Computational and Structural Biotechnology Journal 27 (2025) 149–159

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

MTIOT: Identifying HPV subtypes from multiple infection data

Qi Zhao [a,d,1], Tianjun Zhou [b,1], Lin Li [a,1], Guofan Hong [b,*], Luonan Chen [a,c,**]

[a] Key Laboratory of Systems Biology, Center for Excellence in Molecular Cell Science, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai 200031, China
[b] State Key Laboratory of Molecular Biology, Center for Excellence in Molecular Cell Science, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai 200031, China
[c] Key Laboratory of Systems Health Science of Zhejiang Province, School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Hangzhou 310024, China
[d] University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Keywords:
HPV
Machine Learning
Cervical carcinoma
Convolution

ABSTRACT

Persistent infection with high-risk human papillomavirus (hrHPV) is a major cause of cervical cancer. The effectiveness of current HPV-DNA testing, which is crucial for early detection, is limited in several aspects, including low sensitivity, accuracy issues, and the inability to perform comprehensive hrHPV typing. To address these limitations, we introduce MTIOT (Multiple subTypes In One Time), a novel detection method that utilizes machine learning with a new multichannel integration scheme to enhance HPV-DNA analysis. This approach may enable more accurate and rapid identification of multiple hrHPV types within a single sample. Compared to traditional methods, MTIOT has the potential to overcome their core limitations and offer a more efficient and cost-effective solution for cervical cancer screening. When tested on both simulated samples (to mimic real-world complexities) and clinical samples, MTIOT achieved F1 scores (the harmonic mean of sensitivity and specificity) of 98 % and 92 % respectively for identifying subtypes with a sample size $\geq$ 50, suggesting that it may significantly improve the precision of cervical cancer screening programs. This work with MTIOT represents a significant step forward in the molecular diagnosis of hrHPV and may suggest a promising avenue for enhancing early detection strategies and potentially reducing the incidence of cervical cancer. This study also underscores the importance of methodological innovation in tackling public health challenges and sets the stage for future clinical trials to validate MTIOT's efficacy in practice.

1. Introduction

Cervical cancer is the fourth most common cancer among women. In 2020, an estimated 604,000 women worldwide were diagnosed with cervical cancer, and about 342,000 women may have died from the disease [1,2]. However, when diagnosed early and managed effectively, cervical cancer is one of the most potentially successfully treatable forms of cancer. Screening may allow pre-cancerous lesions to be identified at stages when they can be relatively easily treated. Cancers diagnosed in late stages can also potentially be controlled through appropriate treatment and palliative care. With a comprehensive approach encompassing prevention, screening, and treatment, cervical cancer may be eliminated as a public health problem within a generation [3–5]. Consequently, screening for cervical cancer is of utmost importance. The primary cause of pre-cancerous and cancerous cervical lesions is infection with a high-risk or oncogenic HPV type. Nearly all cervical cancer cases (99 %) are linked to infection with high-risk human papillomaviruses (HPV), an extremely common virus transmitted through sexual contact. Just two high-risk HPV strains (16 and 18) may cause more than 70 % of cervical cancers, but they can potentially be treated if detected

* Corresponding author.
** Corresponding author at: Key Laboratory of Systems Biology, Center for Excellence in Molecular Cell Science, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai 200031, China.
E-mail addresses: zhaoqi2021@sibcb.ac.cn (Q. Zhao), tjzhou@sibs.ac.cn (T. Zhou), lilin6@sibcb.ac.cn (L. Li), gfhong@sibcb.ac.cn (G. Hong), lnchen@sibcb.ac.cn (L. Chen).
[1] These authors contributed equally: Qi Zhao, Tianjun Zhou, Lin Li

https://doi.org/10.1016/j.csbj.2024.12.005
Received 13 May 2024; Received in revised form 6 December 2024; Accepted 6 December 2024
Available online 16 December 2024

early enough [4,5]. Based on this understanding, effective cervical cancer screening essentially involves the precise detection of high-risk HPV types, with the aim of early identification and management of viral infections associated with cervical cancer. This approach is intended to potentially reduce the incidence of cervical cancer or facilitate early intervention in its development. As recommended by the WHO as the top choice method for cervical cancer screening, HPV-DNA testing is an objective diagnostic that leaves no room for interpretation of results. It has been demonstrated to be simpler, may prevent more pre-cancers and cancers, and may save more lives. It is also potentially more cost-effective than visual inspection techniques or cytology (commonly known as 'pap smears') [6,7]. Among the numerous HPV-DNA testing methods, the most traditional and regarded as the gold standard is based on a PCR-Sanger sequencing framework [8], which captures specific sequences and compares them with standard sequences to identify infections [9,10]. Although PCR-Sanger sequencing can provide relatively accurate genotyping, in cases of multiple HPV infections, overlapping peaks in Sanger sequencing results can render it impossible to read as base sequences. This may necessitate the use of different primers for multiple experiments to determine the number of subtypes present, making detection extremely time-consuming, labor-intensive, and costly [11–13]. At the same time, the market offers a wide variety of HPV testing kits capable of detecting and differentiating multiple HPV subtypes in a single experiment. These testing kits are mainly based on: second-generation hybrid capture [14,15], real-time fluorescent PCR [16], enzyme-linked immunosorbent assay [17], and gene chip technology [18–20]. Essentially, these methods detect specific nucleotide sequences through probe hybridization. However, probe hybridization methods face significant challenges in accurately typing HPV [21], and they may not be able to distinguish highly similar HPV subtype nucleotide sequences. Additionally, the uneven amplification efficiency of different genotypes in PCR-based hybridization methods can lead to errors in HPV detection and genotyping. Therefore, these methods may suffer from low detection sensitivity, poor accuracy, and the inability to perform comprehensive subtype testing [22]. To address these challenges, we have developed a novel method named MTIOT, which is based on the PCR-Sanger sequencing framework and enhanced by advanced machine learning technology for predictive analysis. MTIOT employs random convolutional kernels and a new multi-channel integration approach to analyze Sanger sequencing data from samples with multiple infections. By training a machine learning model to recognize complex patterns, MTIOT may overcome the unreadability issues associated with overlapping peaks, enabling accurate differentiation of various HPV subtypes without sacrificing detection speed or increasing costs. The primary objective of this research is to enhance the accuracy and efficiency of detecting high-risk HPV subtypes. Specifically, genotyping tests for at least 13 high-risk HPV subtypes are essential, and our method may allow for the detection of these coinfections in a single-tube PCR reaction. We anticipate that this will potentially facilitate the early diagnosis of precancerous lesions, significantly improving the early detection rates for high-risk HPV subtypes and contributing to cervical cancer prevention. Thus, combining PCR-Sanger sequencing technology with advanced machine learning analysis represents a crucial step towards early intervention in cervical cancer, potentially supporting the long-term goal of eliminating it as a public health issue.

2. Results

2.1. MTIOT on simulation dataset and clinical dataset

To validate the effectiveness of MTIOT, we conducted evaluations using both a simulated HPV dataset and real clinical samples. From the simulation dataset, we selected 453 valid samples across 18 subtypes, each with a minimum of ten samples. A stratified 10-fold cross-validation approach was employed [23–26]. For each subtype, its samples were designated as positive cases, while all other samples were treated as negative cases. During the training, the weights were adjusted based on the ratio of positive to negative cases.

For subtypes with ten or more samples, the average sensitivity and specificity achieved were 0.67 and 0.99, respectively. Subtypes with twenty or more samples demonstrated average sensitivity and specificity of 0.81 and 0.98, respectively. Notably, for subtypes with more than fifty samples, the average sensitivity and specificity reached an impressive 0.98 each, as shown in the "Simulation" column of Table 1.

Fig. 1 illustrates that the model's performance tends to improve as the sample size increases. This suggests that MTIOT may achieve nearly perfect sensitivity and specificity, approaching 100 %, given a sufficient large sample size.

