

## Research Article

# Predicting Cross-Species Infection of Swine Influenza Virus with Representation Learning of Amino Acid Features

Zheng Kou , Junjie Li, Xinyue Fan, Saeed Kosari, and Xiaoli Qiang 

*Institute of Computing Science and Technology, Guangzhou University, Guangzhou 510006, China*

Correspondence should be addressed to Zheng Kou; [kouzheng@gzhu.edu.cn](mailto:kouzheng@gzhu.edu.cn) and Xiaoli Qiang; [qiangxl@gzhu.edu.cn](mailto:qiangxl@gzhu.edu.cn)

Received 24 August 2021; Revised 27 September 2021; Accepted 28 September 2021; Published 11 October 2021

Academic Editor: Tao Huang

Copyright © 2021 Zheng Kou et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Swine influenza viruses (SIVs) can unforeseeably cross the species barriers and directly infect humans, which pose huge challenges for public health and trigger pandemic risk at irregular intervals. Computational tools are needed to predict infection phenotype and early pandemic risk of SIVs. For this purpose, we propose a feature representation algorithm to predict cross-species infection of SIVs. We built a high-quality dataset of 1902 viruses. A feature representation learning scheme was applied to learn feature representations from 64 well-trained random forest models with multiple feature descriptors of mutant amino acid in the viral proteins, including compositional information, position-specific information, and physicochemical properties. Class and probabilistic information were integrated into the feature representations, and redundant features were removed by feature space optimization. High performance was achieved using 20 informative features and 22 probabilistic information. The proposed method will facilitate SIV characterization of transmission phenotype.

## 1. Introduction

The influenza A virus (family: Orthomyxoviridae) genome contains eight segmental RNAs [1]. The hemagglutinin (HA) gene is located in the fourth segment, and the neuraminidase (NA) gene is in the sixth segment. According to the antigenic characteristics of HA and NA, the influenza A virus has 18 HA subtypes and 11 NA subtypes [2–4]. Besides the fast mutation rates of viral genes, segmental reassortments of viral genomes facilitate the occurrence of novel virus with the changes of host barriers [5, 6]. The 1957 pandemic was caused by an influenza A (H1N1) virus, which has a genome that contains segments 2, 4, and 5 from the avian influenza virus, whereas the 1968 pandemic was caused by an influenza A (H3N2) virus, which has a genome that contains segments 2 and 4 from the avian influenza virus [1].

Swine influenza virus (SIV) is an influenza A virus that causes acute respiratory infectious disease of swine [7]. There are three main SIV subtypes circulating worldwide (H1N1, H1N2, and H3N2), and these subtypes can be subdivided into different genotypes, including the classical H1N1,

avian-like H1N1, human-like H3N2, reassorted H3N2, and H1N2 subgroups [8–10]. SIV genotypes are diverse, and its gene pool in nature is heterogeneous. SIV can cross the species barriers unforeseeably and directly infect humans. The 2009 H1N1 pandemic killed more than 18,000 people, and the viral pathogen during the pandemic was a novel SIV that was produced by genome reassortment between genome-reassorted swine viruses from America and Europe [11–13]. H3N2 variant viruses (H3N2v), which have segment 7 from the 2009 H1N1 pandemic virus in their genomes, were identified in swine in 2010 and first detected in people in 2011 [14]. Clinical cases of H1N1 variant viruses (H1N1v) and H1N2 variant viruses (H1N2v) were also been reported after 2011 [15]. SIVs are a huge challenge for human public health and may trigger pandemic risk.

Computational bioinformatics tools are needed to predict transmission phenotype and pandemic risk of SIVs. For this purpose, machine-learning methods may be ideal tools [16–18]. Machine learning techniques have great potential for virus screening because they can use viral protein sequences as input without the need for prior knowledge. In this paper, we propose a feature representation

algorithm to predict cross-species infection of SIVs. The algorithm includes the sequence-based feature descriptors to build a comprehensive predictive model with sufficient information from different aspects. Sequence-based features with class information or probabilistic information are learnt from well-trained random forest (RF) classifiers that can learn a set of features. The dimension of the feature space was reduced using the minimum redundancy maximum relevance (mRMR) method to obtain the most informative features and distinguish SIVs with different transmission phenotypes.

To identify SIVs capable of interspecies transmission, we constructed a predictor with two predictive models that were trained using 20 features based on class information or 22 features based on probabilistic information under the RF classifier. The predictor with the feature representation learning achieved a high prediction performance. This study provides an important tool in predicting cross-species infection of SIVs for public health.

## 2. Materials and Methods

**2.1. Data.** Viral sequences of influenza viruses isolated from swine and human were downloaded from the GISAID EpiFlu public database (<http://platform.gisaid.org/epi3/frontend>) [2, 3]. GISAID deposits high-quality genomic sequences along with their clinical information in the database. Since sequence redundancy was very high and genome coverage varied greatly, raw data were filtered using public bioinformatics tools and algorithms (Table S1).

We obtained 5860 SIVs and 44,623 human influenza viruses from the GISAID database on 21 March 2019. The dataset included all of the 11 influenza virus proteins (PB2, PB1, PB1-F2, PA, HA, NP, NA, M1, M2, NS1, and NEP) encoded in eight genome segments. Strains without any of the 11 protein sequences or without subtype information were excluded. Amino acid positions in the 11 proteins were determined using the multiple sequence alignment tool MUSCLE [19]. Strains with more than three amino acids missing at the terminal ends of the viral proteins were removed, and if there were only a few missing residues, they were added according to those in viral proteins with highest identity. We used the fast-clustering algorithm of the CD-Hit tool to reduce the redundancy in the dataset [20]. Ambiguous amino acid residues, such as X and B, were likely caused by sequencing error and were replaced by those in viral protein with highest identity. Strains with large numbers of ambiguous residues in viral protein were also removed.

The final dataset for predicting cross-species infection contained two categories of viruses: (1) 769 viruses isolated from human (positive sample; H1N1, H1N2, H2N2, and H3N2 subtypes); (2) 1133 influenza viruses isolated from swine (negative sample; H1N1, H1N2, and H3N2 subtypes). The positive samples were composed by seasonal human influenza virus, 2009 pandemic swine virus, and variant swine virus isolated from human. Since these viruses could be also isolated from swine [1], they were excluded from the negative samples according to the similarity of genome

sequence. Information about the 1902 strains is summarized in Table S1.

**2.2. Signature Amino Acid Positions Based on Entropy.** Most of the amino acid residues in the viral proteins were conserved. To reduce the computing complexity, amino acid residues were filtered by the entropy measure at each position of the 11 viral proteins. For a given position  $i$ , the entropy value was computed using the formula [21]:  $E_i = -\sum_{j=1}^{20} P_{i,j} \log(P_{i,j})$ , where  $P_{i,j}$  is the observed probability of amino acid  $j$  at position  $i$ . High entropy values indicate high amino acid mutation rates at the corresponding position. We set the threshold of entropy difference as 1.5 and obtained 36 signature positions, and therefore, each strain was represented by a list of 36 amino acid residues in the screened positions.

**2.3. Representation of Signature Amino Acid Set.** Mutations in the viral proteins determine the pathogenicity or virulence of SIVs [1]. After obtaining the entropy ranking for each position, 36 significant amino acids were screened. Six encoding algorithms for compositional information, position-specific information, and physicochemical properties of amino acids were used to explore the key information required for high-quality predictions [16]. The encoding algorithms for the signature amino acid set to transform SIV into fix-length vectors are detailed below.

