using validation dataset-1

Predictor	Acc (%)	Sn (%)	Sp (%)	MCC
AVPpred	85.70	88.30	82.20	0.71
Chang et al	89.50	91.70	86.70	0.79
AntiVPP	–	–	–	–
FIRM-AVP	92.40	93.30	91.10	0.84
Meta-iAVP	95.20	96.70	93.20	0.90
Akbar et al	–	–	–	–

However, Chang et al. improved the sensitivity to 86.60% and decreased it to 83.00%. AntiVPP predictor was not used the training dataset-1, while FIRM-AVP boosted the performance by yielding 92.40% Acc, 93.30% Sn, 91.10% Sp, and 0.84 MCC. Meta-iAVP method outperformed AVPpred and Chang et al. However, lower performance is obtained by Meta-iAVP than FIRM-AVP. The best performance was achieved by Akbar et al. model with Acc of 97.33%, Sn of 92.36%, Sp of 98.85%, and MCC of 0.89. The performance of existing predictors in terms of accuracy has been provided in Fig. 3.

3.2 Comparison of Existing Studies on Training Dataset-2 (544 + 544)

Training dataset-2 was used only by three methods namely AVPpred, Chang et al., and Meta-iAVP. The Acc, Sn, Sp, and MCC of AVPpred are 90.00%, 89.70%, 90.30%, and 0.80, respectively as listed in Table 4. Chang et al. predictor improved accuracy, specificity, and MCC, however, generated slightly lower sensitivity than AVPpred. On training dataset-2, Meta-iAVP achieved remarkable performance and produced 93.20% Acc, 89.00% Sn, 97.40% Sp, and 0.87 MCC. These results are also higher than both AVPpred and Chang et al. methods. The performance of FIRM-AVP and Akbar et al. is promising on training dataset-1, however, they do not implement training dataset-2 to evaluate their models. On training dataset-2, the best performance was shown by Meta-iAVP approach. Accuracy comparison is depicted in Fig. 4.

3.3 Comparison of Existing Studies on Validation Dataset-1 (60 + 45)

In addition to training datasets, model performance can be evaluated by independent/testing datasets. In this connection, several existing methods determined the generalization power of their models by validation dataset-1 and validation dataset-2. The Acc, Sn, Sp, and MCC produced by AVPpred predictor on validation dataset-1 are 85.70%, 88.30%, 82.20%, and 0.71, respectively as shown in Table 5. Chang et al. boosted the prediction values of Acc, Sn, Sp, and MCC which are 3.8%, 3.4%, 4.5%, and 0.8, respectively higher than AVPpred. Compare with AVPpred and Chang et al. methods, FIRM-AVP model also secured better performance by attaining 92.40% Acc, 93.30% Sn, 91.10% Sp, and 0.84 MCC. Among all methods, the highest results are generated by Meta-iAVP. It means that this method can discriminate AVPs from non-AVPs more efficiently. AntiVPP and Akbar et al. models have not assessed their proposed predictors by validation dataset-1. Graphical representation of accuracies of existing methods has shown in Fig. 5.

