frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 08 February 2023
DOI 10.3389/fgene.2022.1103783

OPEN ACCESS

EDITED BY
Suyan Tian,
Jilin University, China

REVIEWED BY
Bin Yang,
Zaozhuang University, China
Yuan Gao,
China University of Mining and Technology, China

*CORRESPONDENCE
Zhiyuan Li,
zhiyuanleeaa@126.com

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 21 November 2022
ACCEPTED 30 December 2022
PUBLISHED 08 February 2023

CITATION
Zhang Y and Li Z (2023), RF_phage virion: Classification of phage virion proteins with a random forest model.
Front. Genet. 13:1103783.
doi: 10.3389/fgene.2022.1103783

RF_phage virion: Classification of phage virion proteins with a random forest model

Yanqin Zhang[1] and Zhiyuan Li[2]*

[1]School of Finance, Xuzhou University of Technology, Xuzhou, China, [2]School of Artificial Intelligence and Software College, Jiangsu Normal University Kewen College, Xuzhou, China

Introduction: Phages play essential roles in biological procession, and the virion proteins encoded by the phage genome constitute critical elements of the assembled phage particle.

Methods: This study uses machine learning methods to classify phage virion proteins. We proposed a novel approach, RF_phage virion, for the effective classification of the virion and non-virion proteins. The model uses four protein sequence coding methods as features, and the random forest algorithm was employed to solve the classification problem.

Results: The performance of the RF_phage virion model was analyzed by comparing the performance of this algorithm with that of classical machine learning methods. The proposed method achieved a specificity (Sp) of 93.37%%, sensitivity (Sn) of 90.30%, accuracy (Acc) of 91.84%, Matthews correlation coefficient (MCC) of .8371, and an F1 score of .9196.

KEYWORDS

phage virion proteins, classification, bioinformatics, machine learning, random forest

1 Introduction

Phages integrate their DNA sequences with bacterial genomes following infection and play a role in maintaining the diversity of microorganisms (Shen et al., 2007; Xia et al., 2010; Wetie Ngounou et al., 2014; Zou et al., 2016). If the abundance of a particular type of bacteria increases rapidly in a bacterial population, the corresponding phage specifically infects and kills the rapidly proliferating bacteria. The entire bacterial population returns to equilibrium following this process. Phages also participate in the Earth's material cycle and are essential to the human microbiome (Brohee and Van Helden, 2006; Shen et al., 2019; Zhang and Quan, 2020). There are approximately 10^{14} bacteria in each individual's gut, while the number of bacteriophages is 10^{15-16}, which is ten times higher than the number of bacteria. These findings indicate that phage proteins play several crucial roles in biological processes (Ngo et al., 1994; Godzik et al., 1995; Whisstock and Lesk, 2003; Wu et al., 2009; De Las Rivas and Fontanillo, 2010; Awais et al., 2019).

Phage proteins can be classified as virion and non-virion proteins. The virion proteins encoded by the phage genes are essential components of the assembled phage particle and include the capsid protein, envelope protein, and virion enzymes (Chatterjee et al., 2011; You et al., 2013; Peng et al., 2017). These virion proteins determine the specificity for recognizing host bacteria and play essential roles in the recombination of phage viruses, receptor recognition, bacterial attachment, and penetration. The non-virion proteins of phages are synthesized in the infected cells and are also encoded by the phage genome. However, the non-virion proteins cannot be packaged into mature phage particles. The non-virion proteins primarily include enzymes and regulatory proteins, which play important roles in the processes

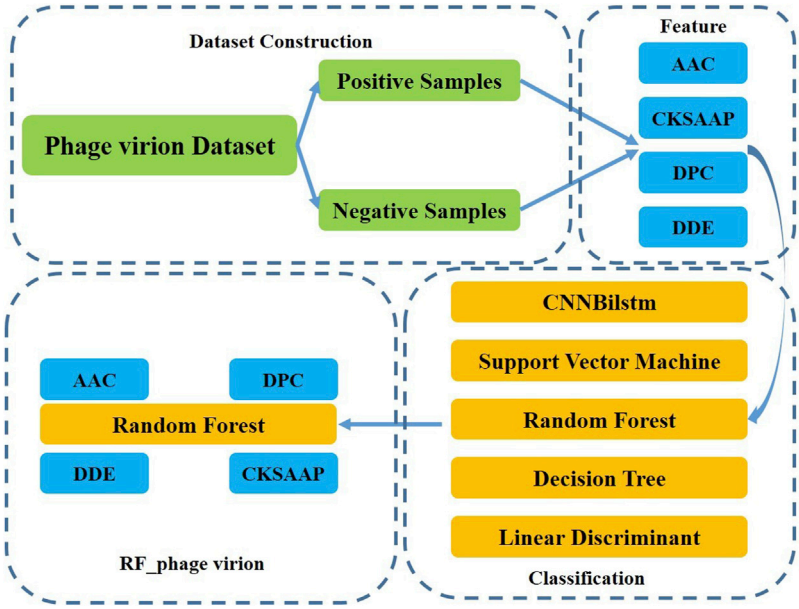

FIGURE 1
The outlines of RF_phage virion.

TABLE 1 The information of dataset.

