Methods and Applications

Development of a Subsequence Correlation Coefficient Feature Vector Method for High-Resolution HIV-1 Subtype Classification — China, 2004–2022

Shuyan Han[1,&]; Lily He[1,&]; Yihang Tang[2]; Kun Peter Li[3]; Yuhua Ruan[4]; Hengjian Cui[5,#]

Summary

What is already known about this topic?

Current HIV-1 subtype classification tools often rely on time-consuming alignment, whereas new non-alignment methods typically target single genes. China lacks a model for specifically predicting the non-B subtype strains prevalent domestically.

What is added by this report?

We developed a fast alignment-free method (SCCFV-RDA) for building multigene models. It achieved over 99.85% accuracy in classifying an international dataset and 99.7% accuracy in classifying Chinese *pol* gene data and showed superior recall for key circulating recombinant form subtypes.

What are the implications for public health practice?

This tool provides accurate and efficient computational support for the precise molecular surveillance of HIV-1 in China, thereby facilitating the formulation of targeted prevention and control strategies.

ABSTRACT

Introduction: The subtype classification of human immunodeficiency virus type 1 (HIV-1) is vital for its prevention and control. Current methods often rely on time-consuming sequence alignments, whereas new alignment-free approaches typically focus on single genes. The prevalent HIV strains in China are mainly non-B subtypes. However, no subtype prediction model exists for local sequences. Therefore, we aimed to develop a fast and accurate method for building multigene models specifically tailored to predict HIV-1 subtypes on the basis of Chinese data.

Methods: Herein, we propose a novel sequence-feature extraction method, named Subsequence Correlation Coefficient Feature Vector (SCCFV), which captures the spatial distribution and correlations of nucleotides and converts DNA sequences into high-dimensional numerical features. It then combines these features with a regularized discriminant analysis (RDA) classifier to build a dedicated classification model.

Results: The proposed SCCFV-RDA model exhibited robust and generalizable capabilities, maintaining more than 99.85% accuracy across an international dataset covering 10 gene regions. It also achieved 99.7% classification accuracy on an independent test set of Chinese HIV-1 *pol* gene data, showing significantly higher recall for mainstream circulating recombinant form subtypes than that achieved by traditional tools.

Conclusion: We developed a classification tool for HIV-1 subtypes and built a specialized model for HIV-1 strains prevalent in China. The accuracy and efficiency of the tool surpassed those of existing traditional methods, providing reliable computational support for the precise molecular epidemiological surveillance of HIV in China. This method holds significant practical value for facilitating the formulation of targeted prevention and control strategies.

According to recent statistics, approximately 40.8 million people worldwide were living with the human immunodeficiency virus (HIV) in 2025 (*1*). Increasingly complex and diverse HIV subtypes have emerged from the large-scale and evolving pandemic (*2*), posing numerous prevention and treatment challenges. As China is one of the key HIV-affected countries, the classification of its indigenous viral subtypes warrants attention. Through surveillance efforts from 2004 to 2022, the Chinese Center for Disease Control and Prevention (CDC) has established one of the largest national molecular epidemiology databases for acquired immunodeficiency syndrome (AIDS), currently providing data from 57,902 HIV-infected individuals (*3*). Domestically prevalent non-B

subtypes, such as CRF01_AE and CRF07_BC, exhibit significant genetic differences from the B subtype dominant in Europe and America (*4*). However, because the current mainstream classification tools are primarily trained on datasets dominated by European and American B subtypes and known circulating recombinant forms (CRFs), their performance in classifying the strains circulating in China may be limited, in particular for new CRFs that are not represented in their training data. Consequently, novel computational models that are specifically tailored to the genetic sequences prevalent in China and capable of efficiently processing large-scale data are urgently needed.