Additionally, we utilized the entire simulation dataset for training and tested the model on 95 real clinical samples, comprising 90 single HPV infections and five compound HPV infections. The results for the clinical data are presented in the "Experiment" column of Table 1. The performance on the actual samples closely mirrored that of the simulated data, albeit slightly lower. This outcome further validates the effectiveness of MTIOT and underscores the positive impact of increased data volume on enhancing its detection performance.

We observed that in certain HPV subtypes, such as HPV-31 and HPV-56, the diagnostic accuracy on the clinical sample dataset may exceed that on the simulated sample dataset. There are the following possible explanations for these observations. Firstly, there is a distributional disparity between simulated and real (clinical) datasets. On these specific subtypes, the model may have adjusted its parameter space to reduce the error on simulated data, resulting in better statistics for the experimental samples. This suggests that while simulated data is helpful for training, the difference from real-world clinical data can affect the model's generalizability. Secondly, our test dataset contained a high proportion of single HPV infection samples. For subtypes like HPV-31 and HPV-56, the model may have obtained sufficient features to accurately identify these single-infection sequence data. This phenomenon indicates a specific strength of the model in recognizing these subtypes. The incorporation of randomness in the feature extraction steps of MTIOT may prove advantageous. This randomness allows a wider range of to be explored during feature extraction, potentially enhanceing model performance in real-sample identification.

The MTIOT results for the five real composite infection samples are shown in Table 2. At first glance, they are not ideal. But the MTIOT results were good for the two subtypes HPV-16 and HPV-52 with a train size greater than 50. MTIOT has been shown to be effective for simulated samples, but will need more data to overcome the learning cost for the deviation between simulated and real samples. The data volume for HPV-16 and HPV52 is, however, sufficient for MTIOT to perform effectively. Obtaining sufficient real composite samples and verifying their effectiveness on real samples will be the focus of future work.

In the upcoming comparative analysis, we will focus on subtypes with a sample size of 20 or more within the simulation dataset to ensure sufficient sample volume to support variations in experimental conditions and obtain optimal results. Through this approach, we aim to more accurately assess the efficacy of MTIOT in detecting various HPV-DNA subtypes, ensuring the reliability and representativeness of the experimental outcomes. Furthermore, setting the sample size to 20 or more will help reduce biases caused by insufficient sample numbers, thereby enhancing the robustness of the experimental design and the universality of the conclusions.

2.2. MINIROCKET and other feature extraction methods

In the simulated dataset where N is greater than or equal to 20, we compared the effects of the ROCKET 'family' methods (namely, ROCKET [27] and its variants MINIROCKET [28] and MultiRocket [29]), Hydra [30], and their combined usage under the same settings as in Section 'MTIOT on simulation dataset and clinical dataset'. The methods of the ROCKET "family" and Hydra may be among the most accurate and

Table 1
Results of MTIOT on simulation and experiment datasets.

Subtype	Sample size(N)	Simulation			Experiment			
		Sensitivity	Specificity	F_1-Score[a]	Sensitivity	Specificity	F_1-Score[a]	Test size
HPV−16	53	1.00	0.99	1.00	0.88	0.99	0.93	8
HPV−52	51	0.96	0.97	0.97	0.86	0.96	0.91	7
HPV−18	37	0.95	1.00	0.97	0.83	1.00	0.91	6
HPV−31	30	0.70	0.96	0.81	1.00	0.98	0.99	5
HPV−58	28	0.60	0.97	0.74	0.33	0.96	0.84	4
HPV−66	28	0.77	0.98	0.86	0.75	0.95	0.50	4
HPV−6	27	0.77	1.00	0.87	1.00	0.98	0.99	1
HPV−68	26	0.77	0.97	0.86	0.43	0.89	0.58	7
HPV−33	25	0.93	0.99	0.96	0.67	1.00	0.80	6
HPV−56	22	0.83	0.99	0.91	1.00	1.00	1.00	6
HPV−53	20	0.60	0.97	0.74	0.00	1.00	0.00	5
HPV−39	17	0.25	0.98	0.40	1.00	0.31	0.48	6
HPV−81	16	0.80	1.00	0.89	1.00	0.98	0.99	4
HPV−35	15	0.45	0.98	0.62	1.00	0.97	0.98	5
HPV−11	14	0.65	0.99	0.79	1.00	1.00	1.00	3
HPV−54	11	0.30	0.99	0.46	0.50	1.00	0.67	4
HPV−62	11	0.30	1.00	0.46	0.20	1.00	0.33	5
HPV−84	11	0.50	1.00	0.67	0.75	1.00	0.86	4
Mean($N \geq 10$)	25	0.67	0.99	0.78	0.73	0.94	0.76	4
Mean($N \geq 20$)	32	0.81	0.98	0.88	0.70	0.98	0.77	5
Mean($N \geq 50$)	52	0.98	0.99	0.98	0.87	0.98	0.92	5

[a] The F1 Score mentioned in this article refers to the harmonic mean of sensitivity and specificity, rather than the accuracy and recall rate

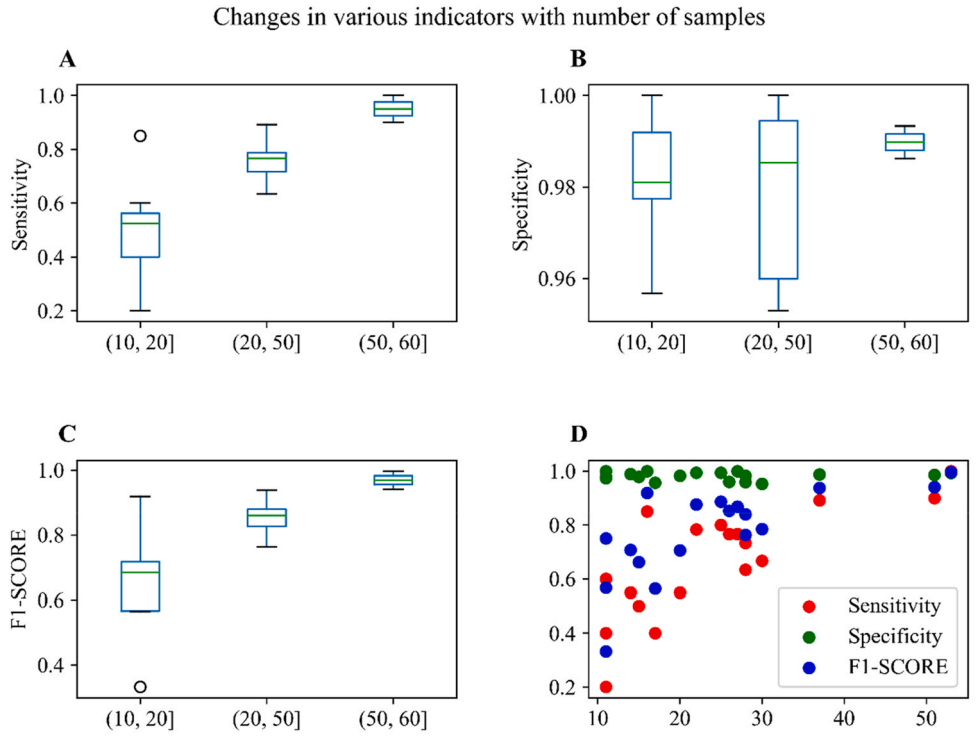

Fig. 1. Changes in various indicators with number of samples. (A) Sensitivity changes with different sample sizes; (B) The variation of specificity in different sample sizes; (C) The change of F1 score in different sample sizes; (D) In the overall scatter diagram, the x-axis is the sample size, and the y-axis is the value of the three indicators.

fastest ones on the UCR archive dataset currently. The ROCKET series methods are based on random convolution kernels, while Hydra is a method that potentially bridges ROCKET and dictionary methods. Both are used for feature extraction, and the extracted features are employed to train a ridge regression classifier or logistic regression (when dealing with larger datasets).