Non-phage virion proteins	Phage virion proteins
500	500

of gene replication, transcription, and gene expression in phages (Sato et al., 1994; Schwikowski et al., 2000; Wei et al., 2017).

Several computational methods have been reported for classifying the functions of phage genes and virion proteins over the past few decades. Li et al. proposed a novel tool named SynFPS for classifying closely related genomes in whole genome comparison studies (Coates and Hall, 2003). The method employs a support vector machine (SVM) classifier and uses gene-to-gene distances as a feature. Feng et al. proposed a naïve Bayes method for classifying phage virion proteins based on the composition of primary amino acids and dipeptides as coding schemes (Free et al., 2009). Ding et al. proposed a method for classifying virion proteins using an SVM-based approach (Kim and Subramaniam, 2006). In these models, the key features among g-gap dipeptide compositions were initially determined by analysis of variance. Yang et al. described an ensemble algorithm-based method for classifying organellar proteins, in which the amino acid composition, physicochemical properties, sequence distribution, and structural characteristics of the sequences were used as features (Zhang et al., 2012). Han et al. proposed a two-layer multi-class SVM model for classifying subcellular localizations (Vazquez et al., 2003). After the first layer of SVM classification is completed, each amino acid sequence is represented by a k-dimensional vector, and each element in the vector corresponds to a classification result of the classifier (Yang et al., 2020). The output of the first layer is used as the input for the next layer, and the second layer uses SVM to determine the final result. Jia et al. proposed a random forest algorithm-based method that used different features extracted from protein sequences (You et al., 2017). The method used a voting system for computing the final classification results, which depended on seven independent models. Bahri et al. proposed an ensemble method named Greedy-Boost based on the adaptive combination, which improves the accuracy of detection (Guo et al., 2008). Although the smoothing method improves the stability of the classification system, the method has a high computational cost. Zhang et al. proposed a method based on logistic models for classifying samples using the amino acid composition, transformation, and distribution features and pseudo-amino acid composition as features (Koike and Takagi, 2004). The final results were computed based on the results obtained from the classification models. Liu et al. used different weights for classifying the four SVMs used in their study (You et al., 2015a). The method determined the final classification by traversing and selecting appropriate parameters. These findings indicate that ensemble algorithms can improve the accuracy of the final classification.

This study aimed to develop a method for the classification of phage virion proteins using machine learning methods. A novel method, RF_phage virion, is proposed herein for the effective classification of the virion and non-virion proteins. The method uses four protein sequence coding methods as features, and the random forest algorithm is used for solving the classification problem. The performance of the RF_phage virion model was determined by comparing the performance of this algorithm with some classical machine learning methods. A schematic representation of the RF_phage virion method is provided in Figure 1.

2 Materials and methods

2.1 Dataset

Ding's dataset, which primarily focuses on phage virion proteins, was used for classifying the phage proteins in this study (Bradford and

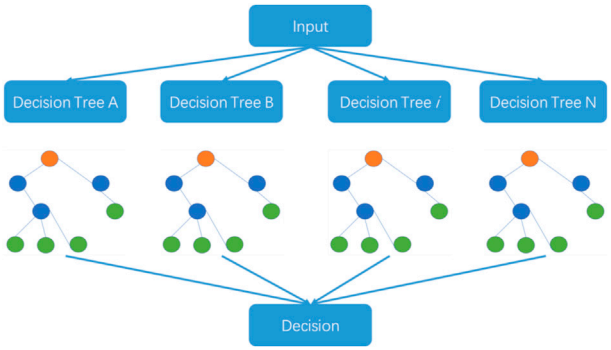

FIGURE 2
The structure of the random forest algorithm.

Westhead, 2005; Cui et al., 2012; Romero-Molina et al., 2019). Ding's dataset comprises 1000 samples, of which phage virion proteins constitute one-half, and the other half comprises non-phage virion proteins. The dataset can be treated as a typical ideal dataset for the classification of phage virion proteins. There is a large difference between the number of non-phage and phage virion proteins. Therefore, Ding's dataset can be considered an ideal benchmark dataset for phage virion protein classification problems. The detailed information of the employed dataset is demonstrated in Table 1.

2.2 Encoding methods

2.2.1 Amino acid composition (AAC)

The AAC feature describes the distribution of amino acid residues (Li et al., 2012). The feature focuses on the frequency of occurrence of each amino acid residue. At the same time, the AAC feature can provide typical statistical information regarding the identified protein sequences. The formula used for determining the AAC is provided in Eq. 1:

$$\mathrm{AAC} = \frac{\mathrm{aac}(i)}{length}, i \in \{A, C, \cdots, Y\} \quad (1)$$

Where, *length* represents the length of the identified phage virion protein sequence, and aac(i) represents the occurrence of the ith amino acid residue in the protein sequence. The parameter i refers to the twenty amino acids present in protein sequences. The sum of the twenty amino acids equals to 1.