Machine learning and deep learning methods have been extensively applied to bioinformatics classification, demonstrating significant potential in this task (*5*). Existing models are predominantly trained on European and American B subtype data and typically target specific *pol* gene or full-genome regions, lacking a unified framework adaptable to diverse fragments. Herein, we propose a novel modular framework for HIV-1 subtype classification. Based on the Subsequence Correlation Coefficient Feature Vector (SCCFV) method, this framework establishes a model system adaptable to multiple gene fragments. To assess the applicability of the model beyond the *pol* gene region, we additionally collected 821 full-genome sequences from Chinese public databases for independent model validation. The model demonstrated outstanding performance in classifying complex CRFs in China, providing a unified and efficient solution for global HIV-1 genotyping and evolutionary surveillance based on arbitrary fragments.

METHODS

Dataset

The HIV-1 genome, which spans approximately 9,800 bp, comprises the highly conserved structural genes *gag*, *pol*, and *env* along with multiple regulatory and accessory genes. In this study, the sequence data for each gene segment were downloaded from the HIV Sequence Database, totaling 10 datasets. Categories with sample sizes exceeding 9 were selected to establish the models.

Of the 57,902 HIV *pol* gene sequences in the Chinese CDC AIDS database, 5% are labeled as "unique recombinant forms (URFs)," "Other," and "Other CRFs." Given their limited epidemiological significance, only sequences of the major HIV subtypes circulating in China were selected (Supplementary Figure S1, available at https://weekly.chinacdc.cn/).

Subsequence Correlation Coefficient Feature Vector Method

To analyze the base position distribution of the nucleotides in DNA sequences, we used a feature mapping method to transform the sequences into vector representations within a multidimensional feature space. For a given HIV DNA sequence, $G=(g_1,g_2,\cdots,g_n)$, where each nucleotide variable is $g_i = \{G, T, A, C\}$ and the sequence length is n. For a specific base, where $g \in \{A, T, C, G\}$, each base position is converted into a numerical feature:

$$v_g(i) = \begin{cases} 1, g_i = g \\ 0, g_i \neq g \end{cases} \quad i = 1,2,\ldots n. \tag{1}$$

Digital representation of the sequences is achieved by constructing four base-specific time series: $V_g = [V_g(1), V_g(2), \ldots, V_g(n)]$. For example, the sequence "AGCTAAG" can be converted as follows:

$$\begin{aligned} &V_A = [1, 0, 0, 0, 1, 1, 0] \\ &(\text{A appears in the 1st, 5th, and 6th positions.}) \end{aligned} \tag{2}$$

V_G, V_C, and V_T are similarly derived.

Two key statistical measures are calculated from the obtained numerical data: the average frequency and correlation coefficient. Let N_g denote the total count of nucleotide g in the sequence, where $g \in \{A, T, C, G\}$. The average frequency f_g represents the global proportion of base g, reflecting its compositional characteristics, and is defined as follows (Figure 1):

$$f_g = \frac{N_g}{n} \tag{3}$$

For a more in-depth analysis of the spatial distribution patterns of the bases, autocorrelation and cross-correlation functions were introduced. The autocorrelation function measures the strength of the association between the same bases after an interval of M positions:

$$\varnothing_{AA}(M) = \frac{1}{n}\sum_{i=1}^{n}(v_A(i) - f_A)(v_{A+M}(i) - f_A) \tag{4}$$

The autocorrelation coefficient $R_{AA}(M)$ of base A is defined as follows:

$$R_{AA}(M) = \frac{\varnothing_{AA}(M)}{\varnothing_{AA}(0)} \tag{5}$$

The M-step-delayed normalized cross-correlation coefficient between nucleotides A and T serves as a crucial metric for quantifying the strength of the