Compared with other time series classification methods, the main advantage of the ROCKET series and Hydra is their speed. This is may be significant for the rapid and accurate detection of HPV-DNA required for clinical practice. For other state-of-the-art methods, training and testing times of other methods using on the 112 UCR datasets task days. The most precise method, HC-2 [31], takes about two weeks, while ROCKET, MINIROCKET, MultiRocket and Hydra respectively takes 2.85 h, 2.44 min, 15.77 min, and 41 min [29].

In terms of specificity, all methods are quite accurate in detecting different HPV subtypes (see Table 3). By approaching a perfect score of 1.00, methods appear to yield no false positives in identifying non-target sequences, critical in reducing misdiagnosis. The sensitivity of all

Table 2
Results of MTIOT on real samples.

	Label	Train_size	Prediction	Train_size
Sample1	HPV−62	11	HPV−39	17
	HPV−61	1		
Sample2	HPV−66	28	HPV−31	30
	HPV−58	28	HPV−68	26
	HPV−59	9		
	HPV−53	1		
Sample3	HPV−16	53	HPV−16	53
	HPV−84	11	HPV−39	17
	HPV−54	11		
Sample4	HPV−16	53	HPV−39	17
	HPV−52	51		
	HPV−68	26		
Sample5	HPV−52	51	HPV−52	51
	HPV−62	11		
	HPV−61	1		

methods also improved with sample size. For example sizes of 20 or more, the sensitivity ranges were 0.68–0.81, but 0.93–0.98 for sample size of 50 or more. Notably, MINIROCKET demonstrated outstanding performance at both the N ≥ 20 and N ≥ 50 thresholds, indicating its effectiveness in detecting HPV infections. For sample sizes of > 50, the combination of MINIROCKET and Hydra and Hydra alone performed almost as well as MINIROCKET.

MINIROCKET is thus the preferred feature extractor for MTIOT in terms of speed, specificity and sensitivity.

2.3. The impact of sequence trimming and noise reduction on performance of MTIOT

In Sanger sequencing, the initial unreadable portion of the sequence is normally trimmed. In our method, however, peaks don't need to be interpreted as bases, as the signal only needs to be sufficiently specific to distinguish the various HPV subtypes. Therefore, we need to consider the relationship between noise reduction and information loss that may be caused by trimming the sequence.

Indices clearly follow a set pattern depending on the trimming length (Fig. 2). Subtypes with sample sizes of 20 or more, indices attain similar peak values when either 200 or 500 nucleotides are trimmed (see Supplementary Figure S1). For subtypes with sample sizes of 50 or more, a trimming length of 500 clearly yields better scores. When the sample size is low, it is best to only remove the least readable part of the sequence (≤200). The information and noise of the remaining initial sequence appear to counterbalance each other, achieving an effect

Table 3
Comparison of Sensitivity, Specificity and F1-Score Across Models at Various Sample Sizes.

Method	N ≥ 20			N ≥ 50		
	Sensitivity	Specificity	F1-SCORE	Sensitivity	Specificity	F1-SCORE
Hydra	0.75	0.98	0.84	0.95	0.98	0.97
ROCKET	0.71	0.99	0.82	0.95	0.99	0.97
MINIROCKET	0.81	0.98	0.88	0.98	0.98	0.98
MultiRocket	0.67	0.99	0.78	0.90	0.99	0.94
Hydra+ROCKET	0.73	0.98	0.83	0.94	0.99	0.96
Hydra+MINIROCKET	0.77	0.98	0.85	0.95	0.99	0.97
Hydra+MultiRocket	0.71	0.98	0.82	0.92	0.99	0.95

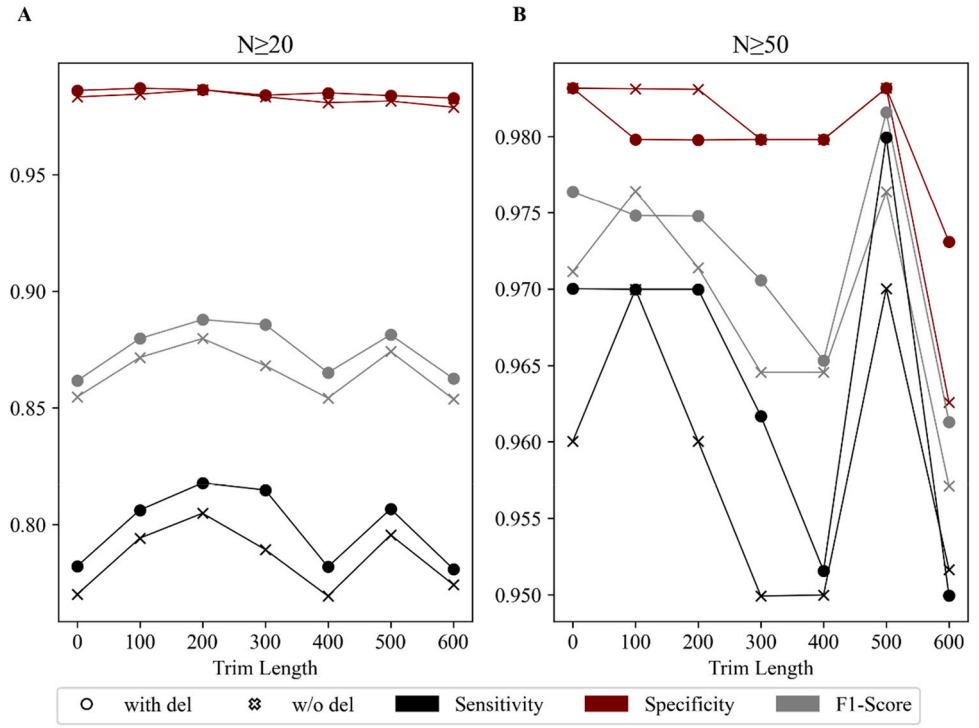

Fig. 2. Impact of trim length and noise deletion or not on sensitivity, Specificity, and F1-Score (A) For subtypes with sample size of 20 or more, the index changes under different pruning lengths and noise removal settings. (B) For the subtypes with a sample size of 50 or more, the index changes under different pruning lengths and noise removal settings.

similar to complete trimming (see Supplementary Figure S1).

In Sanger sequencing, signal fluctuations are mostly ignored, and standard peak shapes are converted into bases for sequence alignment. However, in our method, these signal variations may contain valuable information. We, therefore, compared the outcome after retaining or removing peak height data. For subtypes with sample sizes of 20 or more, retaining peak height information consistently outperformed analysis after peak data (Fig. 2). For subtypes with a sample sizes of 50 or more, the results of retaining peak data are mostly superior to or equivalent to those of noise removal. This suggests that due to the specialized handling of HPV-DNA sequencing data by MTIOT, it may be capable of capturing information from slight signal fluctuations typically considered as noise.

In summary, appropriately adjusting the sequence trimming length and retaining peak height data can improve MTIOT performance. Therefore, for trimming and denoising, we selected a trimming length of 500 and no noise removal, which also represents the potentially best-performing combination in our experiment.

2.4. The impact of multi-channel processing on performance of MTIOT

In the MTIOT process, each ab1 file contains data for four channels that correspond to the four nucleotides: adenine (A), thymine (T), guanine (G), and cytosine (C). Each channel may have significant biological relevance, making it crucial to assess whether the information for each nucleotide is potentially obscured or underutilized.

In the official implementations of ROCKET, MINIROCKET, Multi-Rocket, and Hydra, the method for processing multi-channel data involves randomly selecting some channels for convolution and summation [32–35]. However, this approach may inevitably lead to information loss. Therefore, we have employed a novel method: extracting features separately from the four nucleotide channels and then concatenating them to form the final feature set of the samples. Moreover, we explored several other approaches, including concatenating each channel's features with those obtained from random convolution and summation, and performing soft/hard voting after classifying the four channels separately.