2.2.2 Composition of k-spaced amino acid pairs (CKSAAP)

Although the AAC feature includes the amino acids present in protein sequences, the feature does not provide any positional information regarding the amino acids in protein sequences (Chen and Liu, 2005; You et al., 2015b; Wang et al., 2018). The CKSAAP feature describes the relationship between two amino acid residues in protein sequences, and focuses on the frequency of amino acid residue pairs, which are separated by n number of neighboring

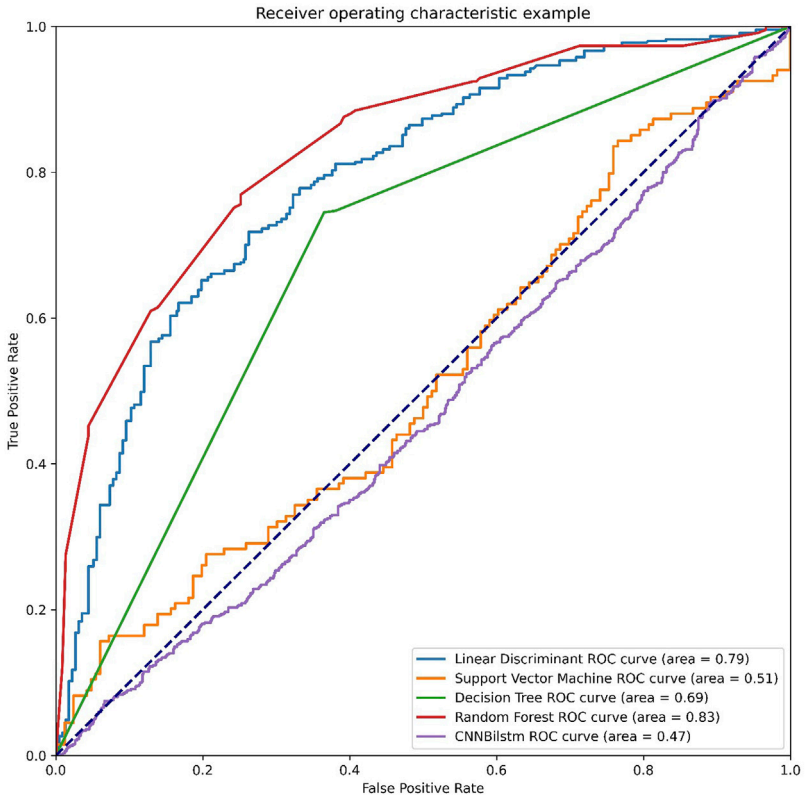

FIGURE 3
The ROC curves of AAC feature. Note: LD means the linear discriminant classifier. SVM means the support vector machine. DT means the decision tree. RF means the random forest and CNNBilstm mean convolution neural network with Bi-Long Short Term Memory.

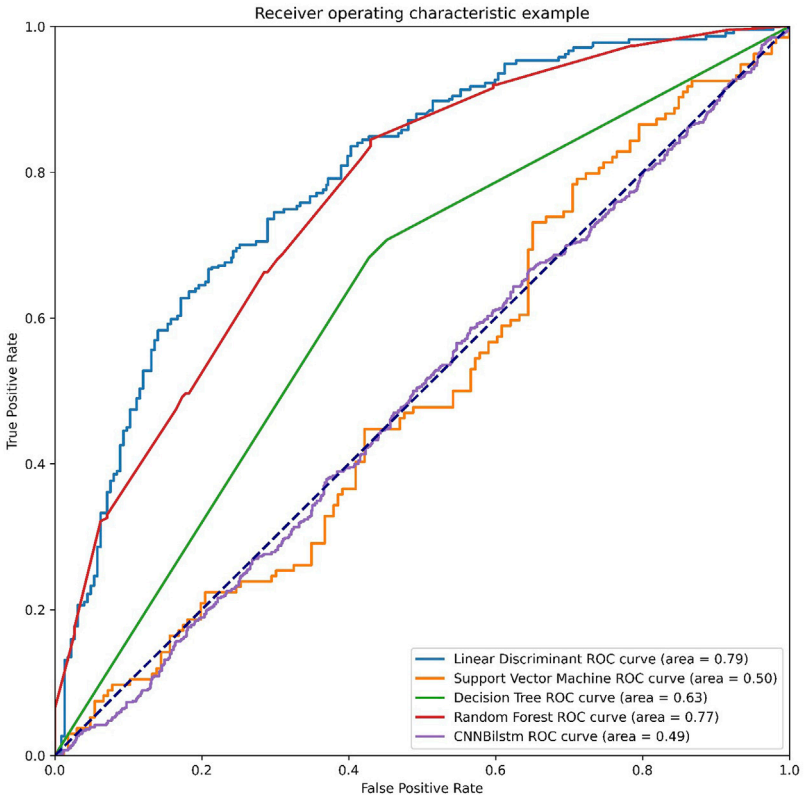

FIGURE 4
The ROC curves of CSKAAP feature.

TABLE 2 The performances of AAC feature.

	SP (%)	SN (%)	Acc (%)	MCC	F1 score
LD	74.16	70.07	72.12	4427	7268
SVM	46.99	52.24	49.61	−0077	4825
DT	63.47	74.50	68.99	3821	6718
RF	74.83	76.94	75.89	5178	7563
CNNBilstm	99.47	00	49.74	−0514	6643

TABLE 3 The performances of CSKAAP feature.

