Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 13, No. 3, 2024

A Novel Feature Fusion Method Based on RSCU and K-mer to Classify the SARS-Cov-2

Fuqiang Ye[1], Jianhua Zhou[2], Hao Zhang[3], Lili Wang[1,*]

[1]School of Physics and Electrical Engineering, Northwest Normal University, Lanzhou 730070, China
[2]Biomedical Research Center, Northwest Minzu University, Lanzhou730010, China
[3]School of Life Science and Engineering, Northwest Minzu University Lanzhou, 730010, China
[*]Corresponding author: Lili Wang (wanglili20021984@126.com)

Abstract: The SARS-Cov-2 virus exhibits a high mutation rate, which makes the prediction and classification of its genetic evolution and variation trends highly significant. Accurate classification methods not only contribute to epidemiological studies of the virus, but also play a crucial role in vaccine development and antiviral drug discovery. This study aims to systematically evaluate the accuracy and generalization capability of the RSCU (Relative Synonymous Codon Usage) and K-mer encoding techniques in the classification of the SARS-CoV-2 genome. We extracted genomic data from two major SARS-CoV-2 variants, Alpha and Beta, and applied the Support Vector Machine (SVM) classification algorithm to train the data and assess the impact of different feature encoding methods on classification performance. Furthermore, we introduce a novel multi-feature fusion method, KRSCU, which combines the sequence position information from K-mer with the synonymous codon compositions from RSCU. This method effectively captures subtle differences in genomic data, significantly improving both the accuracy and generalization capability of the classification model. Experimental results demonstrate that the KRSCU method outperforms traditional single-feature encoding approaches in SARS-CoV-2 subtype classification tasks. Our research offers new insights into genomic data analysis, with potential applications in viral mutation monitoring.

Keywords: SARS-Cov-2, Support vector machine, K-mer, Relative Synonymous Codon Usage.

1. Introduction

Since the outbreak of the SARS-CoV-2 virus in 2019, it has rapidly spread globally, presenting significant challenges to public health, the economy, and society [1]. The high mutation rate and rapid transmission of the virus have made efficient and accurate identification and classification of the virus particularly crucial. Precise classification of the virus not only aids in tracing its origin, vaccine development, and antiviral drug design but also provides a scientific basis for formulating global prevention and control strategies. With the rapid advancement of bioinformatics technologies, the task of classifying nucleic acid sequences has become increasingly important in fields such as genomics, systems biology, and comparative genomics. To efficiently address this challenge, researchers have explored various feature representation methods to extract valuable information from nucleic acid sequences, which is then applied to classification tasks using machine learning algorithms [2]. These methods can reveal the virus's evolutionary trajectory, mutation patterns, and transmission characteristics, thus supporting scientific research and public health decision-making.

In pursuit of accurate and efficient nucleic acid classification, researchers continue to explore and optimize new encoding methods and machine learning algorithms [3]. In nucleic acid sequence analysis, feature extraction is a fundamental and critical step that directly impacts the accuracy of classification results. An ideal feature representation method should be capable of fully capturing the biological information within the sequence to enhance the performance of machine learning models in classification and prediction tasks. Over the past few decades, scholars have proposed various encoding schemes to achieve effective representation of nucleic acid sequences. Among these feature representation methods, RSCU [4-5] and K-mer [6-7] are widely used encoding techniques.

RSCU is an encoding method widely used in nucleic acid sequence analysis, first introduced by Sharp and Li in 1987 [8]. It aims to investigate the usage preference of synonymous codons in genes. By quantifying the frequency of codon usage, RSCU analysis can reveal underlying patterns related to gene expression regulation, genome evolution, and other aspects. RSCU analysis not only uncovers the codon usage bias in genes but also reflects differences in gene expression levels [9]. High-expression genes tend to preferentially use synonymous codons that are more frequently utilized in specific species or environments, a phenomenon known as codon bias. Based on this, RSCU has been widely applied in various fields, including the prediction of gene expression levels, the study of genome adaptive evolution, and the analysis of gene function [10].