**2.3.1. Amino Acid Composition.** The amino acid composition (AAC) is a 20-dimension vector as usual. Because the gaps (deletion or insertion) in viral proteins occurred frequently during the evolution of SIV, we defined the AAC as a 21-dimension vector to represent the frequency of the 20 amino acid residues and one gap in the 36 signature positions of the viral proteins. For example, if the amino acid type  $i$  occurs  $n_i$  times in the amino acid set of a specific virus, the frequency of  $i$  is denoted as  $f(i) = n_i/36$ . A 21-dimensional feature vector that represents the frequencies of the 20 different amino acids and one gap was obtained for each strain.

**2.3.2. Parallel Correlation-Based Pseudo-Amino-Acid Composition.** Parallel correlation-based pseudo-amino-acid composition (PC-PseAAC) method was used to compute the parallel correlation of any two amino acids in the 36 signature amino acid positions in the viral protein sequences [22]. For a virus D, the PC-PseAAC feature vector was defined by

$$\text{PC-PseAAC} = [p_1, \dots, p_{21}, p_{21+1}, \dots, p_{21+\lambda}]^T, \quad (1)$$

where

$$p_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{21} f_i + 0.05 \sum_{j=1}^{\lambda} \theta_j}, & 1 \leq u \leq 21, \\ \frac{0.05 \theta_{u-21}}{\sum_{i=1}^{21} f_i + 0.05 \sum_{j=1}^{\lambda} \theta_j}, & 21 + 1 \leq u \leq 21 + \lambda, \end{cases} \quad (2)$$

where  $u$  is an integer,  $fi(1 \leq i \leq 21)$  represents the normalized occurrence frequency of the 20 amino acids and one gap in virus  $D$ ,  $\lambda$  is the highest tier of the correlation along  $D$ , and  $\theta_j$  is the correlation function that measures the  $j$ -tier sequence-order correlation between all the  $j$ -th most contiguous residues along  $D$ . The  $\theta_j$  function is given as

$$\theta_j = \frac{1}{36} \sum_{i=1}^L \frac{1}{5} \sum_{m=1}^5 [H_m(A_{i+j}) - H_m(A_i)]^2, \quad (3)$$

where  $Hm(A_i)(m = 1, 2, 3, 4, 5)$  represents the five amino acid factors that correspond to the  $i$ -th amino acid  $A_i$  in virus  $D$ , respectively [23]. If  $i + j$  is  $>36$ , then  $i + j$  equals  $i + j - 36$ . The five factors for gap (deletion or insertion) were simply set to zero.

**2.3.3. G-Gap Dipeptide Composition.** The G-gap dipeptide composition (GGAP) is the dipeptide composition coupled with local order information of any two interval residues among the 36 amino acid residues of the 11 viral proteins for each virus. The GGAP is commonly used feature descriptor for sequence analysis and model construction. In this paper, GGAP is a 441-dimension vector that represents the frequency of dipeptide comprising 20 amino acid residues and one gap. It is defined as

$$\text{GGAP}(g) = (p_1^g, p_2^g, \dots, p_{441}^g), \quad (4)$$

where  $p_i^g$  is the occurrence frequency of the  $i$ -th ( $i = 1, 2, \dots, 441$ ) G-gap dipeptide, which is defined as

$$p_i^g = \frac{O_i^g}{\sum_{i=1}^{441} O_i^g}, \quad (5)$$

where  $O_i^g$  is the occurrence number of the  $i$ -th G-gap dipeptide in the 36 signature amino acid residues. The dimension of the GGAP feature vector is  $21 \times 21 = 441$ . Deletion or insertion is also computed.

**2.3.4. Twenty-Bit Features.** Position-specific information and physicochemical properties were used to encode the 36 amino acid residues for each virus. Five physicochemical property descriptors of the standard amino acids were constructed, namely, polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge [23]. For each descriptor, the standard amino acid alphabets were classified into three groups, and the deletion/insertion (indel) was regarded as the fourth group. Representation of 20 standard amino acids and one indel was according to the five physicochemical properties. Each residue was encoded as a 20-bit vector comprising 0/1 elements, where the position of the bit was set to 1 if the residue belongs to the corresponding group; otherwise, it was 0. Given the amino acid augment approach, the top  $k$  residues with the highest entropy values were selected. The dimension of the feature vector was  $20 \times k$ .

**2.3.5. Twenty-One-Bit Features.** Twenty-one-bit feature was like a one-hot encoding. In this algorithm, each amino acid residue is transformed into a 21-bit 0/1 vector. (e.g., Ala by 1,0; indel by 0,1). Given the amino acid augment approach, each strain with the top  $k$  residues was represented by a  $21 \times k$  dimensional feature vector.

**2.3.6. Overlapping Property Features.** This algorithm divided the 20 standard amino acids and one gap (indel) into 11 different groups according to physicochemical properties. The distribution of the 20 stranded amino acids in the 10 groups can overlap [24]. The 10 amino acid groups were aromatic = {F, Y, W, H}, negative = {D, E}, positive = {K, H, R}, polar = {N, Q, S, D, E, C, T, K, R, H, Y, W}, hydrophobic = {A, G, C, T, I, V, L, K, H, F, Y, W, M}, aliphatic = {I, V, L}, tiny = {A, S, G, C}, charged = {K, H, R, D, E}, small = {P, N, D, T, C, A, G, S, V}, and proline = {A, S, G, C}. Indels form the 11th group. Each amino acid residue was represented by an 11-dimensional 0/1 vector. The position of the vector was set to 1 if the residue belongs to the physicochemical property group; otherwise, it was 0. Given the amino acid augment approach, the top  $k$  residues with the highest entropy values were selected. The amino acid augment was encoded with an  $11 \times k$  feature vector.

**2.4. Framework of Feature Representation Learning.** The framework of the feature representation learning algorithm, which includes two main steps, feature representation learning and feature representation optimization, is shown in Figure 1. Firstly, feature representations from a set of feature descriptors are generated using the RF classifier systems. Secondly, the feature representations learnt from the first step are optimized to yield informative feature subsets. The two-step feature representation learning procedure was as follows [16].

**2.4.1. Feature Representation Learning.** The six feature encoding algorithms were AAC, PC-PseAAC, GGAP, 20-bit features (BIT20), 21-bit features (BIT21), and overlapping property features (OLP), all of which are described above. A feature pool was built to generate as much information as possible in the predicting models with different parameters. For example,  $k$  is a common parameter for BIT20, BIT21, and OLP. Because the 36 significant amino acids were screened after the entropy ranking was obtained, we set  $k$  as 4-36 by step 4. The maximum  $k$  value was set as 36 because there were 36 signature positions, and therefore, a total of 27 feature descriptors are obtained for BIT20, BIT21, and OLP. A similar procedure was used for PC-PseAAC and GGAP. With the use of different parameters, a total of 64 feature descriptors were in the feature pool. Information about all the feature descriptors is provided in Table 1.

