Hindawi
Computational and Mathematical Methods in Medicine
Volume 2022, Article ID 1515810, 11 pages
https://doi.org/10.1155/2022/1515810

Research Article

Mathematical Modeling and Computational Prediction of High-Risk Types of Human Papillomaviruses

Junchao Zhang and Kechao Wang

Department of Mathematics, Harbin University, Harbin, Heilongjiang 150001, China

Correspondence should be addressed to Junchao Zhang; zhangjc@hrbu.edu.cn

Received 22 April 2022; Accepted 28 June 2022; Published 21 July 2022

Academic Editor: Lei Chen

Cervical cancer is one of the main causes of cancer death all over the world. Most diseases such as cervical epithelial atypical hyperplasia and invasive cervical cancer are closely related to the continuous infection of high-risk types of human papillomavirus. Therefore, the high-risk types of human papillomavirus are the key to the prevention and treatment of cervical cancer. With the accumulation of high-throughput and clinical data, the use of systematic and quantitative methods for mathematical modeling and computational prediction has become more and more important. This paper summarizes the mathematical models and prediction methods of the risk types of human papillomavirus, especially around the key steps such as feature extraction, feature selection, and prediction algorithms. We summarized and discussed the advantages and disadvantages of existing algorithms, which provides a theoretical basis for follow-up research.

1. Introduction

Human papillomavirus (HPV) usually causes benign papillomas and epithelial malignancies [1]. About 30 years ago, researchers found a close relationship between HPV and cervical cancer. Since then, specific HPV type DNA has been found in almost all cervical cancer biopsies [2]. Epidemiological studies also continue to show that HPV is the main cause of cervical cancer. HPV is a papillomavacuolar virus of the family lactomaviridae. It is a spherical DNA virus with 72 shell particles on the surface, a 20 hedral three-dimensional symmetrical nucleocapsid structure, with a diameter of about 45-55 nm [3]. It contains 8000 base pairs, of which 88% are viral proteins. According to the function of HPV genome, it can be divided into three parts: noncoding region, early gene region, and late gene region. The noncoding regions include promoters, enhancers, and silencers, which play an important role in DNA replication and transcriptional regulation. The early gene region contains six genes E1, E2, E4, E5, E6, and E7, including proteins and oncogenes required in the process of replication [4]. The late gene region contains L1 and L2 genes, which encode the structural proteins that constitute the virus capsid [5].

The early region encodes and produces six proteins. Their functions are as follows: E1 protein mainly controls virus replication and replication inhibition and is expressed in the early stage of virus infection; E2 protein is a specific DNA binding protein [6], which can not only regulate mRNA transcription and DNA replication but also inhibit the expression of E6 and E7 proteins; E4 protein is expressed during virus infection and plays an important role in virus replication and mutation; E5 protein is the smallest transforming protein [7]. It consists of two functional regions, one is the amino terminal hydrophobic region, which is related to the insertion position of E5 protein on the membrane, and the other is the hydrophilic region of carboxyl terminal. If the carboxyl terminal is injected into resting cells, it can induce cell DNA synthesis. In addition, it can also induce the expression of a variety of oncogenes; E6 and E7 proteins are the two most important proteins. They can not only regulate the cell cycle [8] but also play an important role in the proliferation of cancer cells.

More than 200 HPV types have been identified, and about 40 types can invade the female reproductive tract [9]. When the identified HPV has significant homology differences with the defined HPV types, some new types will be defined. Epidemiological studies have shown that there is a strong relationship between reproductive human papillomavirus and cervical cancer, which is not related to other risk factors. According to its relative malignancy, HPV can be divided into two or three types: low-risk type, medium-risk type, and high-risk type. The high-risk HPVs include HPV16, 18, 31, 33, 34, 35, 39, and 45, and the low-risk HPVs include HPV6, 11, 42, 43, and 44 [10–13].

The distribution of HPV types has obvious geographical characteristics. Research shows that among the high-risk types worldwide, HPV16 accounts for 51% [14], followed by HPV16, accounting for 16%, others, such as HPV45 accounts for 9%, HPV31 accounts for 6%, and HPV33 accounts for 3%. The sum of these four common HPV types exceeds 80%. However, in Latin America, HPV33 is the most common, followed by HPV39 and 59. In Asia, in addition to the most common HPV16 and 18 [15], HPV 52 and 58 account for far more cervical cancer than Western and African countries.

Different types of HPV have different clinical manifestations. According to the tissues invaded, HPV can be divided into high (low) risk type of skin and high (low) risk type of mucosa [16]. Low-risk skin types such as HPV1, 2, 3, and 4 are generally related to common warts, flat warts, and other diseases. High-risk skin types such as HPV5, 8, 14, and 17 are associated with anal cancer, prostate cancer, bladder cancer, and so on. For low-risk types of mucosa: HPV6, 11, 13, 32, 34, 40, 42, etc., they are generally related to diseases such as genital, anal, oropharyngeal, and esophageal mucosal infection. Mucosal high-risk types such as HPV16, 18, 30, 31, 33, 35, and 39 can lead to cervical cancer, rectal cancer, tonsillar cancer, and other diseases [17].

Now, there are many epidemiological and experimental methods to identify HPV types [18–21]. They mainly use polymerase chain reaction (PCR) technology to realize the rapid detection of clinical samples. With the rapid accumulation of human papillomavirus data, a reliable and effective calculation method is needed to predict high-risk HPV. In recent years, many computational models have been proposed to predict HPV high-risk types. Eom et al. designed genetic algorithm to predict HPV type through the sequence fragment distribution characteristics of HPV [22]. Joung et al. designed the prediction method of HPV type based on support vector machine prediction and hidden Markov model [23, 24]. Through systematic analysis, Park et al. suggested using decision tree to construct the typing model of human papillomavirus [25]. Kim and Zhang calculated the distance distribution of amino acid pairs and further predicted the risk type of HPV through E6 protein [26, 27]. Kim et al. extracted the differential features of protein secondary structure and designed a set of support vector machine (GSVM) to classify HPV types [28]. Esmaeili et al. calculated Chou's pseudoamino acid composition of HPV sequence and classified HPV types using ROC [29]. Alemi et al. systematically compared the physical and chemical properties of high-risk and low-risk HPV and build a prediction model of high-risk HPV based on the support vector machine [30]. Wang et al. used protein "sequence space" to explore this information to predict high-risk types of HPVs [31]. Xu et al. proposed a HPV high-risk prediction model based on reduced amino acids, and they divided 20 amino acids into several nonoverlapping groups and calculated the structure and chemical patterns of each group as the basis for distinguishing high-risk HPV [32].

In this paper, we introduce the development of HPV typing in cervical cancer, mainly focusing on sequence characteristics and predicted secondary structure characteristics. The feature fusion method and feature selection algorithm are discussed. Finally, we review the multiple classification and various machine learning algorithms in HPV typing of cervical cancer.