Table 4 shows their varying performances in terms of sensitivity, specificity, and F1 scores. Clearly, the feature concatenation method significantly outperformed the other methods. The reason may be that feature concatenation for each channel better preserves its unique properties compared to random channel convolution and summation. The information is thus more effectively combined than by the voting methods.

3. Discussion

The development and validation of the MTIOT method for typing multiple HPV infections through a single experimental process appears to provide a significant improvement in HPV detection and genotyping. This study underscores the complexity and diversity of HPV genotypes and the potential critical need for accurate, efficient, and comprehensive subtyping methods to improve cervical cancer screening and diagnostic procedures.

The utilization of machine learning with a random convolutional kernel scheme, specifically the integration of the MINI-ROCKET feature extraction technique and ridge regression classifier, along with a new multi-channel integration method within the PCR-Sanger sequencing framework, potentially exhibits a novel approach to overcome the limitations faced by traditional HPV genotyping methods.

Our findings demonstrate that MTIOT can achieve F1 scores of 98 % and 92 % for identifying subtypes with sample sizes of 50 or more in both simulation and clinical datasets, respectively. This indicates its potential for improved accuracy in clinical diagnostics.

A critical challenge in HPV genotyping, especially in the case of multiple infections, is to distinguish closely related HPV genotypes. Traditional PCR-Sanger sequencing, while mostly accurate for single infections, struggles in the case of multiple infections. Current methods based on probe hybridization, such as second-generation hybrid capture and real-time fluorescent PCR, though widely used, suffer from lower sensitivity and specificity and often fail to provide comprehensive subtype testing.

MTIOT in part overcomes the some limitations of Sanger sequencing by detecting samples with multiple infections. Sanger sequencing typically uses universal primers for uncharacterised samples. For single infections, it yields a clear and readable signal peak. This signal peaks provide a DNA sequence, which is then compared to a sequence data bank to identify the infectious agent. However, for multiple infections, Sanger sequencing may produce overlapping signal peaks, causing difficulties in identifying the infectious agent. To overcome these limitations, dedicated primers for each HPV subtype are used in the Sanger sequencing. For m subtypes, m sequencing runs would thus be required. While MTIOT shares the initial stages of standard Sanger sequencing, it diverges thereafter. Following the method described in Section 'MTIOT Method Process', we may employ machine learning techniques to extract features from the fluorescent signals of the four different bases. These features are then combined and fed into a number of different HPV subtype-specific classifiers for classification, yielding corresponding binary identification results (Fig. 3). In particular, this approach does not require the samples to be singular or multiple infections.

The application of advanced machine learning techniques to interpret complex Sanger sequencing data presents a significant methodological advancement. By leveraging the distribution of sequencing data from multiple infection samples, the MTIOT method effectively address the challenges such as overlapping peaks and unreadable sequences inherent in Sanger sequencing results, offering a robust solution for accurate HPV genotyping in the presence of multiple infections.

This study also emphasizes the crucial role of feature extraction techniques in enhancing machine learning model performance for biological data analysis. MINIROCKET's ability to efficiently process and extract meaningful features from sequencing data, combined with the predictive power of the ridge regression classifier, exemplifies the synergy between bioinformatics and machine learning. This integration is further enhanced by the innovative multi-channel integration approach, which improves the model's capability to handle complex biological datasets. This interdisciplinary approach not only facilitate accurate typing of HPV infections but also opens avenues for applying similar methodologies in other areas of molecular diagnostics.

However, the study acknowledges limitations, including the small number of multiple HPV infection samples and the need for further

Table 4
Comparison of sensitivity, specificity and F1-score across different multi-channel processing methods with varying sample sizes.

Method	N ≥ 20			N ≥ 50		
	Sensitivity	Specificity	F1-SCORE	Sensitivity	Specificity	F1-SCORE
Official	0.77	0.98	0.86	0.95	0.99	0.97
Concatenate	0.81	0.98	0.87	0.98	0.98	0.98
Concatenate + origin	0.79	0.98	0.87	0.96	0.98	0.97
Soft vote	0.74	0.98	0.84	0.93	0.98	0.95
Hard vote	0.67	0.97	0.78	0.91	0.96	0.93

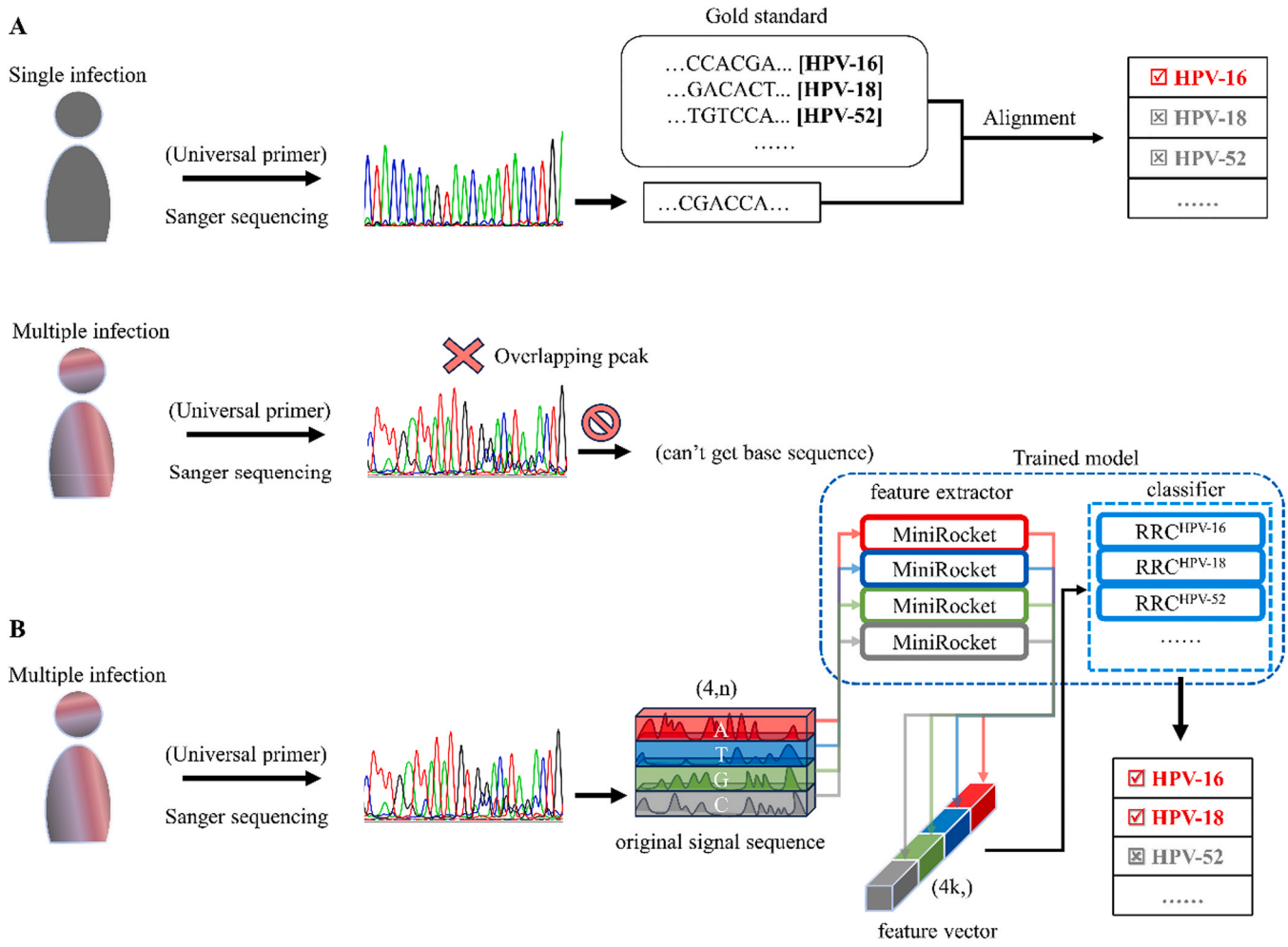

Fig. 3. Differences between the MTIOT and typical PCR Sanger sequencing methods. (A) Typical PCR-Sanger Sequencing method: For a single infection, the universal primer is employed to amplify the target DNA, followed by Sanger sequencing. The clear peaks generated facilitate DNA sequence alignment and identification of the HPV type, such as HPV-16, which can be matched with a gold standard reference sequence. However, in the case of multiple infections, typical Sanger sequencing yields overlapping peaks that obstruct accurate base sequence determination. (B) MTIOT: This method leverages MINIROCKET feature extractors to convert the original signal sequences into feature vectors. These vectors are subsequently classified by a trained model set using Ridge Regression Classifier (RRC), to differentiate between various HPV types. For instance, HPV-16 and HPV-18 can be accurately identified as positive, while HPV-52 is correctly classified as negative, thereby effectively overcoming the limitations of the conventional approach.