Before the optimization of feature representation, two types of predictions were used to fulfill the learning. All the 64 descriptors in the feature pool were used to train and predict with the RF models, and two types of predictions were achieved. The first prediction type was the class label

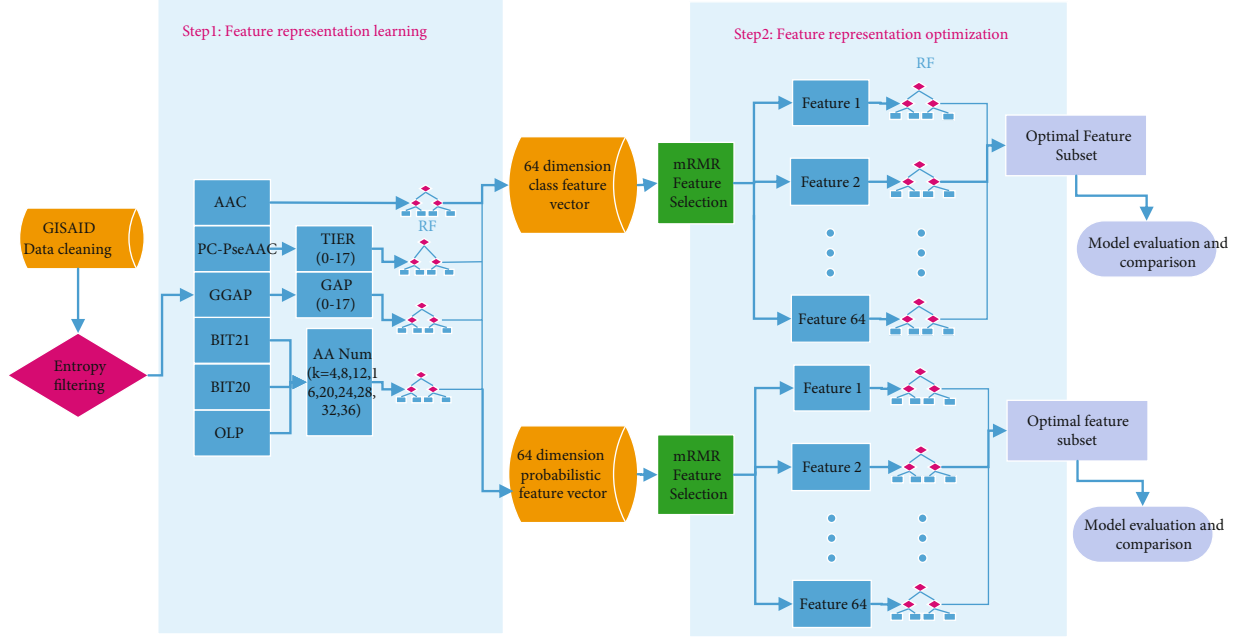


FIGURE 1: Flowchart of representation learning of amino acid features. After data cleaning, 36 signature amino acid positions based on entropy were screened. Six encoding algorithms with the change of parameter were used to explore the key information. All the 64 descriptors in the feature pool were used to train and predict with the RF models, and two types of predictions were achieved to be further optimized. Each swine virus was eventually represented by two optimized feature vectors, “class” and “prob.” Finally, the predictive models were built and compared.

(positive or negative): positive samples (swine viruses with the phenotype of cross-species infection) were marked as 1, and negative samples (swine viruses without the phenotype of cross-species infection) were marked as 0. The second prediction type was the pseudo probability that a sample belongs to a certain class (positive or negative). For each prediction type, all 64 outputs computed by the 64 RF models were concatenated as a new feature vector. Each swine virus was eventually represented by two 64-dimensional feature vectors, which were marked “class” and “prob,” respectively. Feature vector “class” comprised the class information learnt from the original feature pool, and feature vector “prob” comprised the probabilistic information. Fast speed for computation was expected for the first type models while high performance for prediction accuracy was expected for the second type models.

**2.4.2. Feature Representation Optimization.** The two prediction types were further optimized to improve their feature representation ability. A well-known feature selected method, mRMR, was used to rank the features of the “class” and “prob” information [25]. The mRMR method uses the mutual information to maximize the mutual information between the joint distribution of the selected features and the class labels and minimizes the redundancy between the selected features. The mRMR method was used to optimize the feature representations and obtain the feature list ranked by their importance scores. The sequential forward search (SFS) strategy was used to increase the features from the ranked feature list one by one [16]. After training the RF classifier, the feature subset with the best performance was

considered as the optimal subset. We obtained 20 optimal features for “class” and the 25 optimal features for “prob.”

**2.5. RF Algorithm.** An RF algorithm was used to obtain two types of feature vectors and construct models of prediction for cross-species infection of SIVs. RF machine-learning algorithms are robust and have been used widely to model biology data [4]. The RF behaves like an ensemble algorithm and proposes a set of decision trees by random feature selection. We used the RF algorithm in the R environment in this study [26]. All the experiments were done using version 3.5.0 of R with the default parameters (tree number = 500).

**2.6. Evaluation Metrics.** We used four commonly used metrics to evaluate the model performance, namely, sensitivity (SN), specificity (SP), accuracy (ACC), and Mathew’s correlation coefficient (MCC) as follows:

$$\left\{ \begin{array}{l} \text{SN} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%, \\ \text{SP} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100\%, \\ \text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\%, \\ \text{MCC} = \frac{\text{TP} \times \text{TN} + \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FN})(\text{TN} + \text{FP})}} \times 100\%, \end{array} \right. \quad (6)$$

where TP indicates true positive, the number of correctly predicted true strains with the phenotype of human

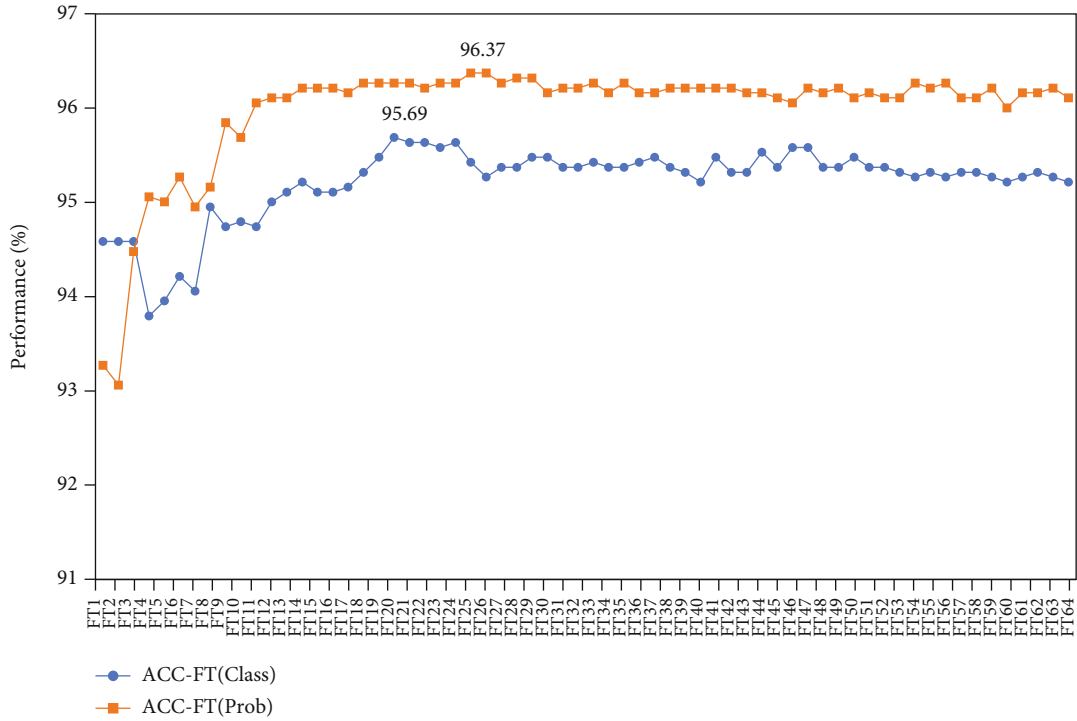
TABLE 1: Summary of feature descriptors and their corresponding feature number.