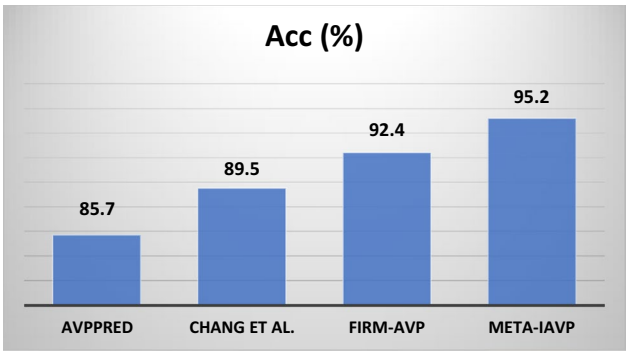

Fig. 5 Graphical view of accuracies using validation dataset-1

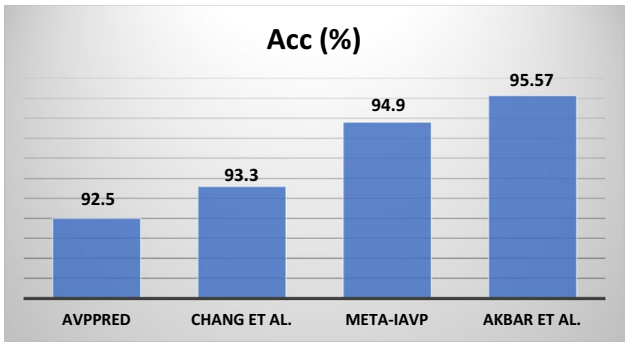

Fig. 6 Accuracy comparison of past studies using validation dataset-2

3.4 Comparison of Existing Studies on Validation Dataset-2 (60 + 60)

Several methods including AVPpred, Chang et al., Meta-iAVP, and Akbar et al. approaches examined their proposed studies using validation dataset-2. The comparative prediction results are summarized in Table 6. AVPpred yielded an accuracy of 92.50%, sensitivity of 93.30%, specificity of 91.70%, and MCC of 0.85. Chang et al. improved the performance on all evaluation parameters i.e., 93.30% Acc, 91.70% Sn, 95.00% Sp, and 0.87 MCC. AntiVPP and FIRM-AVP have not examined their methods by validation dataset-2. With the same dataset, values of accuracy, sensitivity, specificity, and MCC boosted by Meta-iAVP are 94.90%, 91.70%, 98.20%, and 0.90, respectively. Akbar et al. further enhanced the performance using all evaluation parameters. Among all methods, Akbar et al. model is superior to all existing predictor on validation dataset-2. A comparative view of the past studies is drawn in Fig. 6.

3.5 Comparison of Existing Studies on Training Dataset-3 (342 + 312)

The training dataset-3 was only implemented by Zare et al. In this method, authors extracted features by PseAAC and model training and classification were performed by Adaboost (RBF), Adaboost (Naive Bayes), Adaboost (J48), Adaboost (Decision Stump), and Adaboost (REFTree). Among all models, Adaboost (J48) achieved the highest results in terms of 93.26% Acc, 0.926 Sn, 0.939 Sp, and 0.86 MCC. The second-best (87.59%) accuracy was attained by Adaboost (REFTree) and Adaboost (Naive Bayes) showed the third-best (78.87%) accuracy. Although, the performance of this predictor is promising, however, it is not compared with other existing studies due to application of different datasets. Moreover, this method was not validated by a testing dataset to determine its generalization capacity for unseen samples. Therefore, the use of one training dataset can't explore its reliability and can't be considered an effective model.

Table 6 Comparison of previous models using validation dataset-2

Predictor	Acc (%)	Sn (%)	Sp (%)	MCC
AVPpred	92.50	93.30	91.70	0.85
Chang et al	93.30	91.70	95.00	0.87
AntiVPP	–	–	–	–
FIRM-AVP	–	–	–	–
Meta-iAVP	94.90	91.70	98.20	0.90
Akbar et al	95.57	93.64	98.72	0.89

3.6 Advantages of the Past Studies

Keeping in view the crucial utilization and implementation of antiviral peptides in diverse fields of life, identification with high precision is indispensable. However, identification of antiviral peptides via experimental methods is a challenging task due to annotative peptides being multiplied rapidly in the databanks. To overcome the limitations, researchers designed computational methods using machine learning techniques. These computational predictors are more efficient than experimental approaches in terms of fast and accurate prediction. The performance of these methods can be boosted by applications of efficient feature encoding approaches, appropriate feature optimization schemes, and deep learning approaches.

3.7 Limitations of the Past Studies

Although machine learning methods surpassed the experimental methods. However, limitations of the existing methods can degrade a method's performance. For instance, most predictors encoded features by AAC, physicochemical features, PseAAC, DPC, and AA index which avoids important patterns of antiviral peptides. Selection of effective features can boost a model's performance. However, only two existing predictors FIRM-AVP [6] and Akbar et al. [9] considered these schemes. Further, these methods did not use deep

learning frameworks that lead to lower performance. An online web server is useful prediction of AVPs and enriches significance of a model. However, existing approaches were not designed web servers.

4 Conclusion

Antiviral peptides are significant for development of vaccine as they have lower costs, possess good tolerability, are relatively safe, and are highly selective. Many predictors were introduced to enhance prediction of AVPs. Among these predictors, Akbar et al. method secured the highest results on training dataset-1 and validation dataset-2. However, this method has not been implemented training dataset-2 and validation dataset-1. Meta-iAFP predictor also achieved the second-best results on all four datasets.

In the literature, other approaches used some of the datasets and did not train and validated their models by all datasets. Therefore, it is concluded that Meta-iAFP could be reliable predictor for discrimination of AVPs from non-AVPs.

5 Future direction

Accurate identification of AVPs is a challenging task in bioinformatics and drug designing fields. The predictors can identify AVPs. However, it is still highly desirable to predict AVPs with high precision. The prediction results can be boosted by implementations of effective descriptors like biological sub-words, fastText, and bidirectional encoder representations from transformers (BERT). The heterogeneous patterns can be extracted by incorporating segmentation strategies, bigram, PseAAC, and DPC into PSSM. Compression techniques including discrete wavelet transform and discrete cosine transform into PSSM can explore discriminative features.