validation with larger and more diverse clinical samples. Future research should focus on refining the MTIOT algorithm to enhance its utility across a broader range of HPV genotypes and explore its potential with existing screening programs to improve cervical cancer prevention strategies.

In conclusion, the MTIOT method represents a promising advancement in HPV genotyping technology, offering a high-throughput, accurate, and cost-effective alternative to existing methods. Its development is timely, given the increasing recognition of the HPV's role in cervical and other cancers, and highlights the potential of combining traditional molecular techniques with machine learning to tackle complex diagnostic challenges. Moving forward, rigorous clinical validation is essential to confirm its efficacy and reliability before considering its integration into routine clinical practice, ensuring improved outcomes for individuals at risk of HPV-related diseases.

4. Materials and methods

4.1. Sequencing data set of multiple HPV infection samples

Given that the number of patient samples with multiple infections was relatively low and the quantity of such multiple-infected samples required for our study is potentially insufficient, we decided to simulate mixed infection samples by utilizing laboratory-confirmed single-infection samples with identified subtypes. For PCR amplification, we employed the LoTemp HiFi DNA Polymerase Ready Mix System, due to its high efficiency and fidelity. This should ensure that the DNA amplification products from single-infection samples would be as accurate as possible. The amplified products were mixed in equal proportions to simulate a multiple infection environment, followed by sequencing of these mixed samples. This step aimed to generate a set of simulated samples that could potentially serve as input data for further analysis using the MTIOT method. Through this approach, we were able to create laboratory-generated sample data that may closely approximate natural infection scenarios. Although artificially generated under ideal conditions, it could potentially simulate multiple infections in the real world. We thus created a dataset of 200 samples, including 22 unique HPV subtypes, 148 double, 45 triple, 4 quadruple, and 1 quintuple infection sample (Supplementary Table S1). Each multiple infection sample was futuremore considered a representative sample for each infection subtypes it contains. All infection samples containing a distinct HPV subtype were thus classified as positive cases, while all other samples were classified as negative cases. A dual infection sample infected with both HPV-16 and HPV-52, was therefore regarded as a sample for both the

HPV-16 and HPV-52 subtypes. Consequently, our dataset effectively expanded to 460 samples in terms of the number of effective samples, as shown in Supplementary Figure S2 and 6. This approach enables us to conduct a multi-dimensional in-depth analysis and understanding of the interactions between different subtypes and their distribution in cases of multiple infections, potentially providing a rich data foundation for further research.

4.1.1. Clinical sample source

The sample simulation employs single infection clinical samples from Changning District Maternal and Child Health Care Hospital (CMCHCH) [36], Jiading District Maternal and Child Health Care Hospital (JMCHCH), and Fudan University Shanghai Cancer Center. All biological samples were collected in accordance with ethical guidelines. Prior to sample collection, all donors were thoroughly informed about the purpose of the study and the use of their samples for method development, ensuring they understood their rights and the nature of the research. The study protocol was approved by the relevant institutional ethics committee, ensuring compliance with ethical standards in research. Medical staff use a cervical exfoliated cell collector to sample at the squamocolumnar junction within the cervical canal. The collected cells are suspended in PreservCyt, Surepath fixative, or Thinprep cytology test. After collection, the samples are lysed [37]. The lysate is then stored at −20°C for a period of 10–15 years. Lysates from various HPV subtypes are randomly selected for subsequent analysis.

4.1.2. Sample detection and preparation of amplification products

Lotemp HiFi DNA polymerase ready-to-use mix system [38] was employed to detect HPV genotyping of samples and to prepare amplification products. The detection process followed the method outlined in reference [36], though the primers of the nested PCR were replaced by GP6+ and MY11. The nested PCR products were sequenced using GP6 + as the sequencing primer. Sequences exceeding 50 bases were compared to the HPV DNA database in GenBank. Sequences were perfectly consistent with the standard HPV genotype DNA from GenBank, thus enabling the definitive determination of the HPV genotype infecting the selected samples.

4.1.3. Sequencing of amplified products from single infection sample

Nested PCR products were sequenced using GP6 + as the sequencing primer. Sequences longer than 50 bases were selected for comparison with the HPV DNA database in GenBank. The sequencing results demonstrated 100 % consistency with standard HPV genotype DNA from GenBank, enabling definitive determination of the HPV genotype infecting the selected samples.

4.1.4. Sequencing of multiple infection samples

Amplified products from multiple infection samples were mixed and sequenced using the same sequencing method as for single infection samples. During the study, we observed that increasing the number of samples containing various subtypes enhanced the effectiveness of subtype classifiers. This indicates that to improve the detection rate of a specific subtype, the sample size for that and other subtypes should be increased. Our simulated multiple samples included 22 subtypes with varying quantities for each. Notably, types 16 and 52 were predominant, consistent with the multiple infection data from clinical samples analyzed in our laboratory [36]. When selecting the mixed subtypes, we aimed to reflect the proportions of mixed infection subtypes found in clinical samples while also considering subtypes with lower clinical prevalence but high-risk associations. Following the predefined mixture, nested PCR amplification products were combined in equal proportions, and Sanger sequencing was performed using GP6 + as the sequencing primer. The sequencing result file of the simulated multiple infection samples served as the input data for MTIOT.

4.2. MTIOT method process

MTIOT is engineered to discern the composition of HPV genotypes within a sample featuring multiple infections. As illustrated in Fig. 4, MTIOT comprises the following procedural steps:

1. Sample Collection and PCR-Sanger Sequencing;
   First, samples were collected from the subjects. After collection, HPV DNA was amplified in two rounds using the nested-PCR technique. The first round of PCR aimed to amplify a larger DNA fragment, typically targeting the L1 region of the HPV genome. Subsequently, the product from the first round was used as a template for the second round of PCR amplification. A set of nested primers specifically designed were employed to further amplify a smaller, specific region within the first round amplification fragment. This method significantly enhanced the specificity and sensitivity of detection. After completing nested-PCR, the second round amplification products were sequenced using Sanger sequencing to generate sequence data files (usually in ab1 format). The primers used here have been described in Section 'Sequencing data set of multiple HPV Infection Samples', namely GP6 + , MY11 (for nested-PCR) and GP6 + (for Sanger sequencing).
2. Fluorescence Signal Extraction;
   Next, we analysed ab1 files generated by Sanger sequencing. The ab1 file format is a binary file that records the fluorescence intensity values as a function of time. We parse these files to extract fluorescence signal data. Each nucleotide (A, C, G, T) corresponds to a fluorescence signal curve, reflecting the changes in fluorescence intensity for that nucleotide during the sequencing process. Ideally, as shown in Fig. 5, these fluorescence signals can be presented as readable peaks, thereby being translated into a nucleotide sequence. However, in the case of HPV multiple infections, these signals become unreadable. In the MTIOT method, based on the extracted fluorescence signal data, we construct a matrix with dimensions (4, n), where 4 represents the four types of nucleotides (A, C, G, T) in the DNA sequence, and n represents the sequencing read length, that is, the spectral length. Each row represents the fluorescence signal of a nucleotide, while the columns correspond to time points or nucleotide positions during the sequencing process. This matrix is a mathematical representation of the raw data and provides a foundation for subsequent analysis.
3. Matrix Trimming and Data Padding;
   During the matrix preprocessing stage, we initially carry out precise trimming on the raw data matrix extracted from fluorescence signals. The objective is to eliminate low-quality sequencing signals that might be present at the ends of the sequences. This step is important as low-quality signals impair subsequent feature extraction and classification accuracy. The trimming strategy is based on signal quality and uses a threshold to determine which portions should be removed to ensure that only high-quality data is retained for downstream analysis. After the trimming process, data is padded to standardize data lengths following data trimming step and to ensure that the matrix shape meets the input requirements of subsequent models. The padding value is selected to be a static value close to the low signal values in the original signal in order to minimize interference during the feature extraction stage.
4. Feature Extraction