| Descriptors | Type                      | Number | Descriptors | Type               | Number |
|-------------|---------------------------|--------|-------------|--------------------|--------|
| 1           | AAC                       | 20     | 33          | GGAP ( $g = 13$ )  | 441    |
| 2           | PseAAC ( $\lambda = 1$ )  | 21     | 34          | GGAP ( $g = 14$ )  | 441    |
| 3           | PseAAC ( $\lambda = 2$ )  | 22     | 35          | GGAP ( $g = 15$ )  | 441    |
| 4           | PseAAC ( $\lambda = 3$ )  | 23     | 36          | GGAP ( $g = 16$ )  | 441    |
| 5           | PseAAC ( $\lambda = 4$ )  | 24     | 37          | GGAP ( $g = 17$ )  | 441    |
| 6           | PseAAC ( $\lambda = 5$ )  | 25     | 38          | BIT20 ( $k = 4$ )  | 80     |
| 7           | PseAAC ( $\lambda = 6$ )  | 26     | 39          | BIT20 ( $k = 8$ )  | 160    |
| 8           | PseAAC ( $\lambda = 7$ )  | 27     | 40          | BIT20 ( $k = 12$ ) | 240    |
| 9           | PseAAC ( $\lambda = 8$ )  | 28     | 41          | BIT20 ( $k = 16$ ) | 320    |
| 10          | PseAAC ( $\lambda = 9$ )  | 29     | 42          | BIT20 ( $k = 20$ ) | 400    |
| 11          | PseAAC ( $\lambda = 10$ ) | 30     | 43          | BIT20 ( $k = 24$ ) | 480    |
| 12          | PseAAC ( $\lambda = 11$ ) | 31     | 44          | BIT20 ( $k = 28$ ) | 560    |
| 13          | PseAAC ( $\lambda = 12$ ) | 32     | 45          | BIT20 ( $k = 32$ ) | 640    |
| 14          | PseAAC ( $\lambda = 13$ ) | 33     | 46          | BIT20 ( $k = 36$ ) | 720    |
| 15          | PseAAC ( $\lambda = 14$ ) | 34     | 47          | BIT21 ( $k = 4$ )  | 84     |
| 16          | PseAAC ( $\lambda = 15$ ) | 35     | 48          | BIT21 ( $k = 8$ )  | 168    |
| 17          | PseAAC ( $\lambda = 16$ ) | 36     | 49          | BIT21 ( $k = 12$ ) | 252    |
| 18          | PseAAC ( $\lambda = 17$ ) | 37     | 50          | BIT21 ( $k = 16$ ) | 336    |
| 19          | PseAAC ( $\lambda = 18$ ) | 38     | 51          | BIT21 ( $k = 20$ ) | 420    |
| 20          | GGAP ( $g = 0$ )          | 441    | 52          | BIT21 ( $k = 24$ ) | 504    |
| 21          | GGAP ( $g = 1$ )          | 441    | 53          | BIT21 ( $k = 28$ ) | 588    |
| 22          | GGAP ( $g = 2$ )          | 441    | 54          | BIT21 ( $k = 32$ ) | 672    |
| 23          | GGAP ( $g = 3$ )          | 441    | 55          | BIT21 ( $k = 36$ ) | 756    |
| 24          | GGAP ( $g = 4$ )          | 441    | 56          | OLP ( $k = 4$ )    | 44     |
| 25          | GGAP ( $g = 5$ )          | 441    | 57          | OLP ( $k = 8$ )    | 88     |
| 26          | GGAP ( $g = 6$ )          | 441    | 58          | OLP ( $k = 12$ )   | 132    |
| 27          | GGAP ( $g = 7$ )          | 441    | 59          | OLP ( $k = 16$ )   | 176    |
| 28          | GGAP ( $g = 8$ )          | 441    | 60          | OLP ( $k = 20$ )   | 220    |
| 29          | GGAP ( $g = 9$ )          | 441    | 61          | OLP ( $k = 24$ )   | 264    |
| 30          | GGAP ( $g = 10$ )         | 441    | 62          | OLP ( $k = 28$ )   | 308    |
| 31          | GGAP ( $g = 11$ )         | 441    | 63          | OLP ( $k = 32$ )   | 352    |
| 32          | GGAP ( $g = 12$ )         | 441    | 64          | OLP ( $k = 36$ )   | 396    |

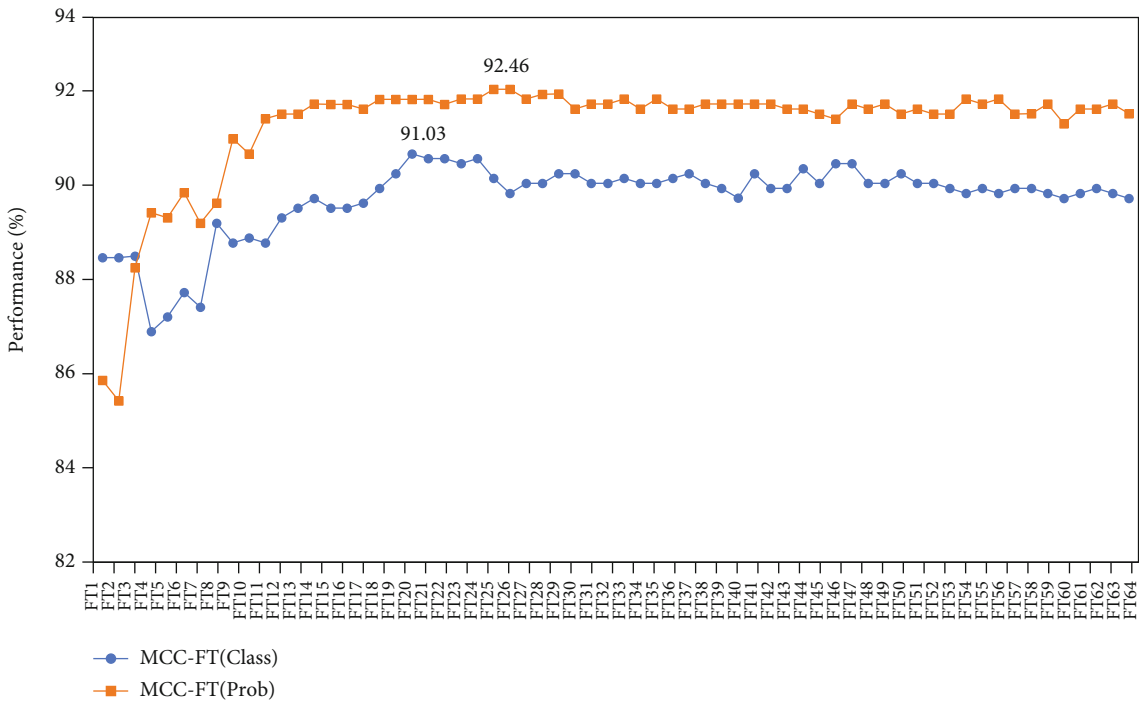
TABLE 2: Amino acid set for predicting SIVs.