	SP (%)	SN (%)	Acc (%)	MCC	F1 score
LD	77.73	67.18	72.46	4516	7384
SVM	21.69	84.33	53.01	0772	3158
DT	57.24	68.29	62.77	2569	6059
RF	69.04	68.74	68.89	3778	6894
CNNBilstm	98.95	00	49.47	−0727	6620

amino acid residues. For instance, n = 0 indicates that the two amino acids are successive. There are 400 types of AACs, and CKSAAP can compute the frequency of occurrence for each combination. The formula used for determining the CKSAAP is provided in in Eq. 2:

$$CKSAAP\ (n=0) = \left(\frac{N_{AA}}{N_{total}}, \frac{N_{AC}}{N_{total}}, \cdots, \frac{N_{AY}}{N_{total}}, \cdots, \frac{N_{YY}}{N_{total}}\right)_{400} \quad (2)$$

In this study, the value of n was set to 3, and the scale of the CKSAAP feature can reach 1600.

2.2.3 Di-peptide composition (DPC)

The DPC feature focuses on the correlation between two successive amino acid residues (Sun et al., 2017). The scale of this feature can reach 400. The DPC feature was calculated using the formula in Eq. 3:

$$\mathrm{DPC} = \frac{bipeptide\,(i)}{length}, i \in \{AA, AC, \cdots, YY\} \quad (3)$$

Where, the sum of the whole elements equals 1. In other words, the DPC can be treated as a second-order term of amino acid pairs.

2.2.4 Dipeptide deviation extraction (DDE)

The DDE feature focuses on a binomial and uniform distribution theoretical sequence, but does not consider the alignment of protein relationships (Zhang et al., 2019). The feature can elucidate the interrelationships within a set of proteins. There DDE feature comprises three key parameters, namely, the size of the dipeptide

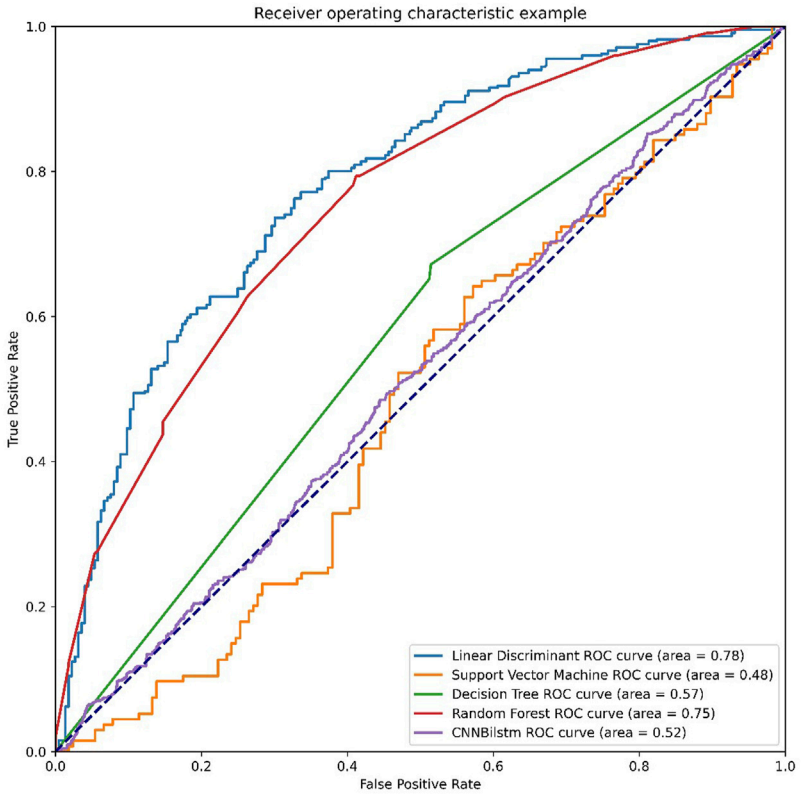

FIGURE 5
The ROC curves of DPC feature.

TABLE 4 The performances of DPC feature.

	SP (%)	SN (%)	Acc (%)	MCC	F1 score
LD	97.33	13.08	55.20	1932	6848
SVM	36.75	67.16	51.96	0411	4334
DT	48.78	65.19	56.98	1416	5314
RF	73.72	62.75	68.23	3669	6989
CNNBilstm	95.90	00	47.95	−1446	6482

composition (Dc), the means of theoretical values (T_m), and the theoretical value of variance (T_v). The formula used for calculating the DDE is depicted in Eq. 4:

$$DDE(t) = \frac{D_c(i) - T_m(i)}{\sqrt{T_V(i)}} \tag{4}$$

For instance, two pairs of successive amino acid residues have a DPC of 400. The scale of the DDE feature is 400, as depicted in Eq. 5:

$$DDE(t = 2) = \{dde_i\}, i \in [0, 400] \tag{5}$$

The formulae used for estimating the D_c, T_m, and T_v are provided in Eq. (6) (7) (8), provided hereafter.

$$D_c(i) = \frac{n_i}{N} \tag{6}$$

There are 400 combinations of amino acid pairs in each dipeptide. Therefore, the $D_c(i)$ can be treated as an element in related DPC features.

$$T_M(i) = \frac{C_{i1}}{C_N} \times \frac{C_{i2}}{C_N} \tag{7}$$

Where, T_m represents the theoretical average, C_{i1} represent the occurrence of the first amino acid residue, C_{i2} represents the occurrence of the second amino acid residue, and C_N represents the entire set of amino acids.

$$T_v(i) = \frac{T_M(i)(1 - T_M(i))}{N} \tag{8}$$

Where, T_v represents the theoretical variations in dipeptides.