2. Benchmark Datasets

HPV database is from Los Alamos National Laboratory (LANL), with a total of 72 HPV types. HPV risk types were manually determined based on the HPV profile [33]. If HPV belongs to skin related or skin group, HPV is classified as a low-risk type. On the other hand, it is classified as a high-risk type if HPV is known to be a high-risk type of cervical cancer [28]. Some researchers may build their own HPV dataset from SWISS-PROT. For these sequences, some researchers will complete their own wonderful processing. Here, we focus on using mathematical models to detect the risk types of human papillomavirus. Therefore, the detailed processing of data sets is no longer our important discussion content.

3. Methods

3.1. Feature Extraction. The extraction of HPV sequence information is transforming from complex structure to extraction of mathematical features [34]. Directly extracting information from the complex structure of HPV protein will cause large errors, and the extracted information may not be comprehensive. However, it is very easy to directly extract the mathematical features from the protein sequence, there is no influence of error on the connection, and the extracted information contains all the protein information.

3.1.1. Sequence Features

(1) K-Mer. Protein sequences and peptides can be regarded as a set of signs [35], and we can analyze their characteristics through the frequency of k-mer to obtain the difference between the two HPV sequences. If k is 1, it means that it is the composition of amino acids, which is calculated as follows [36]:

$$F(S) = (f_1, f_2, \cdots, f_{20})^T, \tag{1}$$

$$f_i = \frac{n_i}{\sum_{i=1}^{20} n_i}, \tag{2}$$

where n_i means the number of the amino acid i in the protein sequence.

(2) Order-Based Features. Order-based features reflect the physical and chemical interaction among the amino acids pairs, which can be described by sequence coupling score and quasisequence score [37].

Suppose S is a protein sequence with a length of L, τ_j is denoted as the sequence interaction factor describing the sequence influence:

$$\tau_j = \frac{1}{L-j}\sum_{i=1}^{L-j} J_{i,i+j}(j = 1, 2, \cdots, k, k < L), \tag{3}$$

where $J_{i,i+j}$ is the physical and chemical distance between the amino acid Ri and R_{i+j}, which can be calculated:

$$J_{i,i+j} = D^2\left(R_i, R_{i+j}\right). \tag{4}$$

The calculation method of D^2 can consider the hydrophobicity value, polarity value, and side chain volume of amino acids. Thus, S can be defined by the following features:

$$V_{QSOE}(S) = \left[\nu_1, \nu_2, \cdots, \nu_{20}, \nu_{20+1}, \cdots, \nu_{20+k}\right]^T. \tag{5}$$

The first 20 features are k-mer features, and the latter k -dimensional features represent the interaction information of amino acid sequences.

(3) Amino Acid Distribution. Amino acid distribution is composed of amino acid composition, transformation, and distribution. This component can be regarded as the amino acid component of HPV, which can be original sequence or reduced sequence to describe the sequence component [38]. The transformation can be described by calculating the number of conversions between a base and a subsequent base. It can be calculated by the following formula:

$$Trs = \frac{\text{Count}_{rs} + \text{Count}_{sr}}{N-1}, \tag{6}$$

where N is the number of amino acids, and Count_{sr} is the number of conversions between the base s and a subsequent base r.

(4) Pseudoamino Acid Composition (PseAAC). In order to avoid the loss of many important information hidden in HPV sequences, Chou proposed pseudoamino acid composition (PseAAC), which is not a simple AAC to represent protein samples [39]. Pseudoamino acid composition is a vector with size $R + \lambda$, where R is the composition size of the amino acids, and the latter λ dimension is the information of pseudoamino acid composition [29]. They can be calculated by the following formula:

$$\nu_u = \begin{cases} \dfrac{f_u}{\sum_{u=1}^{R} f_u + w\sum_{k=1}^{\lambda}\tau_k} (u \le R), \\ \dfrac{w\tau_{u-20}}{\sum_{u=1}^{R} f_u + w\sum_{k=1}^{\lambda}\tau_k} (R \le u \le R + \lambda), \end{cases} \tag{7}$$

where f_u is the frequency of RedAA, w is the weight, and τ_k can be calculated by the following formula:

$$\tau_k = \frac{1}{L-k}\sum_{i=1}^{L-k} J_{i,j+k}(k < L), \tag{8}$$

$$J_{i,j+k} = \frac{1}{3}\left\{\left[H_1(R_i) - H_1(R_{i+k})\right]^2 + \left[H_2(R_i) - H_2(R_{i+k})\right]^2 + \left[M(R_i) - M(R_{i+k})\right]^2\right\}. \tag{9}$$

$H_i(R_i)$ represents molecular weight of amino acid residue R_i, and $M(R_i)$ is the molecular weight of amino acid residue R_i. Then, the protein sequence can be mapped to the following feature vectors by this way:

$$V_{\text{PseAAC}}(S) = \left[\nu_1, \nu_2, \cdots, \nu_{20}, \cdots, \nu_{20+\lambda}\right]^T. \tag{10}$$

3.1.2. Evolutionary Profile

(1) Position Specific Scoring Matrix (PSSM). PSSM is developed on the basis of position frequency matrix [40], which can quantitatively describe the evolutionary characteristics of HPV sequences. At present, it is the most commonly used matrix feature model, especially for protein structural class prediction [41]. When a protein s with length L is given, its PSSM is defined as

$$\text{PSSM}_s = \begin{bmatrix} P_{1\longrightarrow 1} & P_{1\longrightarrow 2} & \cdots & P_{1\longrightarrow j} & \cdots & P_{1\longrightarrow 20} \\ P_{2\longrightarrow 1} & P_{2\longrightarrow 2} & \cdots & P_{2\longrightarrow j} & \cdots & P_{2\longrightarrow 20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{i\longrightarrow 1} & P_{i\longrightarrow 2} & \cdots & P_{i\longrightarrow j} & \cdots & P_{i\longrightarrow 20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{L\longrightarrow 1} & P_{L\longrightarrow 2} & \cdots & P_{L\longrightarrow j} & \cdots & P_{L\longrightarrow 20} \end{bmatrix}, \tag{11}$$

where $i \longrightarrow j$ represents that the residue at the position of i $-$ th mutated into the type of j during the course of biological evolution. $P_{i\longrightarrow j}$ indicates the score of mutation, and the ordered 20 amino acids are marked as 1, 2, 3… 20. The close evolutionary relationship among these HPV sequences can be described by their PSSM [42]. We search homogeneous sequences from SWISS-PROT database with help of PSI-BLAST [43–45] and normalize PSSM using the following sigmoid function:

$$f(x) = \frac{1}{1 + e^{-x}}. \tag{12}$$

(2) Structural Properties of PSSM. When $[0, L-1] \times \sum \longrightarrow \Re$ is used to describe a function M, L is the length of a function M. The PSSM of the sequence Seq_0 will be transformed into

a RedPSSM of its reduced sequence Seq_s,

$$[0, L] \times \sum_{Seq_0} \longrightarrow [0, L] \times \sum_{Seq_r}, \tag{13}$$

where $\sum_{Seq_0}$ and $\sum_{Seq_r}$ are finite alphabet sets of the original and reduced sequences. $[\text{RedPSSM}]_{ij}$ is defined as

$$[\text{RedPSSM}]_{ij} = \sum_{j=1}^{g_s(j)} \frac{P_{i \longrightarrow j}}{g_s(j)}, \tag{14}$$

where $g_s(j)$ represents the size of the reduced group which is made up with the reduced amino acid j, and j belongs to $1 \le j \le |\sum_{seq_r}|$.