| Num | Pro <sup>1</sup> | Pos <sup>2</sup> | Entropy | Num | Pro | Pos | Entropy | Num | Pro    | Pos | Entropy |
|-----|------------------|------------------|---------|-----|-----|-----|---------|-----|--------|-----|---------|
| 1   | HA               | 9                | 1.57    | 13  | HA  | 163 | 1.56    | 25  | HA     | 401 | 1.51    |
| 2   | HA               | 53               | 1.74    | 14  | HA  | 169 | 1.65    | 26  | NA     | 42  | 1.75    |
| 3   | HA               | 78               | 1.56    | 15  | HA  | 173 | 1.62    | 27  | NA     | 43  | 1.78    |
| 4   | HA               | 82               | 1.51    | 16  | HA  | 189 | 2.17    | 28  | NA     | 52  | 1.61    |
| 5   | HA               | 131              | 1.59    | 17  | HA  | 192 | 1.58    | 29  | NA     | 93  | 1.77    |
| 6   | HA               | 135              | 1.67    | 18  | HA  | 193 | 1.63    | 30  | NA     | 332 | 1.65    |
| 7   | HA               | 137              | 1.57    | 19  | HA  | 196 | 1.76    | 31  | NA     | 344 | 1.55    |
| 8   | HA               | 140              | 1.68    | 20  | HA  | 199 | 1.62    | 32  | NA     | 369 | 1.87    |
| 9   | HA               | 142              | 1.90    | 21  | HA  | 219 | 1.65    | 33  | NA     | 385 | 1.74    |
| 10  | Ss               | 144              | 2.15    | 22  | HA  | 261 | 1.76    | 34  | NA     | 400 | 1.72    |
| 11  | HA               | 156              | 1.75    | 23  | HA  | 269 | 1.54    | 35  | NA     | 435 | 1.69    |
| 12  | HA               | 159              | 1.65    | 24  | HA  | 276 | 1.62    | 36  | PB1-F2 | 21  | 1.52    |

<sup>1</sup>Viral protein. <sup>2</sup>Position of amino acid residue as H3 subtype numbering.



(a)



(b)

FIGURE 2: Feature representation optimization with the mRMR algorithm. (a) The SFS curves for ACC of “class” and “prob” feature. The feature number (1-64) and the accuracy were represented by the x- and y-axis. (b) The SFS curves for MCC of “class” and “prob” feature. The feature number (1-64) and the coefficient were represented by the x- and y-axis. The “class” and “prob” features were marked by the blue and yellow color.

infection; TN indicates true negative, the number of correctly predicted true strains without the phenotype of human infection; FP indicates false positive, the number of

strains without the phenotype of human infection predicted to be strains with the phenotype of human infection; and FN is false negative, the number of strains with the phenotype of

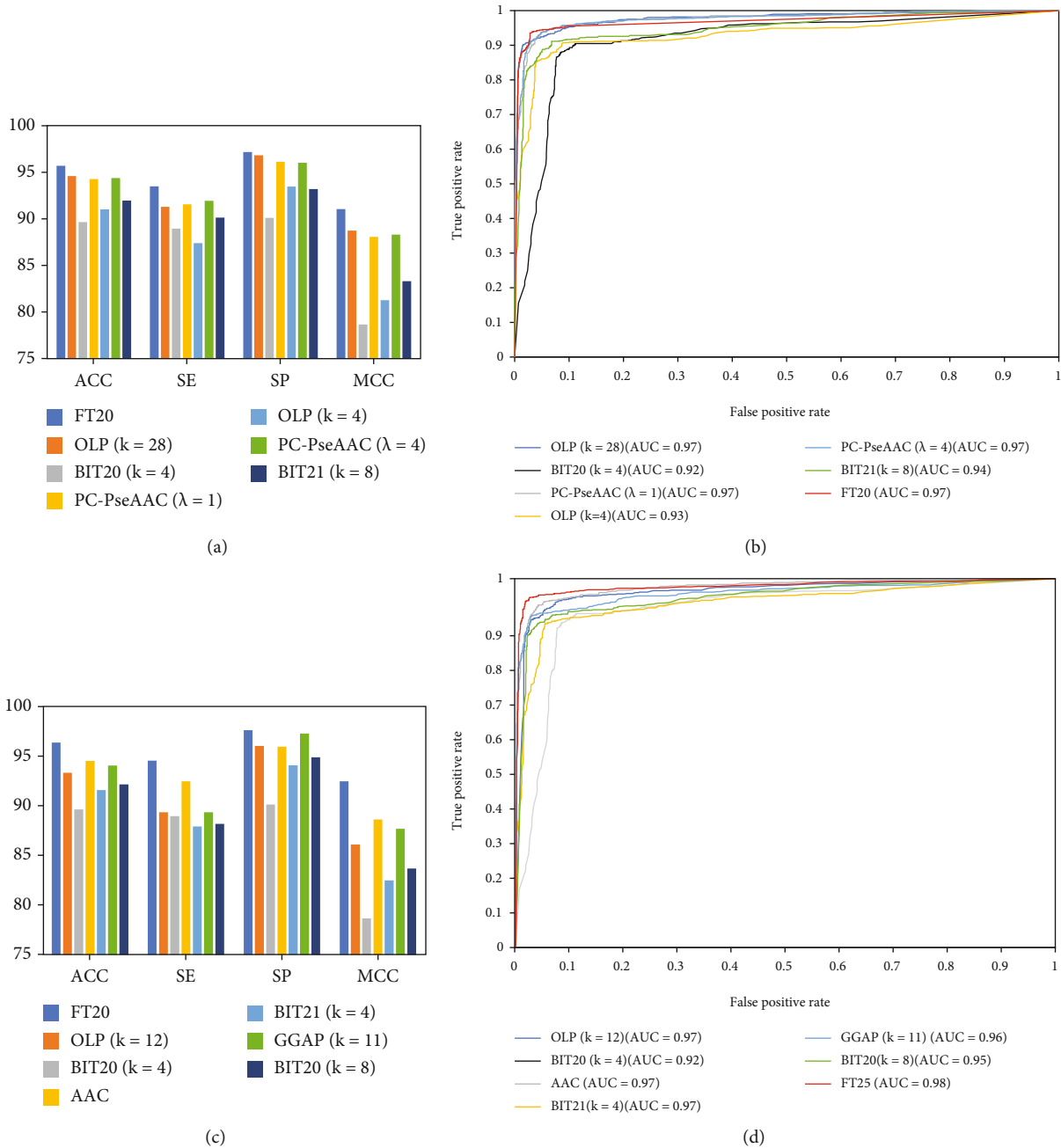


FIGURE 3: Performance of the optimal features: (a) performances of the optimal “class” features and the top 6 individual descriptors; (b) ROC curves of the optimal “class” features and the top 6 individual descriptors; (c) performances of the optimal “prob” features and the top 6 individual descriptors; (d) ROC curves of the optimal “prob” features and the top 6 individual descriptors.

human infection predicted to be strains without the phenotype of human infection. The sensitivity and specificity metrics measure the predictive ability of a model in positive and negative cases, respectively. The other two measures, ACC and MCC, were used to evaluate the overall performance of the models. For these four metrics, high scores indicate high performance of the models.