2.3 Random forest algorithm

The random forest algorithm was proposed by L. Breiman at the beginning of this century and has been successfully used for dealing with classification and regression problems in related areas (Saha et al., 2014; Liu et al., 2018). The algorithm combines randomized decision trees and subsequently aggregates the average results from the decision trees. This algorithm can deal with high-dimensional small-sample problems. In other words, the algorithm performs well in identification problems using datasets where the scale of variables is much larger than the number of samples. The random forest algorithm is also used in big dataset

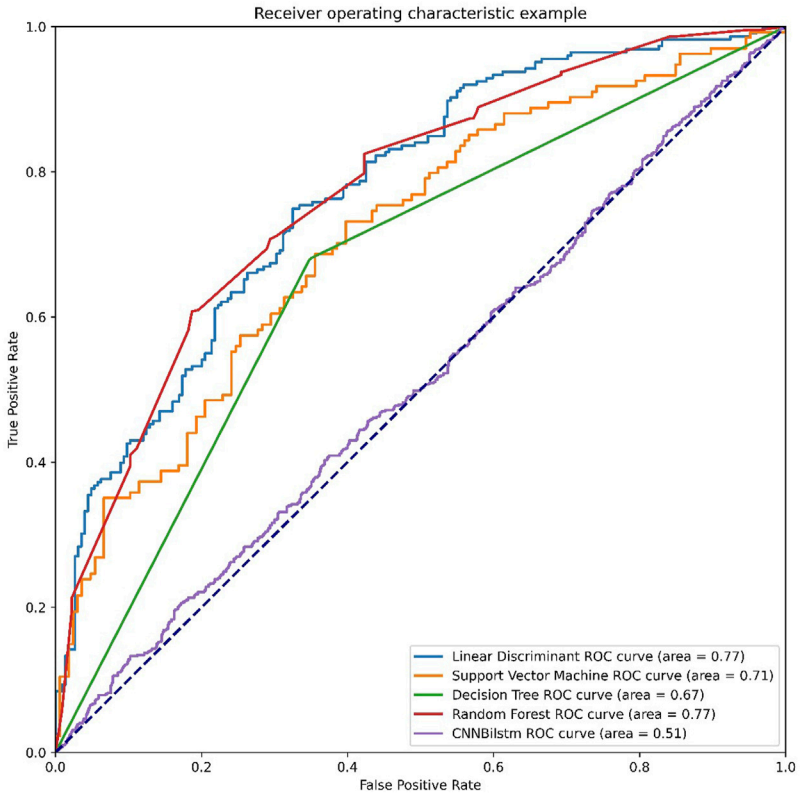

FIGURE 6
The ROC curves of DDE feature.

TABLE 5 The performances of DDE feature.

	SP (%)	SN (%)	Acc (%)	MCC	F1 score
LD	97.33	13.08	55.20	1932	6848
SVM	36.75	67.16	51.96	0411	4334
DT	48.78	65.19	56.98	1416	5314
RF	73.72	62.75	68.23	3669	6989
CNNBilstm	95.90	00	47.95	−1446	6482

problems. The steps of the random forest algorithm are outlined in Figure 2.

2.4 Measurement of performance

The samples in the classification problem in this study could be categorized into two, namely, phage and non-phage virion protein sequences. The defined positive samples comprised the virion protein sequences, while the defined negative samples comprised the non-phage protein sequences of phages. According to the definition, classified samples can produce four results under common conditions. These formulations, including the sensitivity (Sn), specificity (Sp), accuracy (ACC), F1 scores, and Matthews correlation coefficient (MCC), were obtained using the formulae in Eq. (4) (5) (6) (7) (8), provided hereafter.

$$Sn = \frac{TP}{TP + FN} \tag{4a}$$

$$Sp = \frac{TN}{TN + FP} \tag{5a}$$

$$\mathrm{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6a}$$

$$F1 = \frac{2TP}{2TP + FN + FP} \tag{7a}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{8a}$$

Where, P and N represent the scale of positive and negative samples, respectively. T and F represent sets of true and false predicted results, respectively.

The F1 score is used to evaluate the distribution of positive and negative samples in two-types problems. Performance measures should consider several parameters, including the four basic parameters, namely, TP, FP, TN, and FN. The performance measure can be treated as a harmonic average of model accuracy and recall. Another important measure of performance is the MCC, and the values of this performance measure ranges from −1 to 1.