Auto covariance (AC) is a correlation factor, which has been widely used in various fields of bioinformatics [46]. It can connect adjacent residues according to protein sequence and convert PSSM into equal length vector [47]. The AC variable can represent the average interaction between residues with a series of lags, which is defined as

$$AC_{j,g}(S) = \frac{1}{L-g} \sum_{i=1}^{L-g} \left(P_{i,j} - P_j\right)\left(P_{i,j+g} - \mathrm{P}_j\right), \tag{15}$$

where $AC_{j,g}(S)$ can indicate the interaction between two spacer g residues. But it can just calculate the residue interaction in the same column [48]. In order to solve this problem, we calculated total correlations among all the columns by the following formula:

$$\begin{aligned} \mathrm{RAC}_g(h, j) &= \frac{1}{L-g} \sum_{i=1}^{L-g} \left| [\text{RedPSSM}]_{i,h} - \frac{[\text{RedPSSM}]_{i,h} + [\text{RedPSSM}]_{i+g,j}}{2} \right| \\ &\quad \times \left| [\text{RedPSSM}]_{i+g,h} - \frac{[\text{RedPSSM}]_{i,h} + [\text{RedPSSM}]_{i+g,j}}{2} \right| \\ &= \frac{1}{4(L-g)} \sum_{i=1}^{L-g} \left([\text{RedPSSM}]_{i,h} - [\text{RedPSSM}]_{i+g,h} \right)^2. \end{aligned} \tag{16}$$

3.1.3. Secondary Structure Features. In order to improve the accuracy of HPV high-risk types, some researchers analyzed the protein secondary structures of HPV, extracted its characteristics, and constructed prediction models [28, 49]. Given a protein sequence, we predict its secondary structure sequence. Its secondary structure content $(\text{content}_{\text{SE}})$ can be calculated as

$$\text{content}_{\text{SE}} = \frac{\text{Count}_{\text{SE}}}{\sum_{x \in \{C,H,E\}} \text{Count}_x}, \tag{17}$$

where SE belong to $\{C, H, E\}$. H is α-helix, E is β-strand, and C is coil. Gk and Herand calculated the first and second order composition moment vector (CMV) as the structure features [50], which is defined as the following formula:

$$\text{CMV}_{\text{SE}}^k = \frac{\sum_{j=1}^{\text{Count}_{\text{SE}}} \text{PO}_{\text{SE}j}^k}{\prod_{d=1}^{k} (N-d)}, \tag{18}$$

where $\text{PO}_{\text{SE}j}$ represents the element at position j in the secondary structure sequence with length N, which k is the vector order.

In order to further study the arrangement of different structural elements, some important structural fragments or patterns have been proposed one after another [51]. The length of the longest segment $(\text{MaxSeg}_{\text{SE}})$ is defined as the following formula:

$$\text{MaxSeg}_{\text{SE}} = \text{MaxLen}(\text{SEG} : \text{SEG}_{\text{SE}}), \tag{19}$$

where MaxLen represents the function of the maximal segment length, and SEG_{SE} is the composed of each segment of the structure element SE [52]. In order to reduce the length effect, a normalized length of the longest segment $(\text{NMaxSeg}_{\text{SE}})$ is defined

$$N\text{MaxSeg}_{\text{SE}} = \frac{\text{MaxLen}(\text{SEG} : \text{SEG}_{\text{SE}})}{N}, \tag{20}$$

where N is the sequence length. In addition to the maximal segment length, the average length of the segment $(\text{AvgSeg}_{\text{SE}})$ is also an important feature of protein secondary structure, which is defined as

$$\text{AvgSeg}_{\text{SE}} = \frac{\sum \text{Len}(\text{SEG} : \text{SEG}_{\text{SE}})}{\text{Content}_{\text{SEG}_{\text{SE}}}}, \tag{21}$$

where $\text{Content}_{\text{SEG}_{\text{SE}}}$ represents the substance of SEG_{SE}, and Len is a function of segment length. The normalized average length of the segment $(\text{NAvgSeg}_{\text{SE}})$ is

$$N\text{AvgSeg}_{\text{SE}} = \frac{\sum \text{Len}(\text{SEG} : \text{SEG}_{\text{SE}})}{\text{Content}_{\text{SEG}_{\text{SE}}} \times N}, \tag{22}$$

where N represents the length of protein sequence.

Zhang et al. analyzed the secondary structure of protein, only considered the helical and folded structural units, ignoring the irregular curl [53]. According to this simplification rule, they got a simplified protein secondary structure sequence and defined the transition probability matrix (TPM) as follows:

$$\text{TPM} = \begin{pmatrix} p_{\alpha\alpha} & p_{\alpha\beta} \\ p_{\beta\alpha} & p_{\beta\beta} \end{pmatrix}, \tag{23}$$

where

$$p_{a_i a_j} = \begin{cases} \dfrac{\text{Content}_{a_i a_j}}{\sum_{t=1}^{2} \text{Content}_{a_i a_t}} & \sum_{t=1}^{2} \text{Content}_{a_i a_t} \neq 0, \\ 0 & \sum_{t=1}^{2} \text{Content}_{a_i a_t} = 0, \end{cases} \tag{24}$$

where a_i belongs to $\{\alpha, \beta\}$, a_i is followed by the a_j, and the size of the event is described by $\text{Content}_{a_i a_j}$.

Most researchers pay attention to the content of structural units, so sometimes they do not know the location information of these structural units in the protein structure. Dai et al. proposed some location-based structural features [54]. If the distance $\text{Dis}(\delta)$ of the given structural element δ is 1, they will be divided into a structural domain. If not, they will be divided into two different structural domains. If the $\text{Dis}(\delta)$ is given, its probability distribution can be achieved. The definitions of digital feature semimean $\text{Semi} - E_{(k)}(\delta)$ and semivariance $\text{Semi} - D_{(k)}(\delta)$ are as follows:

$$\text{Semi} - E_{(k)}(\delta) = \sum_{Dis(\delta)=1}^{k} \text{Dis}(\delta) \times P(\text{Dis}(\delta)), \tag{25}$$

$$\text{Semi} - D_{(k)}(\delta) = \sum_{\text{Dis}(\delta)=1}^{k} (\text{Dis}(\delta))^2 \times P(\text{Dis}(\delta)) - \left[\sum_{\text{Dis}(\delta)=1}^{k} \text{Dis}(\delta) \times P(\text{Dis}(\delta))\right]^2. \tag{26}$$

The ratio of the standard $\text{Semi} - D_{(5)}$ to $\text{Semi} - E_{(5)}$ can be calculated by:

$$\text{PSSF}(\delta) = \frac{\text{Semi} - E_{(5)}(\delta)}{\sqrt{\text{Semi} - D_{(5)}(\delta)}}. \tag{27}$$

$\text{PSSF}(\delta)$ is used to describe the position distribution of predicted secondary structural elements.