The K-mer method is a widely used technique in nucleic acid sequence analysis, primarily aimed at uncovering sequence patterns and structural features within genomes [11]. A K-mer refers to a subsequence composed of K consecutive nucleotides, and its frequency distribution within the genome provides valuable insights into the genome's structure, function, and evolution [12]. The main advantages of the K-mer method lie in its efficiency, flexibility, and broad applicability, enabling it to handle large-scale genomic data and play a crucial role in various genomic studies [13]. By calculating the frequency of each K-mer in the genome, characteristic patterns of the genome can be identified, which are then utilized in gene prediction, sequence alignment, genome assembly, and gene expression analysis, among other research areas. In gene expression analysis, the K-mer method aids in identifying sequence features associated with specific gene expression by revealing enrichment patterns of certain

K-mers [14]. For example, changes in the frequency of specific K-mers may be linked to transcription factor binding sites [15], gene promoter regions [16], or regulatory elements of genes [17], thus providing clues to the underlying gene regulatory mechanisms [18].

In this study, we employed the SVM algorithm to investigate the performance of the RSCU and K-mer encoding methods in the classification of the COVID-19 virus. To further enhance the classification performance, we innovatively proposed a multi-feature fusion approach called KRSCU. The KRSCU method combines the compositional features of RSCU with the positional information of K-mers, aiming to leverage the strengths of both techniques for a more comprehensive representation of viral sequences. Specifically, the RSCU method focuses on the relative usage frequencies of codons in gene sequences, while the K-mer method captures local structural patterns and motifs by considering short sequence fragments. By integrating these two features, the KRSCU approach provides a richer and more accurate sequence representation, improving classification accuracy and robustness. Future research could explore the potential of this method in the classification of other viruses or biological entities.

2. Materials and Methods

2.1. Data

In this study, we downloaded 398 complete SARS-CoV-2 genomes from the National Center for Biotechnology Information (NCBI) GenBank (Table 1). To ensure that the data used were of high quality and representativeness, we implemented a rigorous selection strategy, which included the following steps:

Variant Strain Coverage: We selected multiple SARS-CoV-2 sequences from the SARS-CoV-2 Data Hub, including those from the Alpha and Beta variants. These variants represent different evolutionary lineages and encompass a wide range of viral characteristics [19].

Quality and Integrity Filtering: During the selection process, we applied strict criteria to ensure that the genome sequences had high quality and integrity. Sequences with obvious gaps, errors, or low-quality segments were excluded, preventing biases in the analysis due to sequence contamination or incompleteness [20].

Diversity and Representativeness: To enhance the generalizability and applicability of our findings, we ensured the inclusion of virus samples from diverse geographical locations, time points, and host sources. This broad sampling strategy helped capture various viral mutations and epidemic dynamics [21].

Data Balance: During data collection and filtering, we paid particular attention to maintaining a balanced representation of different SARS-CoV-2 variants. This was done to avoid the potential impact of class imbalance on subsequent analyses and classification tasks. Ensuring this balance helped maintain the robustness and accuracy of our classification models [22].

Table 1. The data of SARS-Cov-2

	training data	test data	sum
Alpha	140	59	199
Beta	140	59	199

Through this series of strict selection and filtering steps, we obtained 398 high-quality and complete SARS-CoV-2 genome sequences, providing a solid foundation for further analysis. These meticulously curated sequences will offer reliable data support for our classification research and viral mutation analysis.

2.2. RSCU

In this study, we employed the Relative Synonymous Codon Usage (RSCU) as an encoding method for feature extraction from nucleotide sequences. RSCU is a metric used to quantify codon usage bias, providing insights into the preference for certain synonymous codons over others within a nucleotide sequence[8].