3 Results

The random forest model was used in this study for classifying the virion and non-virion proteins of phages using four typical protein

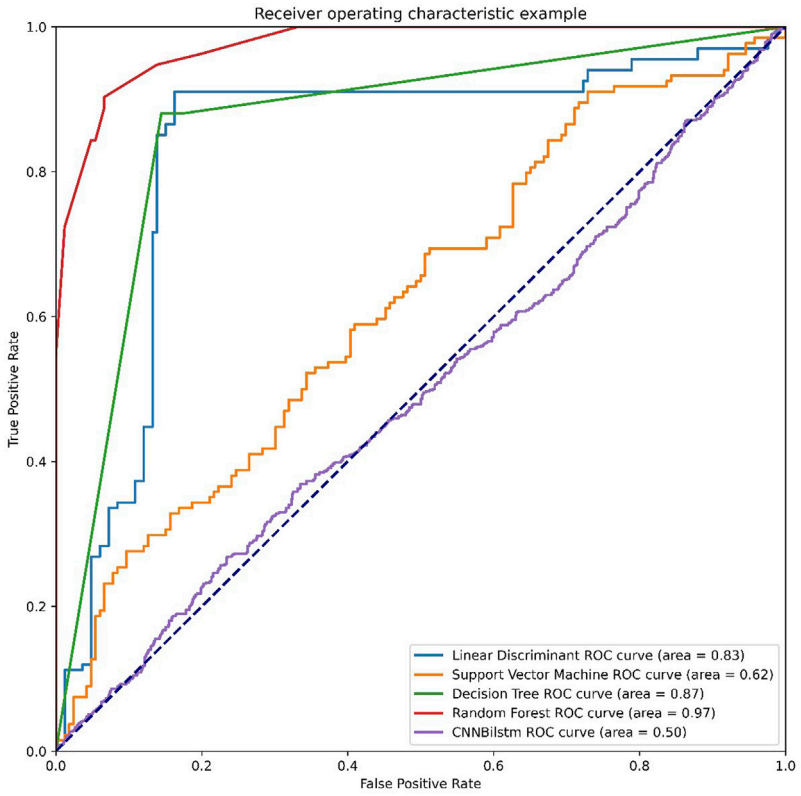

FIGURE 7
The ROC curves of combination feature.

TABLE 6 The performances of combination feature.

	SP (%)	SN (%)	Acc (%)	MCC	F1 score
LD	84.34	86.57	85.45	7092	8529
SVM	18.07	91.79	54.93	1460	2862
DT	83.73	88.06	85.90	7186	8559
RF	93.37	90.30	91.84	8371	9196
CNNBilstm	98.45	00	49.22	−0885	6597

features, namely, the AAC, CSKAAP, DPC, and DDE. The performance of the method was determined by comparing with state-of-the-art methods.

As depicted in Figure 3 and Table 2, the values of Sp, Sn, Acc, MCC, and F1 score for the SVM-based method were 46.99%, 52.24%, 49.61%, −.0077, and .4825, respectively, while the values of these indices for the decision tree model were 63.47%, 74.50%, 68.99%, .3821, and .6718, respectively. The values of Sp, Sn, Acc, MCC, and F1 score for the random forest algorithm using the AAC feature were 74.83%, 76.94%, 75.89%, .5178, and .7563, respectively, while the values of these indices for the deep learning algorithm, which is a convolution neural network, were 99.47%, 0%, 49.74%, −.0514, and .6643, respectively.

As depicted in Figure 4 and Table 3, the values of Sp, Sn, Acc, MCC, and F1 score for the SVM-based method were 21.69%, 84.33%, 53.01%, .0772, and .3158, respectively, while the values of these indices for the decision tree model were 57.24%, 68.29%, 62.77%, .2569, and .6059, respectively. The values of Sp, Sn, Acc, MCC, and F1 score for the random forest algorithm using the CSKAAP feature were 69.04%, 68.74%, 68.89%, .3778, and .6894, respectively, while the values of these indices for the convolution neural network were 98.95%, 0%, 49.47%, −.0727, and .6620, respectively.

As depicted in Figure 5 and Table 4, the values of Sp, Sn, Acc, MCC, and F1 score for the SVM-based method were 54.82%, 75.37%, 65.10%, .3085, and .611, respectively, while the values of these indices for the decision tree model were 65.26%, 67.85%, 66.55%, .3312, and .6611, respectively. The values of Sp, Sn, Acc, MCC, and F1 score for the random forest algorithm using the DPC feature were 65.26%, 67.85%, 66.55%, .3312, and .6611, respectively, while the values of these induces for the convolution neural network were 88.44%, 0%, 44.22%, −.2477, and .6132, respectively.

As depicted in Figure 6 and Table 5, the values of Sp, Sn, Acc, MCC, and F1 score for the SVM-based method were 36.75%, 67.16%, 51.96%, .0411, and .4334, respectively, while the values for the decision tree model were 48.78%, 65.19%, 56.98%, .1416, and .5314, respectively. The values of Sp, Sn, Acc, MCC, and F1 score for the random forest algorithm using the DDE feature were 73.72%, 62.75%, 68.23%, .3669, and .6989, respectively, while the values of these indices for the convolution neural network were 95.90%, 0%, 47.95%, −.1446, and .6482, respectively.

4 Discussions

In the section of results, we merely employed the AAC, CSKAAP, DPC, and DDE features, respectively. Therefore, we combined the four features to evaluate the performances in this work.