3.2. Feature Selection. Feature selection (SF) is one of the important steps in data processing, which plays an important role in data mining, pattern recognition, and machine learning [32]. It solves the problem of how to select the input features corresponding to the optimal prediction results. Through feature selection, the complexity of the problem can be reduced, and the prediction accuracy, robustness, and interpretability of the learning algorithm can be improved [55]. Therefore, feature selection can be used to select the characteristics of HPV risk types [56]. More importantly, it helps to have a deeper understanding of HPV sequences when we analyze their characteristics. This section summarizes some feature selection methods.

3.2.1. Mutual Information. The feature selection method based on mutual information can capture the nonlinear relationship between variables, which is more suitable for dealing with complex classification problems [57]. Mutual information can be expressed as characteristic matrix X, and the relevant category is C. Mutual information can be defined as

$$\text{Re l}(X_i) = I(X_i; C) = \sum_{X_i, C} P(X_i, C) \log \frac{P(X_i, C)}{P(X_i)P(C)}. \tag{28}$$

When the set of S is selected, the redundant formula Eq. (29) is

$$\text{Red}(X_i|S) = \frac{1}{S} \sum_{X_j \in S} I(X_i; X_j). \tag{29}$$

As can be seen from the above definition, mutual information aims to select the features that are most relevant to the target category and have the least redundancy between the selected features [58]. Therefore, feature selection can be realized directly according to the value of mutual information.

3.2.2. Support Vector Machine Recursive Feature Elimination (SVM-RFE). Support vector machine has been widely used in the field of pattern recognition and is very suitable for small samples with high-dimensional data. SVM-RFE is a sequential reverse selection (SBS) algorithm based on the maximum interval principle of SVM [59]. SVM-RFE can be divided into linear SVM-RFE and nonlinear SVM-RFE.

(1) Linear SVM-RFE. There is a training sample set $\{xi, yi\}$, $yi \in \{-1, 1\}$, $i = 1, 2, \cdots, n$. The linear SVM can be calculated by the following formula:

$$f(x) = a \cdot x + b, \tag{30}$$

where a is the weight vector of the optimal hyperplane, and b is the threshold.

By introducing Lagrange's formula, the optimization problem of SVM can be transformed into the following dual programming problem:

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i x_j, \tag{31}$$

where α_i can be calculated by solving the maximum value of L_D under the range of $\alpha_i \geq 0$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$. So a Eq. (32) can be defined as follows:

$$a = \sum_{i=1}^{n} \alpha_i y_i x_i. \tag{32}$$

The ranking criterion score of the k-th feature is defined as the following equation:

$$J(k) = w_k^2. \tag{33}$$

During the training process, the feature with the smallest score of the ranking criterion is removed, and the remaining features are used to train the SVM for the next iteration.

(2) Nonlinear SVM-RFE. When the sample size is larger than the number of features, nonlinear SVM-RFE can obtain better results [60]. Features can be mapped to new spaces in higher dimensions by nonlinear SVM-RFE:

$$x \in R^d \mapsto \varphi(x) \in R^h. \tag{34}$$

After being mapped to the new space, the samples can be

divided into linear features. It can be calculated by the formula of Lagrangian:

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \varphi(x_i) \varphi(x_j). \tag{35}$$

Here, $\Phi(x_i)\Phi(x_j)$ can be changed into a Gaussian kernel formula $K(x_i, x_j)$:

$$K(xi, xj) = \ell^{-\lambda \left|\left|x_i - x_j\right|\right|^2}. \tag{36}$$

The ranking criteria of feature k can be described by the following formula:

$$J(k) = \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K\left(x_i^{(-k)}, x_j^{(-k)}\right), \tag{37}$$

where $x_i^{(-k)}$ represents that k has been dropped.

3.2.3. Genetic Algorithm. Genetic algorithm is an adaptive search strategy, and its principle is similar to the survival mechanism of the fittest in nature [61]. It has strong adaptability, strong independence of domain knowledge, and can carry out a large number of parallel computing. Therefore, it is more suitable for processing large-scale complex data, especially for solving multiobjective optimization problems [62]. Given an HPV data set, we construct its characteristic matrix X and use genetic algorithm to obtain an eigenvector, which is an optimal feature set [63].

3.2.4. Kurtosis and Skewness. Kurtosis and skewness are two characteristics describing distribution [64], which can distinguish the distribution of different characteristics. Therefore, many studies calculated the skewness and kurtosis of each feature and selected certain features according to the value of skewness and kurtosis. Given the distribution of a feature $\{x_1, x_2, \cdots, x_n\}$, its kurtosis and skewness are defined as follows:

$$\text{Kurtosis} = \frac{\sum_{i=1}^{n} (x_i - x)^4}{(n-1)SD^4} - 3, \tag{38}$$

$$\text{Skewness} = \frac{\sum_{i=1}^{n} (x_i - x)^3}{(n-1)SD^3}. \tag{39}$$

3.2.5. ReliefF Algorithm. ReliefF is an independent evaluation method, which evaluates each feature separately and assigns weight to the feature [65, 66]. For each time, it randomly selects a sample R from the training set and then calculates the latest instance from the same class R and different sample sets, and then the weight update rules in each step are as follows:

$$W(A) = W(A) - \sum_{j=1}^{k} \frac{\text{diff}(A, R, H_j)}{mk} + \sum_{C \notin \text{class}(R)} \frac{p(C)/1 - p(\text{Class}(R)) \sum_{j=1}^{k} \text{diff}(A, R, M_j(C))}{mk}, \tag{40}$$

where $\text{diff}(A, R_1, R_2)$ represents the difference between the sample R_1 and the sample R_2 on the feature A, and $M_j(C)$ indicates the nearest neighbor of the position of j – th in class C Eq. (41). $\text{diff}(A, R_1, R_2)$ can be calculated by the following formula:

$$\text{diff}(A, R1, R2) = \begin{cases} \dfrac{|R_1[A] - R_2[A]|}{\max(A) - \min(A)}, \\ 0 R_1[A] = R_2[A], \\ 1 R_1[A] \neq R_2[A]. \end{cases} \tag{41}$$

3.3. Prediction Algorithms

3.3.1. Support Vector Machines (SVM). SVM is a discriminant classifier, which is defined by the classification hyperplane [27]. In other words, the labeled training samples are used to train the model, and then the test sample classification is realized by outputting the best hyperplane [67]. When the prediction problem is nonlinear, the objective function of SVM can be defined as

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b, \tag{42}$$

where $K(x_i, x)$ is a kernel function, x_i is a support vector, and α_i belongs to $[0, C]$. The Gaussian radial basis kernel function with strong learning ability and small error is often selected as the kernel function, which is defined as

$$K(x_i, x) = \exp\left(-\frac{\|x_i - x\|^2}{2\sigma^2}\right), \tag{43}$$

where σ is the coefficient of the kernel function and has high flexibility.