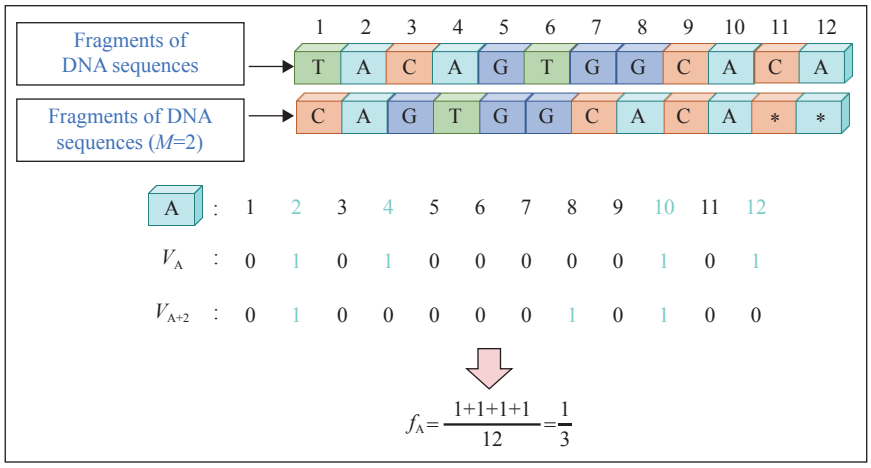


FIGURE 1. Computational methodology for determining V_A, V_{A+2}, and f_A.
Abbreviation: DNA=deoxyribonucleic acid.

correlation between two nucleotides at specific intervals within a sequence. The cross-correlation function is defined as follows:

$$\varnothing_{AT}(M) = \frac{1}{n}\sum_{i=1}^{n}(v_A(i) - f_A)(v_{T+M}(i) - f_T) \quad (6)$$

The correlation coefficient $R_{AT}(M)$ is defined as follows:

$$R_{AT}(M) = \frac{\varnothing_{AT}(M)}{\sqrt{\varnothing_{AA}(0) \cdot \varnothing_{TT}(0)}} \quad (7)$$

This coefficient reveals the periodicity of the base occurrence patterns; a high $R_{AA}(M)$ value indicates a tendency for the A bases to repeat at every M position. Cross-correlation quantifies the spatial dependencies between different bases, with $R_{AT}(M)$ reflecting the cooperative or antagonistic effects between A and T at M position intervals.

To comprehensively capture local feature variations across different regions of the DNA sequence, a segmented processing strategy is used. Specifically, each raw DNA sequence $G = (g_1, g_2, \cdots, g_n)$ is first uniformly divided into J contiguous subsequences. The first r subsequences (Substr$_1$, Substr$_2$, · · ·, Substr$_r$) each contains Y+1 nucleotides, whereas the remaining J-r subsequences (Substr$_{r+1}$, Substr$_{r+2}$, · · ·, Substr$J$) each contain Y nucleotides. The formula is as follows:

$$Y = \left[\frac{n}{J}\right], r = n - J \times Y (0 \le r \le J) \quad (8)$$

The standardized cross-correlation coefficient R for all possible nucleotide combinations (AA, AC, AG, AT, ..., TT; 16 combinations in total) is independently calculated for each subsequent segment at different M delay steps, resulting in a 16 × M-dimensional vector (6). The resulting eigenvalues are concatenated in segment order to form a high-dimensional feature vector with a total dimension of J × 16 × M.

Parameter Tuning and Classifier Selection

Four classifiers from machine learning methods (7) — Random Forest, XGBoost, Regularized Discriminant Analysis (RDA), and LightGBM — were selected for comparative analysis. To comprehensively validate the model performance, the following systematic evaluation framework was used:

$$\text{Accuracy} = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|} \quad (9)$$

$$\text{Precision} = \frac{|TP|}{|TP| + |FP|} \quad (10)$$

$$\text{Recall} = \frac{|TP|}{|TP| + |FN|} \quad (11)$$

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

As an example, we randomly divided 55,261 *pol* gene data points from the Chinese CDC database into 75% and 25% for use as training and independent test sets, respectively. Five-fold cross-validation was performed within the training set. This process simultaneously optimizes the hyperparameters of the classifier and the core parameters of the SCCFV feature extraction method. The classification performance of the final model was objectively evaluated using the independent test set, which was not involved in either the training or parameter selection processes.

According to the accuracy heatmaps of the four classifiers (Supplementary Figure S2, available at

https://weekly.chinacdc.cn/), RDA achieved the highest accuracy (99.7%) under the parameters J = 3 and M = 5 and was therefore selected as the final model with this parameter combination.

Figure 2 illustrates the overall framework of the SCCFV method. First, the input sequence is converted into a four-dimensional vector representation via a numerical mapping layer to capture base position information. Subsequently, the sequence is uniformly partitioned into J subsegments, and both local and global features are extracted via a statistical feature computation layer to construct a multidimensional feature vector. After input to the RDA classifier optimized via 5-fold cross-validation, the optimal model is obtained by parameter tuning, ultimately achieving precise HIV-1 subtype classification.