As depicted in Figure 7 and Table 6, the values of Sp, Sn, Acc, MCC, and F1 score for the SVM-based method were 18.07%, 91.79%, 54.93%, .1460, and .2862, respectively, while the values for the decision tree model were 83.73%, 88.06%, 85.90%, .7186, and .8559, respectively. The values of Sp, Sn, Acc, MCC, and F1 score for the random forest algorithm using the combination feature were 93.37%, 90.30%, 91.84%, .8371, and .9196, respectively, while the values of these indices for the convolution neural network were 98.45%, .00%, 49.22%, −.0885, and .6597, respectively.

5 Conclusion

The present study uses machine learning methods to classify phage virion proteins. Four protein sequence coding methods, namely AAC, CSKAAP, DPC, and DDE, were used as features for the effective classification of the virion and non-virion proteins. The random forest algorithm was subsequently used to solve the classification problem. By combining each of the four features with the classification algorithm, we observed that the performance of the model was best when the combination feature was used.

When it comes to the problem of classification of phage virion proteins, such an issue can be regarded as a typical binary classification problem in the field of machine learning. In this work, we employed Ding's dataset, which is a balanced dataset. Actually, the size of positive samples can hardly be equal to the size of the negative ones. In this work, the AAC, CSKAAP, DPC, and DDE feature and their combination feature can be employed as the input of the RF_phage virion model. There are several other features in the field of protein research. Therefore, these features can also be employed in future work. On the other hand, the other typical classification algorithm can be utilized in future work. The size of the combination feature can reach 2420. Considering such a situation, some reduced useless information approaches can be utilized in this future work.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

YZ designed the algorithm and ZL edited the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Awais, M., Hussain, W., Khan, Y. D., Rasool, N., Khan, S. A., and Chou, K. C. (2019). iPhosH-PseAAC: Identify phosphohistidine sites in proteins by blending statistical moments and position relative features according to the Chou's 5-step rule and general pseudo amino acid composition. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 18, 596–610. doi:10.1109/TCBB.2019.2919025

Bradford, J. R., and Westhead, D. R. (2005). Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics* 21.8, 1487–1494. doi:10.1093/bioinformatics/bti242

Brohee, S., and Van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinforma.* 7.1, 488. doi:10.1186/1471-2105-7-488

Chatterjee, P., Basu, S., Kundu, M., Nasipuri, M., and Plewczynski, D. (2011). PPI_SVM: Prediction of protein-protein interactions using machine learning, domain-domain affinities and frequency tables. *Cell. Mol. Biol. Lett.* 16.2, 264–278. doi:10.2478/s11658-011-0008-x

Chen, X-W., and Liu, M. (2005). Prediction of protein–protein interactions using random decision forest framework. *Bioinformatics* 21.24, 4394–4400. doi:10.1093/bioinformatics/bti721

Coates, P. J., and Hall, P. A. (2003). The yeast two-hybrid system for identifying protein–protein interactions. *J. Pathology A J. Pathological Soc. G. B. Irel.* 199.1, 4–7. doi:10.1002/path.1267

Cui, G., Fang, C., and Han, K. (2012). Prediction of protein-protein interactions between viruses and human by an SVM model. *BMC Bioinforma.* 13, S5. doi:10.1186/1471-2105-13-S7-S5

De Las Rivas, J., and Fontanillo, C. (2010). Protein–protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLoS Comput. Biol.* 6, e1000807. doi:10.1371/journal.pcbi.1000807

Free, R. B., Hazelwood, L. A., and Sibley, D. R. (2009). Identifying novel protein-protein interactions using co-immunoprecipitation and mass spectroscopy. *Curr. Protoc. Neurosci.* 46.1, Unit 5.28–28. doi:10.1002/0471142301.ns0528s46

Godzik, A., Skolnick, J., and Koliński, A. (1995). Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci.* 4.10, 2107–2117. doi:10.1002/pro.5560041016

Guo, Y., Yu, L., Wen, Z., and Li, M. (2008). Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic acids Res.* 36.9, 3025–3030. doi:10.1093/nar/gkn159

Kim, Y., and Subramaniam, S. (2006). Locally defined protein phylogenetic profiles reveal previously missed protein interactions and functional relationships. *Proteins Struct. Funct. Bioinforma.* 62.4, 1115–1124. doi:10.1002/prot.20830

Koike, A., and Takagi, T. (2004). Prediction of protein–protein interaction sites using support vector machines. *Protein Eng. Des. Sel.* 17.2, 165–173. doi:10.1093/protein/gzh020

Li, B-Q., Feng, K. Y., Chen, L., Huang, T., and Cai, Y. D. (2012). Prediction of protein-protein interaction sites by random forest algorithm with mRMR and IFS. *PloS one* 7.8, e43927. doi:10.1371/journal.pone.0043927

Liu, Q., Chen, P., Wang, B., Zhang, J., and Li, J. (2018). Hot spot prediction in protein-protein interactions by an ensemble system. *BMC Syst. Biol.* 12.9, 132–199. doi:10.1186/s12918-018-0665-8

Ngo, J. T., Marks, J., and Karplus, M. (1994). "Computational complexity, protein structure prediction, and the Levinthal paradox," in *The protein folding problem and tertiary structure prediction* (Birkhäuser Boston), 433–506. doi:10.1007/978-1-4684-6831-1_14