3.3.2. Principal Component Analysis (PCA). PCA is a statistical analysis method that converts multiple indicators into a few comprehensive indicators. The idea of PCA is to replace the original more with fewer comprehensive variables [68]. The first step of PCA is to normalize the matrix which can also be expressed as the following mathematical formula:

$$Zij = \frac{x_{ij} - x_j}{s_j} (i = 1, 2, \cdots, n; j = 1, 2, \cdots, p), \tag{44}$$

$$x_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}, s_j^2 = \frac{\sum_{i=1}^{n} (x_{ij} - x_j)^2}{n-1}. \tag{45}$$

The second step of PCA is to obtain the correlation of

coefficient matrix for Z:

$$R = \left[r_{ij}\right]_p xp = \frac{Z^T Z}{n-1}, \tag{46}$$

$$rij = \frac{\sum zkj \cdot zkj}{n-1} (i, j = 1, 2, \cdots, p). \tag{47}$$

3.3.3. K-Nearest Neighbor Algorithm (KNN). KNN is an important nonparametric classification method, which classifies the samples according to the categories of most k -nearest neighbor samples in the feature space [69]. However, the basic k-nearest neighbor classification algorithm needs global search, which has high computational complexity and slow computing speed. In order to reduce the influence of the same feature function in the traditional KNN algorithm, different weights can be assigned to the features in the distance formula to measure the similarity. For example, in Euclidean distance formula, different weights are assigned to different features, as shown in the formula:

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{48}$$

3.3.4. Partial Least Square Discriminant Analysis (PLS-DA). PLS-DA is a standard high-dimensional data analysis method, and it is especially suitable for situations where there are a large number of explanatory variables [70], multicollinearity samples, few observations, and large interference noises. PLS-DA first treats the sample category with dummy variables which can be calculated by the following formula:

$$\begin{cases} Yk = 1, Y = k \\ Yk = 0, Y \neq k \end{cases}, (k = 1, 2, \cdots, q), \tag{49}$$

where Y represents categorical variables, and q is a dummy variable. The relationship model between explanatory variables, response variables is established by least square regression, and then, the category of each sample is determined by comparing the predicted values of model response variables. If the predicted value of the dummy variable component is the largest, it is determined that the sample belongs to the category corresponding to the dummy variable.

3.3.5. Classification and Regression Tree (CART). CART is a nonparametric statistical process of data analysis. Its characteristic is to make full use of the binary tree structure in the calculation process, that is, the root node contains all samples, and the root node is divided into two children under certain partition rules [71]. The process of node is repeated on the subbook point and becomes a regression process until it can no longer be divided into leaf nodes [72]. When all nodes can be classified into class $C\,(k = 1, 2, \cdots, C)$, the Gini impurity of node A can be expressed by the following formula:

$$\text{Gini}(A) = 1 - \sum_{k=1}^{C} p_k^2, \tag{50}$$

where p^k is the proportion of the sample which belongs to class k. If class A can be divided into B and C. The probability of B in the sample A is p and C is q. And the size of impurities will be expressed by the following formula:

$$\text{Gini}(A) - p \cdot \text{Gini}(B) - q \cdot \text{Gini}(C). \tag{51}$$

3.3.6. Linear Discriminant Analysis (LDA). LDA is a statistical analysis method used to determine the type of sample [73], which has been widely used in the prediction of protein structural classes. By finding the feature vector w, the k sets of m metadata are mapped into another lower-dimensional direction. Then classify with the sample in the new space.

$$w_{opt} = \arg \max_{w} \frac{\left|w^T S_B w\right|}{\left|w^T S_w w\right|}. \tag{52}$$

S_B is an interclass deviation matrix:

$$S_B = \sum \left(\text{m}_x - m_y\right)\left(m_x - m_y\right)^T. \tag{53}$$

S_w is an intraclass deviation matrix:

$$S_w = \sum_{i=1}^{N_x} (x_i - m_x)(x_i - m_x)^T + \sum_{i=1}^{N_y} \left(y_i - m_y\right)\left(y_i - m_y\right)^T, \tag{54}$$

where $\{x_i, i = 1, 2, \cdots, N_x\}$ represents class p, $\{y_i, i = 1, 2, \cdots, N_y\}$ represents class n. Where m_x and m_y, respectively, represent the average of class p and class n.

3.3.7. Extreme Gradient Boosting (XGBoost). The gradient boosting algorithm framework supports a variety of different loss functions [74]. In addition to the exponential loss function, it also includes the mean square error loss function and logarithmic loss function. XGBoost algorithm is an enhanced version of gradient boosting algorithm. It is more efficient, flexible, and portable [75]. It used the training data x_i to predict a target variable y_i.

First of all, the objective function can be calculated as

$$obj = \Sigma_{1=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{i=1}^{t} \Omega(f_i), \tag{55}$$

where n represents the number of trees, l is the training loss function, and Ω is the regularization term.

Then, the XGBoost takes the Taylor expansion of the loss function up to the second order and removes all the constants, so the specific objective at step t can be describes as

$$L^{(t)} = \Sigma_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t), \tag{56}$$

where g_i and h_i can be described as follows:

$$\begin{cases} g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right), \\ h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right). \end{cases} \tag{57}$$

The value of the objective function only depends on g_i and h_i.

It can optimize every loss function, including logistic regression and pairwise ranking, and it is simple to parallel and can greatly enhance the program efficiency with a fast model exploration [76].

3.4. Evaluation Measure. After realizing HPV prediction, we need to use statistical test method to evaluate the efficiency of prediction model. Leave-one-out cross-validation (LOO-CV) is a more common method of Bayesian model and a special case of k-fold cross validation [77], because when k is equal to sample size n, and it can be regarded as n-fold cross validation.

HPV typing is a two or three classification problem. If the HPV model predicts a positive result (P) and the true result is also positive, it is called true positive (TP). If the predicted result is negative (N), the real result is also positive, which is called false positive (FP). The results can be calculated by the following formula:

$$\text{accuarcy(ACC)} = \frac{\text{TP} + \text{TN}}{\text{TN} + \text{FP} + \text{FN} + \text{TP}}, \tag{58}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{59}$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{60}$$

4. Conclusion

High-risk HPV accounts for a higher proportion of cervical cancer in the world. Therefore, the identification of high-risk HPV is of great significance for the treatment of some major cancers such as cervical cancer [78]. At present, there are common epidemiological detection techniques for HPV typing [79], such as nucleic acid imprinted in situ hybridization or dot blot hybridization, and the detection infection rate is about 10-20% [80]. In addition, there are test methods, such as polymerase chain reaction (PCR). All HPV types can be detected by PCR, and the detection infection rate is more than 40%. In addition to experimental methods, some computational methods have been proposed to predict HPV typing.