This step makes use of the MINIROCKET algorithm [28]. In first initializes 9996 one-dimensional convolutional kernels. With fixed lengths but different weights, dilation factors, and biases, to cover signal characteristics across various scales (Supplementary Table S2). The input data is a pre-processed HPV fluorescence signal matrix with dimensions (4, n), representing the signals for four types of nucleobases, where n is the sequence length. MINIROCKET applies each initialized convolutional kernels to convolve with each dimension of the input

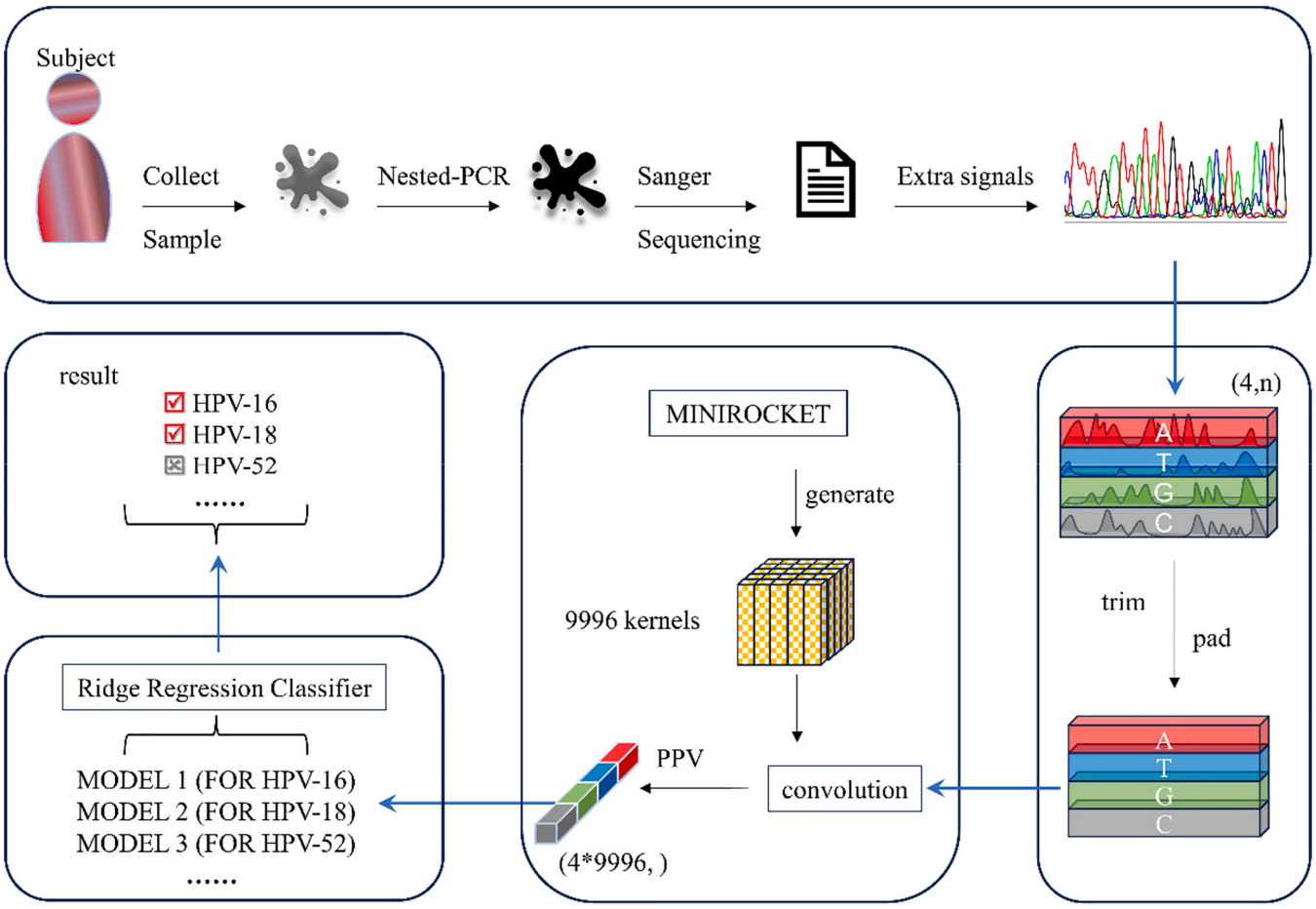

Fig. 4. MTIOT method flow chart, including (sample collection and Sanger Sequencing). Obtain a sample and conduct PCR and Sanger sequencing, generating the ab1 file containing sequencing results; (Fluorescence Signal Extraction) Extract the fluorescence signals corresponding to the four bases from the ab1 file, yielding a raw data matrix with dimensions (4, n), where n represents the length of the spectrum; (Matrix Trimming and Data Padding) Trim the matrix and pad the data to conform to the input requirements of subsequent models; (Feature Extraction) Employ MINIROCKET to individually extract feature from the four channels, concatenating the results to obtain a feature vector with a length of 4 * 9996; (Classification) Standardize the feature vectors, and input them into m ridge regression classifiers corresponding to m subtypes for classification, then obtain binary classification results from the m classifiers. The output of MTIOT comprises the identification result pertaining to the m subtypes.

fluorescence signals, to extract local features of the signal sequence. For the feature map for each convolutional kernel, MINIROCKET pools Proportion of Positive Values (PPV) and calculates the proportion of positive values. Each dimension generates 9996 features through 9996 convolutional kernels. Features from the four dimensions are concatenated to form a feature vector of length 4 * 9996 = 39984. This feature vector captures the key information of the fluorescence signal and represents the characteristics of the DNA sequence of the entire sample, for subsequent classification tasks. In the subsequent steps, MTIOT employs the extracted feature vector to train a ridge regression classifier. After the classifier is trained, the same convolutional kernels and operations are used to extract features from new data, and the trained classifier is used for prediction.

Convolutional kernel weight initialization. Each convolutional kernel has a fixed length of 9, and the weights are drawn from a set (α, β). As a result, the weights of the convolutional kernels can be described by the count of α or β. This leads to a definition where a kernel with a value of 1 consists of a set with one β, for example, [α, α, α, α, α, α, α, α, β]. Similarly, a kernel with a value of 2 includes a set with two βs, such as [α, α, α, α, α, α, α, β, β]. From this framework, it is determined that there are 2^9 = 512 possible combinations of weights. Among these, a subset containing 84 unique combinations, known as the 3-value kernel, was selected for our analysis:

[α, α, α, α, α, α, α, α, β],
[α, α, α, α, α, α, α, β, β],
[α, α, α, α, α, α, α, β, β],
……

This subset effectively balances the accuracy achieved using a minimal number of convolutional kernels with computational efficiency.