Comparison with Existing Tools

The performance of the SCCFV-RDA method was compared with those of COMET (*9*), REGA (*10*), and HIVdb (*11*), which are sequence alignment-based tools that are widely used as authoritative standards in the field. SNV (*12*), an alignment-free method based on single-nucleotide variation features, was also included for comparison. REGA was not included in the comparative analysis of the Chinese *pol* dataset because of its prohibitively long computation time.

Implementation Details

All the computational experiments were performed on a Lenovo Legion Y7000P laptop equipped with an Intel Core i7-14650HX processor, 16 GB RAM, and an NVIDIA GeForce RTX 4050 GPU. The algorithms were implemented using Python (version 3.9) in the Spyder integrated development environment. The key Python libraries NumPy and Pandas were used for data manipulation, whereas Scikit-learn (version 1.2; INRIA, Paris, France) was used for RDA, data preprocessing, model evaluation, and hyperparameter search (RandomizedSearchCV). The model persistence was assessed using Joblib. The source code and data are available at the repository provided in the "Data and code availability" section.

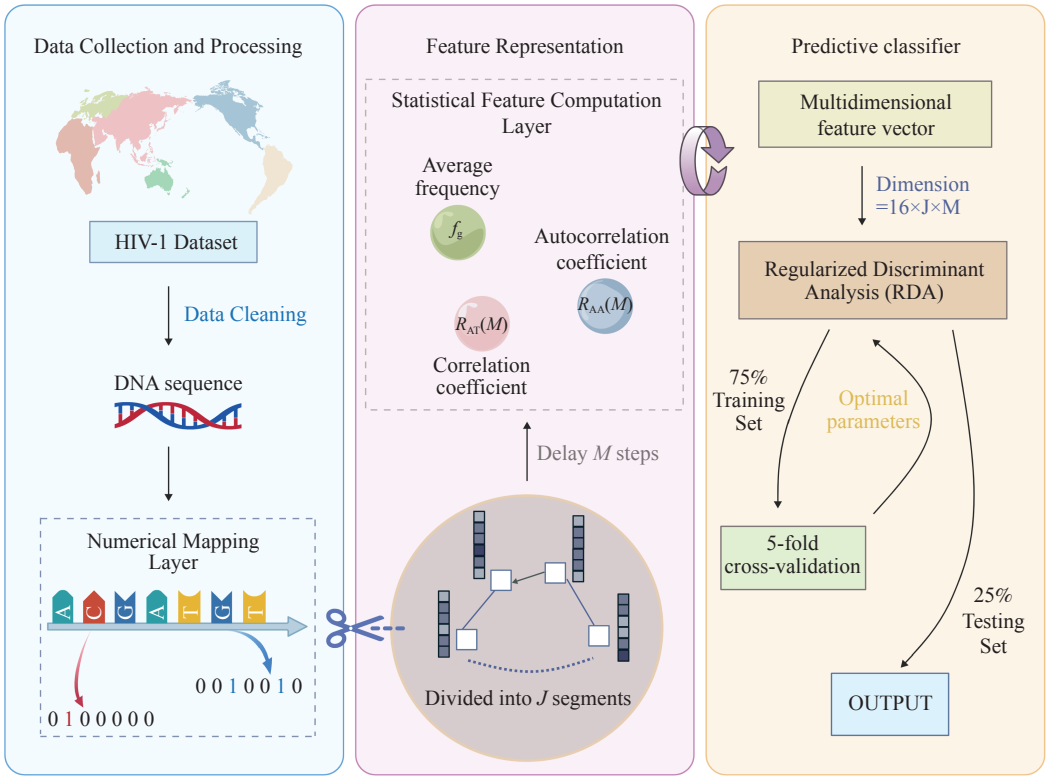


FIGURE 2. Framework diagram of the SCCFV method.
Note: created using the BioGDP tool (https://BioGDP.com) (8)
Abbreviation: HIV=human immunodeficiency virus; RDA=regularized discriminant analysis; SCCFV=subsequence correlation coefficient feature vector.