We summarize the mathematical models and prediction methods of the risk types of human papillomavirus, especially around the key steps such as feature extraction, feature selection, and prediction algorithms. From the above research papers, it can be found that the prediction accuracy of low-risk types is often higher than that of high-risk types. E7 performed better than other HPV proteins in low-risk experiments. However, E6 performs best among all HPV proteins for high-risk prediction and all types of prediction experiments, which is just consistent with experimental studies. E5, E6, and E7 proteins of high-risk HPV play an important role in disease progression and cancer [81]. Therefore, E6 protein sequence is more suitable for HPV high-risk prediction, and E7 protein sequence is more suitable for HPV low-risk prediction. However, there are some exceptions. In the reduced amino acid modes, L1 protein performs better in predicting high-risk HPV, while L2 protein is more suitable for low-risk HPV. These conclusions can provide a theoretical basis for subsequent research and guide the establishment of HPV typing models.

Mutations usually lead to cell dysfunction, cell death, and even cancer in higher organisms. Therefore, mutations in HPV protein may make the virus more easily induced or carcinogenic and increase the chance of reinfecting the host or fleeing the host immune system. For example, the carcinogenicity of HPV is mainly controlled by E6 and E7 proteins, which often produce internal variation. The mutation frequency of E6 in cancer is 20% ~90%, and that of E7 is 60% ~90% [82]. A study in Japan showed that the mutation of aspartate to glutamate (d25e) at position 25 of HPV16 E6 protein was related to the DRB1 $*$ 1502 allele of HLA II. This mutation is considered to be an important mutation in invasive cancer and cervical intraepithelial neoplasia. In the Netherlands, the frequency of HLV DRB1 $*$ 07 allele increased in cancer patients with l83v mutation ($P = 0.08$). If the information of mutation information is considered in HPV classification model, it is also an effective way to improve HPV typing detection.

Data Availability

HPV database can be downloaded from Los Alamos National Laboratory (LANL, https://lanl.gov).

Conflicts of Interest

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

Acknowledgments

The authors thank all the anonymous referees for their valuable suggestions and support. This work is supported by the Natural Science Foundation of Heilongjiang Province of China (LH2019F046).

References

[1] Y. R. Tzenov, P. G. Andrews, K. Voisey et al., "Human papilloma virus (HPV) E7-mediated attenuation of retinoblastoma (Rb) induces hPygopus2 expression via Elf-1 in cervical cancer," *Molecular Cancer Research*, vol. 11, no. 1, pp. 19–30, 2013.

[2] Y. Yao, H. Xu, M. Li, Z. Qi, and B. Liao, "Recent advances on prediction of human papillomaviruses risk types," *Current Drug Metabolism*, vol. 20, no. 3, pp. 236–243, 2019.

[3] J. Haedicke and T. Iftner, "Human papillomaviruses and cancer," *Radiotherapy and Oncology*, vol. 108, no. 3, pp. 397–402, 2013.

[4] Y. J. Choi, E. Y. Ki, C. Zhang et al., "Analysis of sequence variation and risk association of human papillomavirus 52

variants circulating in Korea," *PLoS One*, vol. 11, no. 12, article e0168178, 2016.

[5] E. M. Burd, "Human papillomavirus laboratory testing: the changing paradigm," *Clinical Microbiology Reviews*, vol. 29, no. 2, pp. 291–319, 2016.

[6] M. M. Pan, C. Gao, X. L. Li et al., "The enhancement of DNA binding ability of a mutated E2 (A338V) protein of HPV-2," *Chinese Journal of Virology*, vol. 26, no. 3, pp. 223–227, 2010.

[7] B. H. Horwitz, D. L. Weinstat, and D. Dimaio, "Transforming activity of a 16-amino-acid segment of the bovine papillomavirus E5 protein linked to random sequences of hydrophobic amino acids," *Journal of Virology*, vol. 63, no. 11, pp. 4515–4519, 1989.

[8] M. Tommasino and L. Crawford, "Human papillomavirus E6 and E7: proteins which deregulate the cell cycle," *Bioessays News & Reviews in Molecular Cellular & Developmental Biology*, vol. 17, no. 6, pp. 509–518, 1995.

[9] I. Manini and E. Montomoli, "Epidemiology and prevention of human papillomavirus," *Annali di Igiene : Medicina Preventiva e di Comunita*, vol. 30, 4 Supple 1, pp. 28–32, 2018.

[10] M. R. Pillai, S. Lakshmi, S. Sreekala et al., "High-risk human papillomavirus infection and E6 protein expression in lesions of the uterine cervix," *Pathobiology*, vol. 66, no. 5, pp. 240–246, 1998.

[11] M. L. Tornesello, M. L. Duraturo, G. Botti et al., "Prevalence of alpha-papillomavirus genotypes in cervical squamous intraepithelial lesions and invasive cervical carcinoma in the Italian population," *Journal of Medical Virology*, vol. 78, no. 12, pp. 1663–1672, 2006.

[12] M. Arbyn, M. Tommasino, C. Depuydt, and J. Dillner, "Are 20 human papillomavirus types causing cervical cancer?," *Journal of Pathology*, vol. 234, no. 4, pp. 431–435, 2014.

[13] V. Cogliano, R. Baan, K. Straif, Y. Grosse, B. Secretan, and F. E. Ghissassi, "Carcinogenicity of human papillomaviruses," *Lancet Oncology*, vol. 6, no. 4, pp. 204–204, 2005.

[14] D. A. Wick and J. R. Webb, "A novel, broad spectrum therapeutic HPV vaccine targeting the E7 proteins of HPV16, 18, 31, 45 and 52 that elicits potent E7-specific CD8T cell immunity and regression of large, established, E7-expressing TC-1 tumors," *Vaccine*, vol. 29, no. 44, pp. 7857–7866, 2011.

[15] G. M. Clifford, T. Stephen, and F. Silvia, "Carcinogenicity of human papillomavirus (HPV) types in HIV-positive women: a meta-analysis from HPV infection to cervical cancer," *Clinical Infectious Diseases*, vol. 64, no. 9, pp. 1228–1235, 2017.

[16] T. M. Castro, I. Bussoloti Filho, V. X. Nascimento, and S. D. Xavier, "HPV detection in the oral and genital mucosa of women with positive histopathological exam for genital HPV, by means of the PCR," *Revista Brasileira de Oto-Rino-Laringologia*, vol. 75, pp. 167–171, 2009.

[17] M. J. Lace, J. R. Anson, A. J. Klingelhutz et al., "Human papillomavirus (HPV) type 18 induces extended growth in primary human cervical, tonsillar, or foreskin keratinocytes more effectively than other high-risk mucosal HPVs," *Journal of Virology*, vol. 83, no. 22, pp. 11784–11794, 2009.

[18] H. Furumoto and M. Irahara, "Human papilloma virus (HPV) and cervical cancer," *Journal of Medical Investigation*, vol. 49, no. 3-4, pp. 124–133, 2002.

[19] F. X. Bosch, M. M. Manos, N. Muñoz et al., "Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. International biological study on cervical cancer (IBSCC) study group," *Journal of the National Cancer Institute*, vol. 87, no. 11, pp. 796–802, 1995.