Advance technologies like multi-headed convolutional neural network (MHCNN), ensembles of convolutional neural networks (ECNN), transfer learning, gated recurrent units (GRU), bidirectional long short-term memory (BiLSTM), and recurrent neural network (RNN) can be implemented for model training and prediction. The efficacy of a model can be enhanced using efficient feature selection algorithms. Integration of different features can boost the performance of the model.

Acknowledgements The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University for funding this work under grant number RGP. 2/229/44.

Funding This work was supported by the Deanship of Scientific Research, King Faisal University, Saudi Arabia.

Declarations

Conflict of interest Authors intend no competing interest.

References

1. Thakur N, Qureshi A, Kumar M (2012) AVPpred: collection and prediction of highly effective antiviral peptides. Nucleic Acids Res 40(W1):W199–W204
2. Chang KY, Yang J-R (2013) Analysis and prediction of highly effective antiviral peptides based on random forests. PLoS ONE 8(8):e70166
3. Zare M et al (2015) Using Chou's pseudo amino acid composition and machine learning method to predict the antiviral peptides. Open Bioinform J 9(1):13–19
4. Qureshi A, Kaur G, Kumar M (2017) AVC pred: an integrated web server for prediction and design of antiviral compounds. Chem Biol Drug Des 89(1):74–83
5. Lissabet JFB, Belén LH, Farias JG (2019) AntiVPP 1.0: a portable tool for prediction of antiviral peptides. Comput Biol Med 107:127–130
6. Chowdhury AS et al (2020) Better understanding and prediction of antiviral peptides through primary and secondary structure feature importance. Sci Rep 10(1):1–8
7. Surana S et al (2022) Pandoragan: generating antiviral peptides using generative adversarial network. bioRxiv, p. 2021–02
8. Lin T-T et al (2022) AI4AVP: an antiviral peptides predictor in deep learning approach with generative adversarial network data augmentation. Bioinform Adv 2(1):vbac080
9. Akbar S et al (2022) Prediction of antiviral peptides using transform evolutionary & SHAP analysis based descriptors by incorporation with ensemble learning strategy. Chemom Intell Lab Syst 230:104682
10. Ali F, Hayat M (2015) Classification of membrane protein types using voting feature interval in combination with Chou′ s pseudo amino acid composition. J Theor Biol 384:78–83
11. Ali F et al (2021) AFP-CMBPred: computational identification of antifreeze proteins by extending consensus sequences into multi-blocks evolutionary information. Comput Biol Med 139:105006
12. Chen Z et al (2018) iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics 34(14):2499–2502
13. Swati ZNK et al (2019) Content-based brain tumor retrieval for MR images using transfer learning. IEEE Access 7:17809–17822
14. Ali F et al (2018) DBPPred-PDSD: machine learning approach for prediction of DNA-binding proteins using discrete wavelet transform and optimized integrated features space. Chemom Intell Lab Syst 182:21–30
15. Mohabatkar H et al (2013) Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach. Med Chem 9:133–137
16. Sarangi AN, Lohani M, Aggarwal R (2013) Prediction of essential proteins in prokaryotes by incorporating various physico-chemical Features into the general form of Chou's pseudo amino acid composition. Protein Pept Lett 20(7):781–795
17. Ahmed S et al (2018) Improving secretory proteins prediction in *Mycobacterium tuberculosis* using the unbiased dipeptide composition with support vector machine. Int J Data Mining Bioinform 21(3):212–229
18. Chou K-C (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21(1):10–19
19. Chou KC (2001) Prediction of protein subcellular attributes using pseudo-amino acid composition. Proteins 43:246–255