The receiver operating characteristic (ROC) curve, which is used to evaluate the overall performance of a binary classifier system [27], was also used in this study. The ROC curve is generated by plotting the true positive rate against

the false positive rate under different classification thresholds. We also calculated the area under ROC curve (AUC) to evaluate the predictive performance of the models. AUC values range from 0.5 to 1.

**2.7. Tenfold Cross-Validation Method.** The 10-fold cross-validation method was used to evaluate the predictive performance of the models. The models were trained on 692 positive samples and 1019 negative samples that were selected randomly from the cleaned dataset. The remaining 10% of samples (77 positive and 114 negative) were used

TABLE 3: Results of feature representations using class information and probabilistic information.

| Features                       | ACC   | SE    | SP    | MCC   | TP  | TN   | FP | FN |
|--------------------------------|-------|-------|-------|-------|-----|------|----|----|
| Class features                 | 95.22 | 92.72 | 96.91 | 90.05 | 713 | 1098 | 35 | 56 |
| Probabilistic features         | 95.95 | 93.24 | 97.79 | 91.59 | 717 | 1108 | 25 | 52 |
| Optimal class features         | 95.69 | 93.50 | 97.18 | 91.03 | 719 | 1101 | 32 | 50 |
| Optimal probabilistic features | 96.37 | 94.54 | 97.62 | 92.46 | 727 | 1106 | 27 | 42 |

TABLE 4: Performance of feature representation learning and ensemble learning.

| Learning strategies       | ACC   | SE    | SP    | MCC   | TP  | TN   | FP | FN |
|---------------------------|-------|-------|-------|-------|-----|------|----|----|
| Class information         | 95.69 | 93.50 | 97.18 | 91.03 | 719 | 1101 | 32 | 50 |
| Probabilistic information | 96.37 | 94.54 | 97.62 | 92.46 | 727 | 1106 | 27 | 42 |
| Major voting              | 94.37 | 90.51 | 97.00 | 88.31 | 696 | 1099 | 34 | 73 |
| Probability averaging     | 94.48 | 90.77 | 97.00 | 88.52 | 698 | 1099 | 34 | 71 |

as an independent test dataset to assess the performances of the classifiers. This process was repeated 10 times, and the 10 results are averaged to obtain the final evaluation of prediction performance.

### 3. Results and Discussion

*3.1. Signature Position of SIV.* After elimination of redundancy and other necessary cleaning of viral data from the GISAID database, the final dataset for the prediction of cross-species infection contained two categories of viruses: 769 viruses isolated from human and 1133 viruses isolated from swine. The 769 human viruses were considered as positive samples because they were verified to have the ability of infection among humans. The 1133 swine viruses were considered as negative samples. Information about these virus strains is summarized in Table S1.

To screen the signature position, the entropies in each position of the 11 viral proteins were calculated, respectively. As shown in Table 2, the HA protein contained the highest number of selected amino acid residues (25/36), which is consistent with the known role of HA mainly in receptor-binding and fusion activity for cross-species infection of SIVs. Positions HA102-HA290 are located in or close to the host receptor binding region [28, 29], and HA163 and HA189 are related to the specificity of receptor binding [30, 31]. The signature positions were verified to be related with the mechanism of interspecies transmission or high efficiency of transmission among humans, which would rationalize the model and benefit predicting accuracy.

*3.2. Optimal Feature Representations.* The mRMR feature ranking algorithm was used to select the 64-dimensional feature vector, which comprised the predictions from the 64 RF models (Figure 1). A ranked feature list for the 64 features was generated after sorting their importance scores from the mRMR algorithm. The sequential forward search strategy was proposed to explore the optimal feature representations from the ranked list of 64 features. The features were increased one by one according to the sequence in the list, and the RF classifiers were trained. The influenza virus data

were tested with the 10-fold cross-validation method. The sequential forward search curves for the ACC and MCC metrics were drawn to find the optimized feature (Figure 2). For the class features, the RF classifier performed best with maximum ACC and MCC of 95.69% and 91.03%, respectively, when the feature number 20 was selected (Figures 2(a) and 2(b)). This result indicates that the first 20 features from the ranked feature list had the optimal representation ability to distinguish swine viruses with the ability of cross-species infection. For the probabilistic features, the RF classifier performed best with the first 25 features (ACC of 96.37% and MCC of 92.46%; Figures 2(a) and 2(b)). The screened 20 class and 25 probabilistic features were used to build the predictive classifiers of cross-species infection.

*3.3. Comparison of Optimal Feature Representations with Individual Descriptors.* Using the class feature, optimal representation vectors with 20 dimensions were obtained from 20 individual feature descriptors. The predictive performance of the optimal feature was compared with the six top individual descriptors to evaluate the learning ability of the feature representation. The 10-fold cross-validation tests were fulfilled based on the dataset.

The performances of the optimal class features and the compared individual features are illustrated in Figure 3(a), and the ROC curves were shown in Figure 3(b). The optimal features gave the best predictive performance with maximum ACC and MCC of 95.68% and 91.03%, respectively, which are higher than the values obtained with the second-best feature descriptor BIT20 ( $k=4$ ) (Figure 3(a)). The AUC (0.97) obtained using our feature descriptor was better than that of BIT20 ( $k=4$ ;  $AUC=0.91$ ). Notably, only 20 features were used for optimal feature, whereas BIT20 ( $k=4$ ) used 80 features. Moreover, the 25 optimal feature representations based on probabilistic information were compared with the individual feature descriptors. The performances of the optimal probabilistic features and the compared individual features are shown in Figure 3(c), and the ROC curves are shown in Figure 3(d). The results were consistent and indicated that the probabilistic feature representations