[20] R. D. Burk, G. Y. Ho, L. Beardsley, M. Lempa, M. Peters, and R. Bierman, "Sexual behavior and partner characteristics are the predominant risk factors for genital human papillomavirus infection in young women," *The Journal of Infectious Diseases*, vol. 174, no. 4, pp. 679–689, 1996.

[21] N. Muñoz, F. X. Bosch, S. de Sanjosé et al., "Epidemiologic classification of human papillomavirus types associated with cervical cancer," *The New England Journal of Medicine*, vol. 348, no. 6, pp. 518–527, 2003.

[22] J. H. Eom, S. B. Park, and B. T. Zhang, "Genetic mining of DNA sequence structures for effective classification of the risk types of human papillomavirus (HPV)," *Lecture Notes in Computer Science*, vol. 3316, pp. 1334–1343, 2004.

[23] J. G. Joung, S. June, and B. T. Zhang, "Prediction of the risk types of human papillomaviruses by support vector machines," in *Pacific Rim International Conference on Artificial Intelligence*, pp. 723–731, Springer, Berlin, Heidelberg, 2004.

[24] J. G. Joung, S. June, and B. T. Zhang, "Protein sequence-based risk classification for human papillomaviruses," *Computers in Biology and Medicine*, vol. 36, no. 6, pp. 656–667, 2006.

[25] S. B. Park, S. Hwang, and B. T. Zhang, "Mining the risk types of human papillomavirus (HPV) by AdaCost," in *International Conference on Database and expert Systems Applications*, pp. 403–412, Prague, Czech Republic, September 2003.

[26] K. Sun and J. H. Eom, "Prediction of the human papillomavirus risk types using gap-spectrum kernels," in *International Conference on Advances in Neural Networks*, pp. 710–715, Springer, Berlin, Heidelberg, 2006.

[27] S. Kim and B. T. Zhang, "Human papillomavirus risk type classification from protein sequences using support vector machines," in *Workshops on Applications of Evolutionary Computation*, pp. 57–66, Springer, Berlin, Heidelberg, 2006.

[28] K. Sun, J. Kim, and B. T. Zhang, "Ensembled support vector machines for human papillomavirus risk type prediction from protein secondary structures," *Computers in Biology and Medicine*, vol. 39, no. 2, pp. 187–193, 2009.

[29] M. Esmaeili, H. Mohabatkar, and S. Mohsenzadeh, "Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses," *Journal of Theoretical Biology*, vol. 263, no. 2, pp. 203–209, 2010.

[30] M. Alemi, H. Mohabatkar, and M. Behbahani, "In silico comparison of low- and high-risk human papillomavirus proteins," *Applied Biochemistry and Biotechnology*, vol. 172, no. 1, pp. 188–195, 2014.

[31] C. Wang, Y. Hai, X. Liu et al., "Prediction of high-risk types of human papillomaviruses using statistical model of protein sequence space," *Computational and Mathematical Methods in Medicine*, vol. 2015, Article ID 756345, 9 pages, 2015.

[32] X. Xu, R. Kong, X. Liu, P. He, and Q. Dai, "Prediction of high-risk types of human papillomaviruses using reduced amino acid modes," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 5325304, 10 pages, 2020.

[33] U. C. Megwalu, M. M. Chen, Y. Ma, and V. Divi, "Surrogate for oropharyngeal cancer HPV status in cancer database studies," *Head & Neck*, vol. 39, no. 12, pp. 2494–2500, 2017.

[34] R. P. Bonidia, L. D. H. Sampaio, D. S. Domingues et al., "Feature extraction approaches for biological sequences: a

comparative study of mathematical features," *Briefings in Bioinformatics*, vol. 22, no. 5, 2021.

[35] A. Fiannaca, M. L. Rosa, R. Rizzo, and U. Alfonso, "A _k_ -mer-based barcode DNA classification methodology based on spectral representation and a neural gas network," *Artificial Intelligence in Medicine*, vol. 64, no. 3, pp. 173–184, 2015.

[36] D. Qi, W. Li, and L. Li, "Improving protein structural class prediction using novel combined sequence information and predicted secondary structural features," *Journal of Computational Chemistry*, vol. 32, no. 16, pp. 3393–3398, 2011.

[37] X. Nan, Q. Xu, and D. Cao, "Generating various numerical representation schemes of protein sequence," 2015.

[38] B. Yang, H. S. Yang, J. Li, Z. X. Li, and Y. M. Jiang, "Amino acid composition, molecular weight distribution and antioxidant stability of shrimp processing byproduct hydrolysate," *American Journal of Food Technology*, vol. 6, no. 10, pp. 904–913, 2011.

[39] K. C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins-structure Function & Bioinformatics*, vol. 43, no. 3, pp. 246–255, 2001.

[40] T. K. Attwood, "Profile (Weight Matrix, Position Weight Matrix, Position-Specific Scoring Matrix, PSSM)," in *Dictionary of Bioinformatics and Computational Biology*, American Cancer Society, 2004.

[41] R. Kong, X. Xu, X. Liu, P. He, M. Q. Zhang, and Q. Dai, "2SigFinder: the combined use of small-scale and large-scale statistical testing for genomic island detection from a single genome," *BMC Bioinformatics*, vol. 21, no. 1, p. 159, 2020.

[42] L. Mariño-Ramírez, K. Tharakaraman, O. Bodenreider, J. Spouge, and D. Landsman, "Identification of cis-regulatory elements in gene co-expression networks using A-GLAM," *Computational Systems Biology*, vol. 541, pp. 3–22, 2009.

[43] V. S. Gowri, K. G. Tina, O. Krishnadev, and N. Srinivasan, "Strategies for the effective identification of remotely related sequences in multiple PSSM search approach," *Proteins-structure Function & Bioinformatics*, vol. 67, no. 4, pp. 789–794, 2010.

[44] Z. H. Chen, Z. H. You, L. P. Li, Y. B. Wang, L. Wong, and H. C. Yi, "Prediction of self-interacting proteins from protein sequence information based on random projection model and fast Fourier transform," *International Journal of Molecular Sciences*, vol. 20, no. 4, article 930, 2019.

[45] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices[1]," *Journal of Molecular Biology*, vol. 292, no. 2, pp. 195–202, 1999.

[46] J. Wu, Y. Z. Li, M. L. Li, and L. Z. Yu, "Two multi-classification strategies used on SVM to predict protein structural classes by using auto covariance," *Interdisciplinary Sciences Computational Life Sciences*, vol. 1, no. 4, pp. 315–319, 2009.

[47] T. Liu, X. Geng, X. Zheng, R. Li, and J. Wang, "Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles," *Amino Acids*, vol. 42, no. 6, pp. 2243–2249, 2012.

[48] S. Yang, Y. Wang, Y. Chen, and Q. Dai, "MASQC: next generation sequencing assists third generation sequencing for quality control in N6-methyladenine DNA identification," *Frontiers in Genetics*, vol. 11, article 269, 2020.