20. Arif M et al (2020) TargetCPP: accurate prediction of cell-penetrating peptides from optimized multi-scale features using gradient boost decision tree. J Comput Mol Des 34(8):841–856
21. Mondal S et al (2006) Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J Theor Biol 243(2):252–260
22. Zhou GP, Cai YD (2006) Predicting protease types by hybridizing gene ontology and pseudo amino acid composition. Proteins 63(3):681–4
23. Cao DS, Xu QS, Liang YZ (2013) propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics 29:960–962
24. Nanni L, Lumini A (2008) Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. Amino Acids 34(4):653–660
25. Chen, et al (2007) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo-amino acid composition. J Theor Biol 248(2):377–81
26. Sun XY et al (2012) Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform. Mol BioSyst 8:3178–3184
27. Zhang GY, Fang BS (2008) Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. Amino Acids 34(4):565–572
28. Nanni L et al (2012) Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information. IEEE/ACM Trans Comput Biol Bioinform 9:467–475
29. Khan ZU et al (2019) iPredCNC: computational prediction model for cancerlectins and non-cancerlectins using novel cascade features subset selection. Chemom Intell Lab Syst 195:103876
30. Arif M et al (2020) Pred-BVP-Unb: fast prediction of bacteriophage virion proteins using un-biased multi-perspective properties with recursive feature elimination. Genomics 112(2):1565–1574
31. Ali F, Hayat M (2016) Machine learning approaches for discrimination of extracellular matrix proteins using hybrid feature space. J Theor Biol 403:30–37
32. Ahmad A et al (2021) Deep-AntiFP: prediction of antifungal peptides using distanct multi-informative features incorporating with deep neural networks. Chemom Intell Lab Syst 208:104214
33. Fletcher GL, Hew CL, Davies PL (2001) Antifreeze proteins of teleost fishes. Annu Rev Physiol 63(1):359–390
34. Ahmad A et al (2022) iAFPs-EnC-GA: Identifying antifungal peptides using sequential and evolutionary descriptors based multi-information fusion and ensemble learning approach. Chemom Intell Lab Syst 222:104516
35. Banjar A et al (2022) iDBP-PBMD: a machine learning model for detection of DNA-binding proteins by extending compression techniques into evolutionary profile. Chemom Intell Lab Syst 231:104697
36. Ali F et al (2022) Deep-PCL: a deep learning model for prediction of cancerlectins and non cancerlectins using optimized integrated features. Chemom Intell Lab Syst 221:104484
37. Ali F et al (2022) Deep-GHBP: improving prediction of growth hormone-binding proteins using deep learning model. Biomed Signal Process Control 78:103856
38. Kabir M et al (2018) Improving prediction of extracellular matrix proteins using evolutionary information via a grey system model and asymmetric under-sampling technique. Chemom Intell Lab Syst 174:22–32
39. Sikander R, Ghulam A, Ali F (2022) XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set. Sci Rep 12(1):1–9
40. Ahmad A et al (2020) Identification of antioxidant proteins using a discriminative intelligent model of k-space amino acid pairs based descriptors incorporating with ensemble feature selection. Biocybern Biomed Eng 42:727–735
41. Akbar S et al (2021) iAtbP-Hyb-EnC: prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model. Comput Biol Med 137:104778
42. Akbar S et al (2020) iHBP-DeepPSSM: identifying hormone binding proteins using PsePSSM based evolutionary features and deep learning approach. Chemom Intell Lab Syst 204:104103
43. Khan A et al (2023) AFP-SPTS: an accurate prediction of antifreeze proteins using sequential and pseudo-tri-slicing evolutionary features with an extremely randomized tree. J Chem Inf Model 26:826–834
44. Ghulam A et al (2022) AI and Machine Learning-based practices in various domains: A Survey. V Fast 10:21–41
45. Lundberg SM et al (2018) Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nature Biomed Eng 2(10):749–760
46. Kumar CS et al (2020) Dimensionality reduction based on shap analysis: a simple and trustworthy approach. In 2020 international conference on communication and signal processing (ICCSP). IEEE 558–560
47. Rahu S et al (2022) UBI-XGB: identification of ubiquitin proteins using machine learning model. J Mt Area Res 8:14–26
48. Ghulam A et al (2021) Identification of novel protein sequencing SARS CoV-2 coronavirus using machine learning. Biosci Res 18:47–58
49. Ghulam A et al (2023) DeepImmuno-PSSM: identification of immunoglobulin based on deep learning and PSSM-profiles. V Fast 11:54–66
50. Wong GY, Leung FH, Ling S-H (2013) Predicting protein-ligand binding site using support vector machine with protein properties. IEEE/ACM Trans Comput Biol Bioinform (TCBB) 10(6):1517–1529
51. Khan ZU et al (2019) iRSpot-SPI: deep learning-based recombination spots prediction by incorporating secondary sequence information coupled with physio-chemical properties via Chou's 5-step rule and pseudo components. Chemom Intell Lab Syst 189:169–180
52. Khan IA et al (2021) A privacy-conserving framework based intrusion detection method for detecting and recognizing malicious behaviours in cyber-physical power networks. Appl Intell 51:1–16
53. Ullah M et al (2018) A foreground extraction approach using convolutional neural network with graph cut. In 2018 IEEE 3rd international conference on image, vision and computing (ICIVC), pp. 40–44
54. Khan ZU et al (2021) piEnPred: a bi-layered discriminative model for enhancers and their subtypes via novel cascade multi-level subset feature selection algorithm. Front Comp Sci 15(6):1–11
55. Mandle AK, Jain P, Shrivastava SK (2012) Protein structure prediction using support vector machine. Int J Soft Comput 3:67–78
56. Khan A et al (2022) Prediction of antifreeze proteins using machine learning. Sci Rep 12(1):1–10
57. Khan A et al (2022) Comparative analysis of the existing methods for prediction of antifreeze proteins. Chemom Intell Lab Syst 232:104729
58. Dehzangi, A. and B.G. Khosravi. (2010) Introducing novel physicochemical based features to enhance protein fold prediction accuracy. In computer design and applications (ICCDA), 2010 international conference on. IEEE.
59. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
60. Ali F et al (2022) Target-DBPPred: an intelligent model for prediction of DNA-binding proteins using discrete wavelet transform based compression and light eXtreme gradient boosting. Comput Biol Med 145:105533