RESULTS

Classification of the HIV Sequence Datasets

We analyzed HIV subtypes with more than nine sequence entries for the full genome and *pol*, *env*, *gag*, *nef*, *rev*, *tat*, *vif*, *vpr*, and *vpu* genes. The model built for each gene demonstrated excellent and robust performance (Table 1).

Despite significant variations in the sample size and number of subtypes covered across the datasets, all gene segment models achieved accuracy rates exceeding 99.85%. This demonstrates that the proposed method possesses excellent generalization capability and robustness.

Classification of the Chinese Dataset

The model achieved 99.7% accuracy in classifying 25% of the Chinese HIV *pol* dataset used as an independent test set.

We compared our model with COMET, HIVdb, and SNV using this same *pol* independent test set (Table 2).

Compared with the HIVdb algorithm, which was developed using European and American B subtype data, our method demonstrated comparable performance in classifying the B subtype. However, for the CRF07_BC and CRF08_BC strains prevalent in China, our model achieved recall rates of 0.9882 and 0.9991, respectively, significantly outperforming HIVdb (recall: 0.6002 and 0.7873, respectively). Furthermore, compared with the COMET method that was based on European and American training data, our model clearly demonstrated better recall performance across all categories. The SNV-feature-based linear discriminant analysis classifier performed poorly, with the recall and F1-scores for all subtypes falling below those of the SCCFV method.

We further validated the SCCFV-RDA model using 821 full-genome sequences encompassing the major subtypes circulating in China. The model achieved 100% accuracy, recall, precision, and F1-score for each subtype, indicating classification performance (Supplementary Table S1, available at https://weekly.chinacdc.cn/). These results confirm that the SCCFV-RDA method generalizes effectively to non-*pol* genomic regions, reinforcing its utility for predicting HIV-1 strains prevalent in China.

TABLE 1. Classification metrics for 10 international datasets.

Dataset	ParameterJ	Parameter M	Accuracy	Recall	Precision	F1-score	Number of categories	Sample size
Complete	5	5	1	1	1	1	34	18,147
pol	5	5	0.9988	0.9988	0.9989	0.9988	35	30,708
env	5	5	0.9986	0.9986	0.9986	0.9986	15	159,312
gag	5	5	0.9997	0.9997	0.9997	0.9997	34	95,637
ref	5	4	0.9988	0.9988	0.9989	0.9988	19	56,826
rev	5	5	0.9985	0.9985	0.9985	0.9985	26	94,842
tat	4	5	1	1	1	1	32	41,346
vif	5	4	0.9987	0.9987	0.9988	0.9987	31	47,792
vpr	4	5	0.9993	0.9993	0.9993	0.9993	19	46,621
vpu	5	5	0.9997	0.9997	0.9997	0.9997	24	105,808

TABLE 2. Comparison of the performance of four models in classifying Chinese HIV *pol* data.

Method (Acc)	SCCFV (99.70%)			SNV (98.80%)			HIVdb (82.41%)			COMET (73.51%)		
Subtype	Recall	Precision	F1-score	Recall	Precision	F1-score	Recall	Precision	F1-score	Recall	Precision	F1-score
B	0.9920	0.9964	0.9942	0.9414	0.9833	0.9619	0.9973	0.9574	0.977	0.9574	1	0.9782
C	0.9892	0.9684	0.9787	0.8280	0.8105	0.8191	0.9892	0.7863	0.8762	0.7634	0.9861	0.8606
CRF01_AE	0.9989	0.9972	0.9981	0.9912	0.9862	0.9887	0.9870	0.9970	0.9920	0.7995	0.9996	0.8884
CRF07_BC	0.9982	0.9980	0.9981	0.9885	0.9818	0.9851	0.6002	0.9971	0.7494	0.5459	1	0.7063
CRF08_BC	0.9991	0.9973	0.9982	0.9873	0.9846	0.9859	0.7873	1	0.8810	0.9545	1	0.9767
CRF55_01B	0.9778	0.9925	0.9851	0.9275	0.9573	0.9421	0.8935	1	0.9438	0.892	1	0.9429

DISCUSSION

This is the first study to introduce a novel HIV-1 subtype classification method that integrates SCCFV with an RDA classifier. As an alignment-free approach, it effectively extracts discriminative features from HIV-1 sequences. It proved to be faster and more accurate than sequence alignment-based tools such as COMET and HIVdb and showed improvements across all classification metrics when evaluated against the alignment-free SNV method.