[49] N. Choudhary, B. S. Biswal, and A. Mohapatra, "Prediction of HPV risk types from protein secondary structure," in *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, Jabalpur, India, December 2015.

[50] M. Gk and D. Herand, "Prediction of bacterial virulent proteins with composition moment vector feature encoding method," *MATEC Web of Conferences*, vol. 49, article 07001, 2016.

[51] M. Onesime, Z. Yang, and Q. Dai, "Genomic island prediction via chi-square test and random forest algorithm," *Computational and Mathematical Methods in Medicine*, vol. 2021, 9 pages, 2021.

[52] Y. Wang, W. Cai, L. Gu, X. Ji, and Q. Shen, "Comprehensive analysis of pertinent genes and pathways in atrial fibrillation," *Computational and Mathematical Methods in Medicine*, vol. 2021, 20 pages, 2021.

[53] S. Zhang, S. Ding, and T. Wang, "High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure," *Biochimie*, vol. 93, no. 4, pp. 710–714, 2011.

[54] Q. Dai, Y. Li, X. Liu, Y. Yao, Y. Cao, and P. He, "Comparison study on statistical features of predicted secondary structures for protein structural class prediction: from content to position," *BMC Bioinformatics*, vol. 14, no. 1, pp. 1–14, 2013.

[55] J. Tao, X. Liu, S. Yang, C. Bao, P. He, and Q. Dai, "An efficient genomic signature ranking method for genomic island prediction from a single genome," *Journal of Theoretical Biology*, vol. 467, pp. 142–149, 2019.

[56] J. V. Kuzhali, G. Rajendran, V. Srinivasan, and G. S. Kumar, "Feature selection algorithm using fuzzy rough sets for predicting cervical cancer Risks," *Modern Applied Science*, vol. 4, no. 8, pp. 134–143, 2010.

[57] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Journal of Machine Learning Research*, vol. 5, no. 4941, pp. 1531–1555, 2004.

[58] H. Lim and D. W. Kim, "Using mutual information for selecting features in multi-label classification," *Journal of KISS : Software and Applications*, vol. 39, no. 10, pp. 806–811, 2012.

[59] Y. Ying, "SVM-RFE algorithm for gene feature selection," *Computer Engineering*, 2005.

[60] Y. Mao, X. Zhou, Z. Yin, D. Pi, Y. Sun, and S. T. Wong, "Gene selection using Gaussian kernel support vector machine based recursive feature elimination with adaptive kernel width strategy," in *International Conference on Rough Sets and Knowledge Technology*, pp. 799–806, Springer, Berlin, Heidelberg, 2006.

[61] B. Xi, J. Tao, X. Liu, X. Xu, P. He, and Q. Dai, "RaaMLab: a MATLAB toolbox that generates amino acid groups and reduced amino acid modes," *Biosystems*, vol. 180, pp. 38–45, 2019.

[62] Z. Yang, W. Yi, J. Tao et al., "HPVMD-C: a disease-based mutation database of human papillomavirus in China," *Database: The Journal of Biological Databases and Curation*, vol. 2022, 2022.

[63] E. Topaka, "Application of genetic algorithms and other feature selection techniques in clinical decision support for cervical cancer diagnosis," 2016.

[64] F. J. Rubio, *Modelling of Kurtosis and Skewness : Bayesian Inference and Distribution Theory*, University of Warwick, England, 2013.

[65] R. Durgabai, "Feature selection using ReliefF algorithm," *IJARCCE*, vol. 10, no. 30, pp. 8215–8218, 2014.

[66] H. E. Tao, H. U. Jie, X. I. A. Peng, and G. U. Chaochen, "Feature selection of Emg signal based on ReliefF algorithm and genetic algorithm," *Journal of Shanghai Jiaotong University*, vol. 50, no. 2, article 204, 2016.

[67] I. I. Suni, Y. Huang, and S. Schuckers, *Bioelectronic Tongue for Food Allergy Detection*, US, 2010.

[68] I. M. A. Melo, M. R. P. Viana, B. Pupin, T. T. Bhattacharjee, and R. de Azevedo Canevari, "PCR-RFLP and FTIR-based detection of high-risk human papilloma virus for cervical cancer screening and prevention," *Biochemistry and Biophysics Reports*, vol. 26, article 100993, 2021.

[69] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: a review," *Journal of Data Analysis and Information Processing*, vol. 8, no. 4, pp. 341–357, 2020.

[70] L. L. Chuen, L. Choong-Yeun, and J. A. Aziz, "Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps," *Analyst*, vol. 143, no. 15, pp. 3526–3539, 2018.

[71] Q. Dai, C. Bao, Y. Hai et al., "MTGIpick allows robust identification of genomic islands from a single genome," *Briefings in Bioinformatics*, vol. 19, no. 3, pp. 361–373, 2016.

[72] C. J. Yang, Y. C. Tsai, and J. J. Tien, "Patients with minor diseases who access high-tier medical care facilities: new evidence from classification and regression trees," *The International Journal of Health Planning and Management*, vol. 34, no. 2, 2019.

[73] H. Jones, C. Gardner, E. Hull, K. Nixon, M. Robinson, and Rio Grande Medical Technologies Inc, "Within-sample variance classification of samples," 2003.

[74] N. A. Bokulich, M. Dillon, E. Bolyen, B. Kaehler, G. Huttley, and J. Caporaso, "q2-sample-classifier: machine-learning tools for microbiome classification and regression," *The Journal of Open Source Software*, vol. 3, no. 30, 2018.

[75] X. Wang, Y. Wang, Z. Xu, Y. Xiong, and D. Q. Wei, "Prediction of the classes of anatomical therapeutic chemicals using a network-based label space partition method," *Frontiers in Pharmacology*, vol. 10, article 971, 2019.

[76] T. Chen and C. Guestrin, "A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, United States, August 2016.

[77] T. Sivula, M. Magnusson, and A. Vehtari, "Uncertainty in Bayesian leave-one-out cross-validation based model comparison," 2020, http://arxiv.org/abs/2008.10296.

[78] J. Wu, T. Zhou, J. Tao et al., "Similarity/dissimilarity analysis of protein structures based on Markov random fields," *Computational Biology & Chemistry*, vol. 75, pp. 45–53, 2018.

[79] M. A. Cohenford and B. Lentrichia, "Detection and typing of human papillomavirus using PNA probes," 2007.

[80] M. Kalantari, E. Blennow, B. Hagmar, and B. Johansson, "Physical state of HPV16 and chromosomal mapping of the integrated form in cervical carcinomas," *Diagnostic Molecular Pathology*, vol. 10, no. 1, pp. 46–54, 2001.

[81] L. C. Barrow, "e7 proteins of high-risk (type 16) and low-risk (type 6) human papillomaviruses regulate p130 differently," 2010.

[82] D. M. Da Silva, G. L. Eiben, S. C. Fausch et al., "Cervical cancer vaccines: emerging concepts and developments," *Journal of Cellular Physiology*, vol. 186, no. 2, pp. 169–182, 2001.