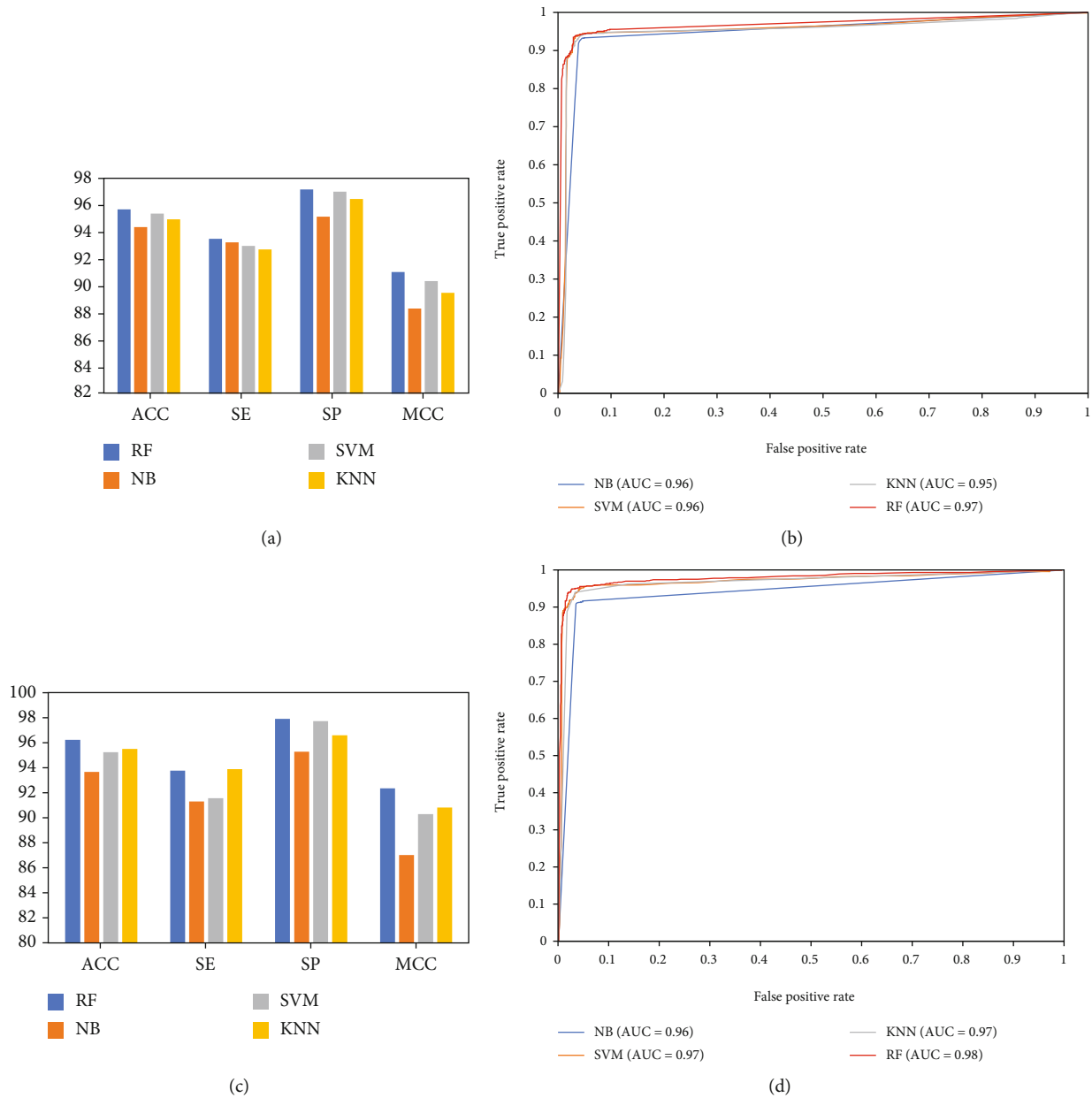


FIGURE 4: Comparison of traditional classifiers: (a) performances of the optimal “class” features with traditional classifiers; (b) ROC curves of the optimal “class” features with traditional classifiers; (c) performances of the optimal “prob” features with traditional classifiers; (d) ROC curves of the optimal “prob” features with traditional classifiers.

outperformed the other six feature descriptors. The optimal features gave the best predictive performance with maximum ACC and MCC of 96.37% and 92.46%, respectively, which are higher than those of the six individual features.

**3.4. Comparison of Class and Probabilistic Information.** Influenza viruses were represented by class and probabilistic information, and their feature vectors comprised the predictions of the 64 RF models to encode signature positions of 36 amino acids. The performance of different information to predict cross-species infection of swine influenza virus was evaluated. As shown in Table 3, the feature vector using probabilistic information outperformed the feature vector

using class information. The overall performance based on probabilistic information had ACC and MCC values of 95.95% and 91.59%, respectively, whereas the overall performance based on class information had ACC and MCC values of 95.22% and 90.05%, respectively. The performances of the two optimal feature vectors also are shown in Table 3. The overall 64-dimensional feature vectors encoded with class and probabilistic information were compared with the optimal features. After ranking by the mRMR algorithm, the performance based on the optimal probabilistic information increased from 95.95% to 96.37% for ACC and from 91.59% to 92.46% for MCC, and the performance based on class information increased from 95.22% to 95.69%

for ACC and from 90.05% to 91.03% for MCC. These results confirmed that the probabilistic feature identified infection better than the class feature. However, both feature types had predictive power for cross-species infection and were used to construct predictive models.

**3.5. Comparison of Feature Representation Learning with Ensemble Learning.** Traditional ensemble learning methods combine predicting results from multiple models to make decisions and for classification. The feature representation learning used and optimized the predictions of the 64 RF models to obtain a predicting model that was similar to traditional ensemble learning models. Two types of feature representation learning (class information learning and probabilistic information) were compared with two classical ensemble learning methods (majority voting and probability averaging). Majority voting considers the majority predictions of the 64 RF models and makes predictions according to the majority rule. Probability averaging simply computes the probabilistic values of the 64 RF models and makes prediction based on the threshold. As shown in Table 4, both types of feature representation learning gave better performances than the two traditional ensemble learning methods. With the probabilistic information, the feature learning strategy had maximum ACC and MCC of 96.37% and 92.46%, respectively. The ACC and MCC obtained with our strategy were about 2% and 3% higher, respectively, than those obtained with the ensemble strategies. Based on class information, the feature learning strategy had the maximum ACC and MCC of 95.69% and 91.03%, respectively. The ACC and MMC obtained with our strategy were about 1% and 3% higher, respectively, than those obtained with the ensemble strategies. Notably, our feature learning strategy achieved a remarkable improvement, even though ensemble learning is considered an effective way to improve predictive performances.

**3.6. Comparison of Our Predictor with Classical Classifiers.** We used the RF algorithm and class or probabilistic information to construct predictor for SIVs. To evaluate the predictive performance of the RF method, we compared our predictor with the traditional classifiers, Support vector machine (SVM), Naïve Bayes (NB), and K-nearest neighbor (KNN), on our dataset with 10-fold cross-validation. The parameters for these classifiers were the same with those in the references [4] and [32]. The results showed that the RF method gave the best overall predictive performance based on the class information with maximum ACC and MCC of 95.69% and 91.03%, which were 1.32% and 2.71% higher, respectively, than those obtained with the NB method (Figures 4(a) and 4(b)). Our AUC (0.97) was better than that of KNN ( $k = 4$ ; AUC = 0.95). We also compared our predictor with the traditional classifiers based on probabilistic information. The probabilistic feature representation outperformed the other three classifiers. The RF method gave the best overall predictive performance based on the probabilistic information with maximum ACC and MCC of 96.37% and 92.46%, which were 2.58% and 5.38% higher, respectively, than those with the NB method (Figures 4(c)

and 4(d)). Our AUC (0.98) was better than that of NB (AUC = 0.96). Overall, the results show that the RF method produced better predictions of infection than the support vector machine, NB, and KNN methods.

## 4. Conclusions

A model for predicting cross-species infection of SIVs was described in the paper. The major contribution of this predictor was the set of informative features of viral proteins that were learned from a total of 64 feature descriptors, including compositional, position-specific, and physico-chemical information. A feature representation learning scheme was proposed. We integrated class and probabilistic information into our feature representations and removed redundant and irrelevant features by feature space optimization to improve the feature representation ability. The ten-fold cross-validation results showed that a high predictive performance was achieved using 20 informative features and 22 probabilistic information. We compared the feature representation learning scheme with those of different learning strategies and confirmed that feature representation learning scheme gave better predictions. We anticipate that our method will be a powerful tool for large-scale identification of swine influenza viruses and will facilitate the characterization of their transmission phenotype and accelerate their applications in virology.

## Data Availability

After the registration for any application (<https://www.gisaid.org/registration/register/>), the public sequences of influenza viruses used in this paper can be downloaded from the GISAID EpiFlu database (<http://platform.gisaid.org/epi3/frontend>) under the database access agreement (<https://platform.epicov.org/epi3/frontend#5aa0ce>) and with the acknowledgment GISAID data contributors (<https://www.gisaid.org/help/publish-with-data-from-gisaid/>). We used the Python programming language to create an easy-to-use tool that implements our predictor and handle massive data, which is freely accessible via <https://github.com/kouzheng/SIVPred-FL>.