Our model — innovatively trained on a domestic sequence database — captures local genetic characteristics and addresses the limitations of international tools in classifying Chinese data. Moreover, unlike sequence alignment-based methods, the feature vector for each sequence need only be computed once, enabling the construction of a comprehensive feature vector database for all HIV-1 strains in China. Notably, the generalizability of our model beyond the *pol* region was confirmed using an independent set of 821 full-genome sequences.

Furthermore, the algorithm is designed to also predict subtypes for strains circulating globally. For instance, it achieved 99.88% classification accuracy on a global HIV-1 *pol* gene dataset encompassing 35 subtypes and 30,708 sequences, including 7,178 from Asia and 6,094 from Africa, demonstrating its robust generalizability.

However, this classification method has several limitations. Model performance depends on the quality of the annotated data. The model capability for discriminating rare subtypes with extremely limited sample sizes or URFs requires further validation with increased data accumulation. The model also lacks integration with relevant sociodemographic factors. Additionally, its ability to identify novel subtypes still requires its integration with more in-depth biological experiments.

In summary, we have generated a much-needed method for subtyping HIV-1 strains in China on the basis of specific gene fragments and full-genome sequences. This alignment-free and generalizable method can be directly applied to build a feature vector database of all circulating HIV-1 variants in China. Studies of its suitability for monitoring other rapidly evolving viruses are warranted.

Conflicts of interest: No conflicts of interest.

Funding: Supported by the National Natural Science Foundation of China (Nos. 12031016 and 12531012), Interdisciplinary Construction of Bioinformatics and Statistics, and Academy for Multidisciplinary Studies, Capital Normal University.

doi: 10.46234/ccdcw2026.080

[#] Corresponding author: Hengjian Cui, hjcui@bnu.edu.cn.

[1] School of Science, Beijing University of Civil Engineering and Architecture, Beijing, China; [2] School of Intelligence Science and Technology, Beijing University of Civil Engineering and Architecture, Beijing, China; [3] Department of Mathematical Sciences, Tsinghua University, Beijing, China; [4] National Key Laboratory of Intelligent Tracking and Forecasting for Infectious Diseases, National Center for AIDS/STD Control and Prevention, Chinese Center for Disease Control and Prevention & Chinese Academy of Preventive Medicine, Beijing, China; [5] School of Mathematical Sciences, Capital Normal University, Beijing, China.
[&] Joint first authors.

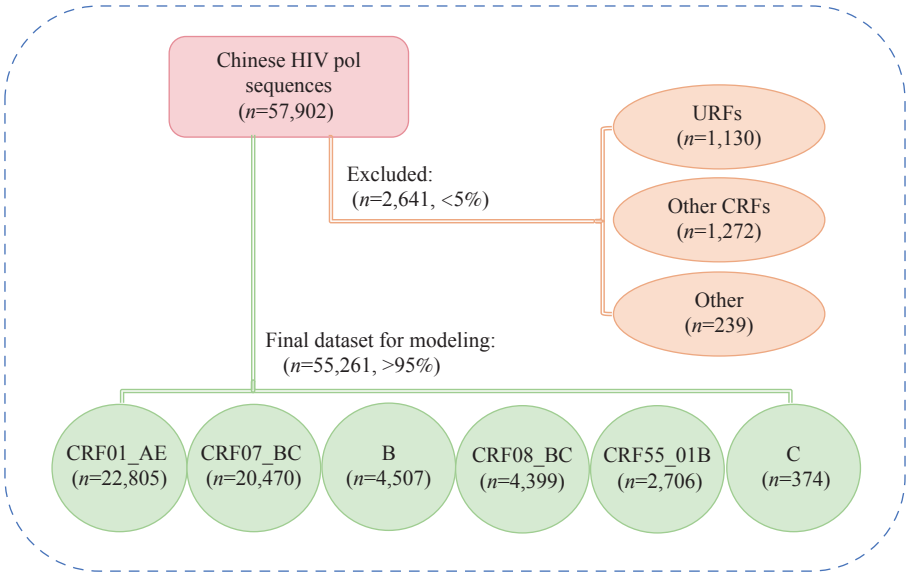
Submitted: January 11, 2026
Accepted: March 25, 2026
Issued: April 17, 2026