Convolutional kernel dilation initialization. In convolutional operations, the dilation coefficient determines the expansion of the convolutional kernel's coverage over the input data. For a given dilation coefficient d, a convolutional kernel performs convolution operations every d element of the input data. Each kernel is assigned a fixed set of dilation collections, adjusted according to the length of the input fluorescence sequences as follows:

$$D = \left(\lfloor 2^0 \rfloor, \ldots, \{\lfloor 2 \}^{\max} \rfloor\right) \quad (1)$$

where the exponents are uniformly distributed between 0 and max:

$$\max = log_2\left(l_{input} - 1\right)/\left(l_{kernel} - 1\right) \quad (2)$$

where l_{input} represents the length of the fluorescence sequence and l_{kernel} is the length of the convolutional kernel (which is 9). This ensure that the maximum effective length of the kernel, including dilation, is equal to the length of the input sequence.

Convolutional kernel bias initialization. Bias values are derived from the convolution outputs. By default, for each kernel/dilation

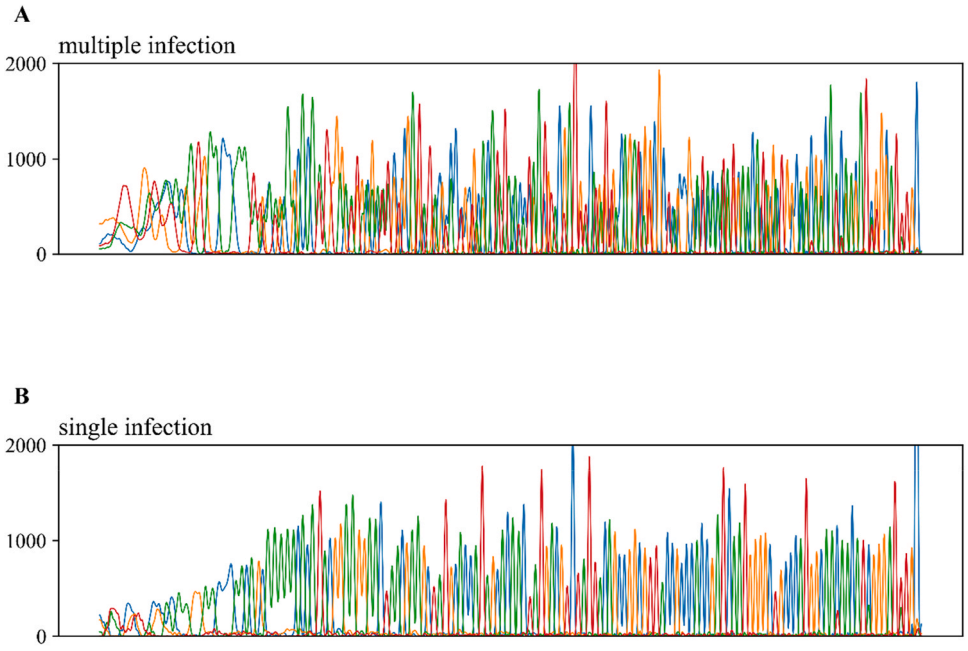

Fig. 5. (A) Sanger sequencing map of multiple infection samples showed overlapping peaks, which could not be read as base sequence; (B) Sanger sequencing map of single susceptible samples, with clear and readable signal peaks, can be transformed into base sequences for comparison.

combination, bias values are extracted from the quantiles of the convolution output of a single randomly selected training example. Specially, for a given convolutional kernel W and dilation coefficient d, we calculate the convolution output of a randomly chosen training example X, denoted as Wd*X. The bias values are taken as the quantiles at [0.25, 0.5, 0.75] of W_d*X.

Padding. MINIROCKET employs standard zero padding, which involves appending zeros to both the start and end of each sequence. This allows the first convolution operation to be centered at the beginning of the sequence and the last operation to be centered at the end of the sequence.

Ppv(Proportion of Positive Values). For the feature map output by each convolutional kernel, we perform PPV (Positive Predictive Value) pooling operations. This involves calculating the proportion of positive values within the feature map. The PPV value can be expressed as:

$$PPV(X * W - b) = \frac{1}{n}\sum[X * W - b > 0] \quad (3)$$

5. Classification

After obtaining the feature vector as described above, an independent ridge regression classifier is constructed for each subtype of HPV. Ridge regression is a statistical learning method used for classification tasks, incorporating a ridge penalty term based on ordinary least squares to reduce overfitting in model parameters. It is an enhanced version of linear regression designed to address issues that arise when the number of features exceeds the number of samples. Unlike ordinary least squares, ridge regression stabilizes parameters estimation by adding a regularization term (the ridge penalty) to the loss function.

5.1. Normalization

To eliminate the discrepancies in feature dimensions and enhance the classifier's generalization ability, the feature vectors are normalized using Z-score. This involves calculating the mean (μ) and standard deviation (σ) of the feature vectors and standardizing them using the following formula:

$$Z = \frac{(X - \mu)}{\sigma} \quad (4)$$

where X is the original feature vector. This step ensures that all features contribute equally to the final classification decision, preventing disproportionate influences from features with larger numerical values. Normalization is particularly crucial for the calculation of the weight vector.

5.2. Loss Function and Training

Each classifier aims to minimize its corresponding loss function. Training involves minimizing the following loss function, which includes a regularization term:

$$L(W) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - W^T x_i\right)^2 + \alpha||\boldsymbol{W}||^2 \quad (5)$$

where N is the number of training samples, y_i is the actual label of sample i (indicating the presence or absence of a subtype), x_i is the feature vector of sample i, W is the weight vector of the model, and α is the regularization strength parameter. Adjusting α controls the degree of regularization.

In Ridge Regression, the analytical solution for the weight vector W is obtained by solving the derivative of the loss function and setting it to zero. This approach allows us to avoid the use of iterative methods:

$$W = (X^T X + \alpha I)^{-1} X^T y \quad (6)$$

where, X is the feature matrix, X^T is its transpose, y is the target value vector, and I is the identity matrix with the same dimensions as $X^T X$. Compared to standard linear regression, the addition of αI ensures that $(X^T X + \alpha I)$ is always invertible, allowing for a direct solution of the analytical solution. This is key to the computational efficiency of Ridge Regression.

Although Ridge Regression is primarily used for regression tasks, it can also be applied to classification by employing specific techniques. In binary classification situations, a threshold is used to determine class categories. During training, for each Ridge Regression classifier, positive samples of the corresponding subtype are assigned $y = 1$, and the

negative samples are assigned $y = 0$. Cross-validation is utilized to identify the optimal regularization parameter α.

5.3. Prediction

After training, the weight vector W can predict the feature vector of new samples. The predicted value $\hat{y}$ is calculated using:

$$\hat{y} = W^T x \tag{7}$$

where x is the feature vector of the new sample.

During prediction, the Ridge Regression model's output can be thresholded to convert continuous prediction values into discrete category labels, thereby completing the classification task. Specifically, for a new sample, each trained Ridge Regression classifier provides a prediction value for its corresponding HPV subtype, representing the probability that the sample belongs to that subtype. By setting a threshold (e.g., 0.5), if the prediction value is greater than or equal to the threshold, the sample is classified as positive for that subtype; otherwise, it is classified as negative.

Ultimately, after thresholding, the outputs of all classifiers are integrated to form a binary vector containing information about the sample's HPV subtype infection. For example, in a classification task involving m subtypes, the output is an m-dimensional vector, where each element corresponds to the detection result for a specific subtype (1 indicates positive, 0 indicates negative).

CRediT authorship contribution statement

Tianjun Zhou: Writing – original draft, Validation, Methodology, Investigation, Data curation. **Qi Zhao:** Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Guofan Hong:** Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Lin Li:** Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis. **Luonan Chen:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Acknowledgements

This study was supported by Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB38040400), National Natural Science Foundation of China (Nos T2341007, T2350003, 31930022, 12131020), National Key R&D Program of China (No. 2022YFA1004800), Science and Technology Commission of Shanghai Municipality (No.23JS1401300), and JST Moonshot R&D (No. JPMJMS2021).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.12.005.