## Ethical Approval

No animals/humans were used for studies that are base of this research.

## Conflicts of Interest

The authors declare no conflict of interest, financial, or otherwise.

## Authors' Contributions

Xiaoli Qiang designed the framework of analysis. Zheng Kou, Junjie Li, and Saeed Kosari performed all computational work. Zheng Kou and Xinyue Fan implemented the code. Zheng Kou and Xiaoli Qiang wrote the manuscript. All authors read and approved the final manuscript.

## Acknowledgments

We would like to acknowledge the originating and submitting laboratories of the viral sequences from the GISAID EpiFlu public database. This work was supported by the National Natural Science Foundation of China (61972109, 62172114).

## Supplementary Materials

Information about the final 1902 strains is summarized in Table S1. (*Supplementary materials*)

## References

- [1] R. G. Webster, W. J. Bean, O. T. Gorman, T. M. Chambers, and Y. Kawaoka, "Evolution and ecology of influenza A viruses," *Microbiological Reviews*, vol. 56, no. 1, pp. 152–179, 1992.
- [2] S. Elbe and G. Buckland-Merrett, "Data, disease and diplomacy: GISAID's innovative contribution to global health," *Global Challenges*, vol. 1, no. 1, pp. 33–46, 2017.
- [3] Y. Shu and J. McCauley, "GISAID: global initiative on sharing all influenza data from vision to reality," *Euro Surveillance*, vol. 22, no. 13, p. 30494, 2017.
- [4] X. Qiang and Z. Kou, "Scoring amino acid mutation to predict pandemic risk of avian influenza virus," *BMC Bioinformatics*, vol. 20, no. S8, p. 288, 2019.
- [5] E. C. Claas, A. D. Osterhaus, R. van Beek et al., "Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus," *Lancet*, vol. 351, no. 9101, pp. 472–477, 1998.
- [6] K. Subbarao, A. Klimov, J. Katz et al., "Characterization of an avian influenza A (H5N1) virus isolated from a child with a fatal respiratory illness," *Science*, vol. 279, no. 5349, pp. 393–396, 1998.
- [7] I. H. Brown, "The epidemiology and evolution of influenza viruses in pigs," *Veterinary Microbiology*, vol. 74, no. 1-2, pp. 29–46, 2000.
- [8] A. I. Karasin, M. M. Schutten, L. A. Cooper et al., "Genetic characterization of H3N2 influenza viruses isolated from pigs in North America, 1977-1999: evidence for wholly human and reassortant virus genotypes," *Virus Research*, vol. 68, no. 1, pp. 71–85, 2000.
- [9] C. W. Olsen, "The emergence of novel swine influenza viruses in North America," *Virus Research*, vol. 85, no. 2, pp. 199–210, 2002.
- [10] N. N. Zhou, D. A. Senne, J. S. Landgraf et al., "Genetic reassortment of avian, swine, and human influenza A viruses in American pigs," *Journal of Virology*, vol. 73, no. 10, pp. 8851–8856, 1999.
- [11] F. S. Dawood, S. Jain, and L. Finelli, "Emergence of a novel swine-origin influenza A (H1N1) virus in humans," *The New England Journal of Medicine*, vol. 360, no. 25, pp. 2605–2615, 2009.
- [12] Z. Kou, S. Hu, and T. Li, "Genome evolution of novel influenza A (H1N1) viruses in humans," *Chinese Science Bulletin*, vol. 54, no. 13, pp. 2159–2163, 2009.
- [13] G. J. D. Smith, D. Vijaykrishna, J. Bahl et al., "Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic," *Nature*, vol. 459, no. 7250, pp. 1122–1125, 2009.
- [14] A. S. Bowman, S. W. Nelson, S. L. Page et al., "Swine-to-human transmission of influenza A(H3N2) virus at agricultural fairs, Ohio, USA, 2012," *Emerging Infectious Diseases*, vol. 20, no. 9, pp. 1472–1480, 2012.
- [15] J. A. Pulit-Penalosa, C. Pappas, J. A. Belser et al., "Comparative in vitro and in vivo analysis of H1N1 and H1N2 variant influenza viruses isolated from humans between 2011 and 2016," *Journal of Virology*, vol. 92, no. 22, p. e01444-18, 2018.
- [16] X. Qiang, C. Zhou, X. Ye, P. Du, S. Ran, and L. Wei, "CPPred-FL: a sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning," *Briefings in Bioinformatics*, vol. 21, no. 1, pp. 11–23, 2020.
- [17] H. Han, "Derivative component analysis for mass spectral serum proteomic profiles," *BMC Medical Genomics*, vol. 7, no. S1, p. S5, 2014.
- [18] X. Zeng, L. Liu, L. Lü, and Q. Zou, "Prediction of potential disease-associated microRNAs using structural perturbation method," *Bioinformatics*, vol. 34, no. 14, pp. 2425–2432, 2018.
- [19] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.
- [20] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.
- [21] J. Wang, C. Ma, Z. Kou, Y. Zhou, and H. Liu, "Predicting transmission of avian influenza A viruses from avian to human by using informative physicochemical properties," *International Journal of Data Mining and Bioinformatics*, vol. 7, no. 2, pp. 166–179, 2013.
- [22] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K. Chou, "Pse-in-one: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Nucleic Acids Research*, vol. 43, no. W1, pp. W65–W71, 2015.
- [23] W. R. Atchley, J. Zhao, A. D. Fernandes, and T. Drüke, "Solving the protein sequence metric problem," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 18, pp. 6395–6400, 2005.
- [24] Y. Dou, B. Yao, and C. Zhang, "PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine," *Amino Acids*, vol. 46, no. 6, pp. 1459–1469, 2014.
- [25] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 185–205, 2005.
- [26] A. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
- [27] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "ROCR: visualizing classifier performance in R," *Bioinformatics*, vol. 21, no. 20, pp. 3940–3941, 2005.
- [28] D. J. Hulse, R. G. Webster, R. J. Russell, and D. R. Perez, "Molecular determinants within the surface proteins involved in the pathogenicity of H5N1 influenza viruses in chickens," *Journal of Virology*, vol. 78, no. 18, pp. 9954–9964, 2004.
- [29] J. Stevens, A. L. Corper, C. F. Basler, J. K. Taubenberger, P. Palese, and I. A. Wilson, "Structure of the uncleaved human H1 hemagglutinin from the extinct 1918 influenza virus," *Science*, vol. 303, no. 5665, pp. 1866–1870, 2004.
- [30] L. Glaser, J. Stevens, D. Zamarin et al., "A single amino acid substitution in 1918 influenza virus hemagglutinin changes receptor binding specificity," *Journal of Virology*, vol. 79, no. 17, pp. 11533–11536, 2005.

- [31] E. M. Sorrell, H. Wan, Y. Araya, H. Song, and D. R. Perez, "Minimal molecular constraints for respiratory droplet transmission of an avian-human H9N2 influenza A virus," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 18, pp. 7565–7570, 2009.
- [32] X. Qiang, Z. Kou, G. Fang, and Y. Wang, "Scoring amino acid mutations to predict avian-to-human transmission of avian influenza viruses," *Molecules*, vol. 23, no. 7, p. 1584, 2018.