REFERENCES

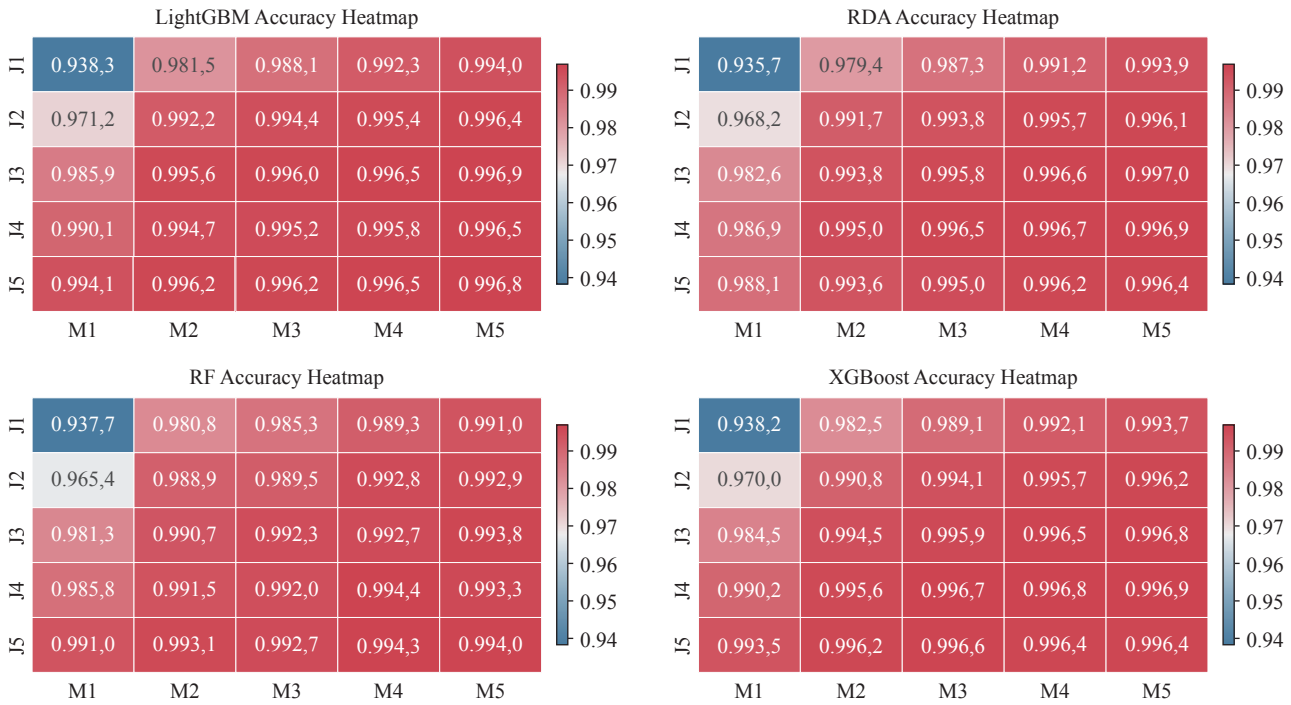
1. UNAIDS. 2025 Global AIDS Update. Geneva: Joint United Nations Programme on HIV/AIDS. 2025. https://www.unaids.org/en/resources/documents/2025/2025-global-aids-update. [2025-7-10]
2. Désiré N, Cerutti L, Le Hingrat Q, Perrier M, Emler S, Calvez V, et al. Characterization update of HIV-1 M subtypes diversity and proposal for subtypes A and D sub-subtypes reclassification. Retrovirology 2018;15(1):80. https://doi.org/10.1186/s12977-018-0461-y.
3. Liu X, Wang D, Hu J, Song C, Liao, LJ, Feng Y, et al. Changes in HIV-1 subtypes/sub-subtypes, and transmitted drug resistance among ART-naïve HIV-infected individuals—China, 2004-2022. China CDC Wkly 2023;5(30):664 – 71. https://doi.org/10.46234/ccdcw2023.129.
4. Hemelaar J, Elangovan R, Yun J, Dickson-Tetteh L, Fleminger I, Kirtley S, et al. Global and regional molecular epidemiology of HIV-1, 1990-2015: a systematic review, global survey, and trend analysis. Lancet Infect Dis 2019;19(2):143 – 55. https://doi.org/10.1016/S1473-3099(18)30647-9.
5. Hu L, Li ZF, Tang ZH, Zhao C, Zhou X, Hu PW. Effectively predicting HIV-1 protease cleavage sites by using an ensemble learning approach. BMC Bioinformatics 2022;23(1):447. https://doi.org/10.1186/s12859-022-04999-y.
6. He L, Sun SY, Zhang QY, Bao XN, Li PK. Alignment-free sequence comparison for virus genomes based on location correlation coefficient. Infect Genet Evol 2021;96:105106. https://doi.org/10.1016/j.meegid.2021.105106.
7. Wade KE, Chen LH, Deng CT, Zhou G, Hu PZ. Investigating alignment-free machine learning methods for HIV-1 subtype classification. Bioinform Adv 2024;4(1):vbae108. https://doi.org/10.1093/bioadv/vbae108.
8. Jiang S, Li HQ, Zhang LWY, Mu WP, Zhang Y, Chen TJ, et al. Generic Diagramming Platform (GDP): a comprehensive database of high-quality biomedical graphics. Nucleic Acids Res 2025;53(D1):D1670 – 6. https://doi.org/10.1093/nar/gkae973.
9. Struck D, Lawyer G, Ternes AM, Schmit JC, Bercoff DP. COMET:

adaptive context-based modeling for ultrafast HIV-1 subtype identification. Nucleic Acids Res 2014;42(18):e144. https://doi.org/10.1093/nar/gku739.
10. Pineda-Peña AC, Faria NR, Imbrechts S, Libin P, Abecasis AB, Deforche K, et al. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. Infect Genet Evol 2013;19:337 – 48. https://doi.org/10.1016/j.meegid.2013.04.032.
11. Liu TF, Shafer RW. Web resources for HIV type 1 genotypic-resistance test interpretation. Clin Infect Dis 2006;42(11):1608 – 18. https://doi.org/10.1086/503914.
12. He L, Dong R, He RL, Yau SST. A novel alignment-free method for HIV-1 subtype classification. Infect Genet Evol 2020;77:104080. https://doi.org/10.1016/j.meegid.2019.104080.

SUPPLEMENTARY MATERIAL



SUPPLEMENTARY FIGURE S1. Screening and composition of the Chinese HIV-1 *pol* dataset.
Abbreviation: HIV=human immunodeficiency virus; URFs=unique recombinant forms; CRF=circulating recombinant form.



SUPPLEMENTARY FIGURE S2. Heatmap of parameter selection for the four classifiers.
Abbreviation: LightGBM=Light Gradient Boosting Machine; RDA=regularized discriminant analysis; RF=random forest; XGBoost=eXtreme Gradient Boosting.

SUPPLEMENTARY TABLE S1. Classification performance of SCCFV-RDA on Chinese full-length genome sequences.

Dataset	Recall	Precision	F1-score	Sample size
CRF01_AE	1	1	1	299
CRF07_BC	1	1	1	298
CRF08_BC	1	1	1	52
CRF103_01B	1	1	1	10
CRF140_0107	1	1	1	19
CRF55_01B	1	1	1	27
CRF85_BC	1	1	1	51
B	1	1	1	65

Abbreviation: SCCFV-RDA=subsequence correlation coefficient feature vector-regularized discriminant analysis; CRF=Circulating recombinant form.