61. Barukab O et al (2022) DBP-CNN: deep learning-based prediction of DNA-binding proteins by coupling discrete cosine transform with two-dimensional convolutional neural network. Expert Syst Appl 197:116729
62. Barukab O, Ali F, Khan SA (2021) DBP-GAPred: an intelligent method for prediction of DNA-binding proteins types by enhanced evolutionary profile features with ensemble learning. J Bioinform Comput Biol 19:2150018
63. Ghulam A et al (2022) Accurate prediction of immunoglobulin proteins using machine learning model. Inform Med Unlocked 29:100885
64. Nanni L et al (2012) Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information. IEEE/ACM Trans Comput Biol Bioinf 9(2):467–475
65. Ghulam A et al (2022) ACP-2DCNN: deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network. Chemom Intell Lab Syst 226:104589
66. Dirvanauskas D et al (2019) Hemigen: human embryo image generator based on generative adversarial networks. Sensors 19(16):3578
67. Cao Y et al (2017) Unsupervised diverse colorization via generative adversarial networks. Joint European conference on machine learning and knowledge discovery in databases. Springer, Cham
68. Antoniou, A., Storkey, A. and Edwards, H., (2017) Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340
69. Akbar S et al (2022) cACP-DeepGram: classification of anticancer peptides via deep neural network and skip-gram-based word embedding model. Artif Intell Med 131:102349
70. Akbar S et al (2020) iRNA-PseTNC: identification of RNA 5-methylcytosine sites using hybrid vector space of pseudo nucleotide composition. Front Comp Sci 14(2):451–460
71. Akbar S et al (2019) iAFP-gap-SMOTE: an efficient feature extraction scheme gapped dipeptide composition is coupled with an oversampling technique for identification of antifreeze proteins. Lett Org Chem 16(4):294–302
72. Akbar S et al (2020) cACP-2LFS: classification of anticancer peptides using sequential discriminative model of KSAAP and two-level feature selection approach. IEEE Access 8:131939–131948
73. Zhang D et al (2017) Sharp and real image super-resolution using generative adversarial network. In international conference on neural information processing. Springer, Cham
74. Zhang K et al (2019) Stock market prediction based on generative adversarial network. Procedia Comput Sci 147:400–406
75. Schapire RE (2003) The boosting approach to machine learning: An overview. Nonlinear estimation and classification. Springer, New York, pp 149–171
76. Schapire, R.E. (1999) A brief introduction to boosting. In Ijcai.
77. Ali F et al (2019) DP-BINDER: machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information. J Comput Aided Mol Des 33(7):645–658
78. Tahir M, Hayat M, Khan SA (2019) iNuc-ext-PseTNC: an efficient ensemble model for identification of nucleosome positioning by extending the concept of Chou's PseAAC to pseudo-tri-nucleotide composition. Mol Genet Genomics 294(1):199–210
79. Akbar S et al (2017) iACP-GAEnsC: evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space. Artif Intell Med 79:62–70
80. Xiao X, Hui M, Liu Z (2016) iAFP-Ense: an ensemble classifier for identifying antifreeze protein by incorporating grey model and PSSM into PseAAC. J Membr Biol 249(6):845–854
81. Liu B et al (2018) iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach. Bioinformatics 34(22):3835–3842
82. Ali F et al (2022) DBP-iDWT: improving DNA-binding proteins prediction using multi-perspective evolutionary profile and discrete wavelet transform. Comput Intell Neurosci 2022:1–18
83. Ali F et al (2022) DBP-DeepCNN: prediction of DNA-binding proteins using wavelet-based denoising and deep learning. Chemom Intell Lab Syst 229:104639