Data Availability

The code and simulated dataset presented in the study are openly available in github at https://github.com/yysj-zq/MTIOT. Experimental and clinical data are also available upon the request.

References

[1] Sung H, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021; 71:209–49. https://doi.org/10.3322/caac.21660.
[2] Ferlay J., E M., Lam F., Laversanne M., Colombet M., Mery L., Piñeros M., Znaor A., Soerjomataram I., Bray F. Global cancer observatory: cancer today, <https://gco.iarc.who.int/today> (2024).
[3] World Health, O. WHO recommendations on self-care interventions: self-administration of injectable contraception. (World Health Organization, Geneva, 2022).
[4] World Health, O. Global strategy to accelerate the elimination of cervical cancer as a public health problem. (World Health Organization, 2020).
[5] Wardak S. Human Papillomavirus (HPV) and cervical cancer. Med Dosw Mikrobiol 2016;68:73–84.
[6] World Health Organization. WHO guideline for screening and treatment of cervical pre-cancer lesions for cervical cancer prevention. 2nd ed. World Health Organization; 2021. p. 97 (xvi).
[7] Walboomers JMM, et al. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. J Pathol 1999;189:12–9. https://doi.org/10.1002/(sici)1096-9896(199909)189:1< 12::Aid-path431> 3.0.Co;2-f.
[8] Bauer HM, et al. Genital human papillomavirus infection in female university students as determined by a PCR-based method. Jama 1991;265:472–7.
[9] Pe G, Mm M. Polymerase chain reaction-based methods for the detection of human papillomavirus DNA. IARC Sci Publ 1992;119:121–33.
[10] de Villiers EM. Cross-roads in the classification of papillomaviruses. Virology 2013; 445:2–10. https://doi.org/10.1016/j.virol.2013.04.023.
[11] Dokianakis DN, Sourvinos G, Sakkas S, Athanasiadou E, Spandidos DA. Detection of HPV and ras gene mutations in cervical smears from female genital lesions. Oncol Rep 1998;5:1195–8. https://doi.org/10.3892/or.5.5.1195.
[12] Sourvinos G, Rizos E, Spandidos DA. p53 Codon 72 polymorphism is linked to the development and not the progression of benign and malignant laryngeal tumours. Oral Oncol 2001;37:572–8. https://doi.org/10.1016/s1368-8375(00)00139-1.
[13] Mammas IN, Zafiropoulos A, Sifakis S, Sourvinos G, Spandidos DA. Human papillomavirus (HPV) typing in relation to ras oncogene mRNA expression in HPV-associated human squamous cervical neoplasia. Int J Biol Markers 2005;20: 257–63. https://doi.org/10.1177/172460080502000409.
[14] Mongia A, et al. Hybrid capture 2 and cobas® 4800: Comparison of performance of two clinically validated tests for human papillomavirus primary screening of cervical cancer. J Med Screen 2021;28:472–9. https://doi.org/10.1177/0969141321992820.
[15] Bhatla N, Singhal S. Primary HPV screening for cervical cancer. Best Pr Res Clin Obstet Gynaecol 2020;65:98–108. https://doi.org/10.1016/j.bpobgyn.2020.02.008.
[16] Simonetti S, Chen X, DiMauro S, Schon EA. Accumulation of deletions in human mitochondrial DNA during normal aging: analysis by quantitative PCR. Biochim Biophys Acta 1992;1180:113–22. https://doi.org/10.1016/0925-4439(92)90059-v.
[17] Veress G, Konya J, Csikymeszaros T, Czegledy J, Gergely L. Human papillomavirus DNA and anti-HPV secretory IgA antibodies in cytologically normal cervical specimens. J Med Virol 1994;43:201–7. https://doi.org/10.1002/jmv.1890430219.
[18] Chen WG, et al. Gene chip technology used in the detection of HPV infection in esophageal cancer of Kazakh Chinese in Xinjiang Province. J Huazhong Univ Sci Tech - Med 2014;34:343–7. https://doi.org/10.1007/s11596-014-1280-6.
[19] Lin M, et al. Genital human papillomavirus screening by gene chip in Chinese women of Guangdong province. Aust NZ J Obstet Gynaecol 2008;48:189–94. https://doi.org/10.1111/j.1479-828X.2008.00844.x.
[20] Han KH. Evaluation of human papillomavirus (HPV) genotyping assays using type-specific HPV L1 reference DNA. Genes Genom 2021;43:775–81. https://doi.org/10.1007/s13258-021-01100-4.
[21] Castle PE, et al. Restricted cross-reactivity of hybrid capture 2 with nononcogenic human papillomavirus types. Cancer Epidemiol Biomark Prev 2002;11:1394–9.
[22] Santos F, Invenção MCV, Araújo ED, Barros GS, Batista MVA. Comparative analysis of different PCR-based strategies for HPV detection and genotyping from cervical samples. J Med Virol 2021;93:6347–54. https://doi.org/10.1002/jmv.27118.
[23] Chang, J.S., Luo, Y.F., Su, K.Y. & Assoc Computat, L. in 30th Annual Meeting of the Assoc for Computational Linguistics. 177-184 (1992).
[24] Jain AK, Duin RPW, Mao JC. Statistical pattern recognition: a review. Ieee Trans Pattern Anal Mach Intell 2000;22:4–37. https://doi.org/10.1109/34.824819.
[25] Devijver, P.A. & Kittler, J.
[26] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Morgan Kaufmann Publishers Inc; 1995.
[27] Dempster A, Petitjean F, Webb G. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. Data Min Knowl Discov 2020;34:1454–95. https://doi.org/10.1007/s10618-020-00701-z.
[28] Dempster, A., Schmidt, D.F., Webb, G.I. & Assoc Comp, M. in 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). 248-257 (2021).
[29] Tan CW, Dempster A, Bergmeir C, Webb GI. MultiRocket: multiple pooling operators and transformations for fast and effective time series classification. Data Min Knowl Discov 2022;36:1623–46. https://doi.org/10.1007/s10618-022-00844-1.
[30] Dempster A, Schmidt DF, Webb GI. Hydra: competing convolutional kernels for fast and accurate time series classification. Data Min Knowl Discov 2023;37:1779–805. https://doi.org/10.1007/s10618-023-00939-3.
[31] Middlehurst M, et al. HIVE-COTE 2.0: a new meta ensemble for time series classification. Mach Learn 2021;110:3211–43. https://doi.org/10.1007/s10994-021-06057-9.
[32] *ROCKET*, <https://github.com/sktime/sktime/blob/df21a0c0275ebf28deb30efac5d469c9f0d178e3/sktime/transformations/panel/rocket/_rocket.py>
[33] *MINIROCKET*, <https://github.com/angus924/minirocket/blob/main/code/minirocket_multivariate.py>

[34] *MultiRocket*, <https://github.com/ChangWeiTan/MultiRocket/blob/main/multirocket/multirocket_multivariate.py>
[35] *Hydra*, <https://github.com/angus924/hydra/blob/main/code/hydra_multivariate.py>
[36] Ge S, et al. Prevent cervical cancer by screening with reliable human papillomavirus detection and genotyping. Cancer Med 2012;1:59–67. https://doi.org/10.1002/cam4.9.
[37] Lee SH, Vigliotti VS, Vigliotti JS, Pappu S. Validation of human papillomavirus genotyping by signature DNA sequence analysis. BMC Clin Pathol 2009;9:3. https://doi.org/10.1186/1472-6890-9-3.
[38] Ye SY, Hong GF. Heat-stable DNA polymerase I large fragment resolves hairpin structure in DNA sequencing. Sci Sin B 1987;30:503–6.