Peng, X., Wang, J., Peng, W., Wu, F. X., and Pan, Y. (2017). Protein–protein interactions: Detection, reliability assessment and applications. *Briefings Bioinforma.* 18.5, 798–819. doi:10.1093/bib/bbw066

Romero-Molina, S., Ruiz-Blanco, Y. B., Harms, M., Munch, J., and Sanchez-Garcia, E. (2019). PPI-detect: A support vector machine model for sequence-based prediction of protein–protein interactions. *J. Comput. Chem.* 40.11, 1233–1242. doi:10.1002/jcc.25780

Saha, I., Zubek, J., Klingstrom, T., Forsberg, S., Wikander, J., Kierczak, M., et al. (2014). Ensemble learning prediction of protein–protein interactions using proteins functional annotations. *Mol. Biosyst.* 10.4, 820–830. doi:10.1039/c3mb70486f

Sato, T., HanadaM., Bodrug, S., Irie, S., IwamaN., Boise, L. H., et al. (1994). Interactions among members of the Bcl-2 protein family analyzed with a yeast two-hybrid system. *Proc. Natl. Acad. Sci.* 91.20, 9238–9242. doi:10.1073/pnas.91.20.9238

Schwikowski, B., Peter, U., and Fields, S. (2000). A network of protein–protein interactions in yeast. *Nat. Biotechnol.* 18.12, 1257–1261. doi:10.1038/82360

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., et al. (2007). Predicting protein–protein interactions based only on sequences information. *Proc. Natl. Acad. Sci.* 104.11, 4337–4341. doi:10.1073/pnas.0607879104

Shen, Z., Yuan, L., and Zou, Q. (2019). Transcription factors–DNA interactions in rice: Identification and verification. *Briefings Bioinforma.* 21, 946–956. doi:10.1093/bib/bbz045

Sun, T., Zhou, B., Lai, L., and Pei, J. (2017). Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinforma.* 18.1, 277–278. doi:10.1186/s12859-017-1700-2

Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nat. Biotechnol.* 21.6, 697–700. doi:10.1038/nbt825

Wang, L., You, Z. H., Yan, X., Xia, S. X., Liu, F., Li, L. P., et al. (2018). Using two-dimensional principal component analysis and rotation forest for prediction of protein-protein interactions. *Sci. Rep.* 8.1, 12874–12910. doi:10.1038/s41598-018-30694-1

Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017). Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intell. Med.* 83, 67–74. doi:10.1016/j.artmed.2017.03.001

Wetie Ngounou, A. G., Sokolowska, I., Woods, A. G., Roy, U., Deinhardt, K., and Darie, C. C. (2014). Protein–protein interactions: Switch from classical methods to proteomics and bioinformatics-based approaches. *Cell. Mol. life Sci.* 712, 205–228. doi:10.1007/s00018-013-1333-1

Whisstock, J. C., and Lesk, A. M. (2003). Prediction of protein function from protein sequence and structure. *Q. Rev. biophysics* 36.3, 307–340. doi:10.1017/s0033583503003901

Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Makela, T. P., and Hautaniemi, S. (2009). Integrated network analysis platform for protein-protein interactions. *Nat. methods* 6.1, 75–77. doi:10.1038/nmeth.1282

Xia, J-F., Han, K., and Huang, D-S. (2010). Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor. *Protein Peptide Lett.* 17.1, 137–145. doi:10.2174/092986610789909403

Yang, X., Yang, S., Li, Q., Wuchty, S., and Zhang, Z. (2020). Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. *Comput. Struct. Biotechnol. J.* 18, 153–161. doi:10.1016/j.csbj.2019.12.005

You, Z-H., Chan, K. C. C., and Hu, Pengwei (2015). Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. *PloS one* 10.5, e0125811. doi:10.1371/journal.pone.0125811

You, Z-H., Lei, Y. K., Zhu, L., Xia, J., and Wang, B. (2013). Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis, *BMC Bioinforma.* 14, S10. doi:10.1186/1471-2105-14-S8-S10

You, Z-H., Li, J., Gao, X., He, Z., Zhu, L., Lei, Y-K., et al. (2015). Detecting protein-protein interactions with a novel matrix-based protein sequence representation and support vector machines. *BioMed Res. Int.* 2015, 867516. doi:10.1155/2015/867516

You, Z-H., Xiao, L., and Chan, K. C. C. (2017). An improved sequence-based prediction protocol for protein-protein interactions using amino acids substitution matrix and rotation forest ensemble classifiers. *Neurocomputing* 228, 277–282. doi:10.1016/j.neucom.2016.10.042

Zhang, L., Yu, G., Xia, D., and Wang, J. (2019). Protein–protein interactions prediction based on ensemble deep neural networks. *Neurocomputing* 324, 10–19. doi:10.1016/j.neucom.2018.02.097

Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., et al. (2012). Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature* 490, 7421556–7421560. doi:10.1038/nature11503

Zhang, Y. P., and Quan, Z. (2020). Pptpp: A novel therapeutic peptide prediction method using physicochemical property encoding and adaptive feature representation learning. *Bioinformatics* 36, 3982–3987. doi:10.1093/bioinformatics/btaa275

Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 173, 346–354. doi:10.1016/j.neucom.2014.12.123