Calculation of RSCU:

The RSCU is computed based on the frequency of synonymous codons for a specific amino acid. The calculation process is as follows:

Codon Frequency Calculation: For a given amino acid (e.g., alanine), the frequencies of all possible synonymous codons (e.g., GCU, GCC, GCA, GCG for alanine) are first recorded.

Average Frequency Calculation: For the synonymous codons of a particular amino acid, the average occurrence frequency is computed as:

$$A = \frac{\sum_i C_i}{N} \tag{1}$$

Where, C_i is the frequency of codon i, and N is the total number of synonymous codons for the amino acid.

RSCU Calculation: For each synonymous codon, the RSCU value is calculated using the formula:

$$RSCU_i = \frac{C_i}{A} \tag{2}$$

Where C_i is the count of codon i, and $A$is the average frequency of all synonymous codons for the amino acid.

If RSCU=1, the usage of codon i is equal to the average frequency of synonymous codons for that amino acid. If RSCU>1, codon i is used more frequently than expected, indicating a preference for that codon. If RSCU<1, codon i is used less frequently than expected, indicating not preference for that codon.

Application of RSCU in Nucleotide Sequence Analysis: By calculating RSCU values, we can gain insights into codon usage preferences within nucleotide sequences. RSCU encoding is a powerful tool for characterizing sequence features, revealing codon usage patterns in different genes or viral genomes. This method has wide applications in genomic classification, functional prediction, and gene expression analysis.

For instance, extracting RSCU features from the SARS-CoV-2 or other viral genomes allows for the identification of potential codon usage patterns and evolutionary trends, providing valuable input for subsequent machine learning

analyses.

2.3. K-mer

In this study, we employed the k-mer encoding method for feature extraction from nucleotide sequences. The k-mer approach is commonly used in bioinformatics to analyze sequences by dividing them into overlapping substrings of length k. Each k-mer serves as a fundamental unit for capturing sequence motifs and can be used to characterize sequence patterns, functional regions, or evolutionary trends.

The k-mer encoding method involves dividing a nucleotide sequence into overlapping subsequences of length k (where k is a positive integer). This approach captures sequence patterns that may be indicative of functional or structural characteristics. Below is the process of computing k-mer frequencies:

Generate k-mers: For a given nucleotide sequence, generate all possible contiguous subsequences of length k. For example, for a sequence "AGCTT", and k = 2, the possible 2-mers are:"AG", "GC", "CT", "TT".

Count the Frequency of Each k-mer: Next, we count how many times each unique k-mer appears in the given sequence. The frequency of each k-mer is recorded in a frequency vector.

For example, in the sequence "AGCTT", the 2-mers may have the following frequencies:

"AG": 1
"GC": 1
"CT": 1
"TT": 1

Normalize the k-mer Frequencies: To make the k-mer frequencies comparable across different sequences, we often normalize them. One common approach is to divide the frequency of each k-mer by the total number of possible k-mers in the sequence. This yields a normalized k-mer frequency that reflects the relative abundance of each k-mer within the sequence. Alternatively, other normalization techniques like z-score normalization or logarithmic transformation can be used.

Construct the k-mer Vector: The frequency or normalized frequency of each possible k-mer (from all possible nucleotide combinations of length k) is stored in a vector. This vector represents the sequence in terms of the presence or absence, as well as the frequency, of each k-mer. For example, for k = 2, the k-mer vector would contain the frequencies of all possible 2-mers (e.g., "AA", "AC", "AG", "AT", ..., "TT").

2.4. Support vector machine

In this study, we chose the Support Vector Machine (SVM) algorithm as the classification method to train and predict the SARS-CoV-2 sequences encoded [23]. SVM is a widely used supervised learning algorithm for pattern recognition, regression analysis, and classification tasks [24]. It is particularly effective at handling high-dimensional data and has demonstrated excellent performance in various practical applications [25].

The fundamental idea behind SVM is to find an optimal hyperplane in the feature space that maximally separates samples of different classes. To achieve this, SVM aims to identify a separating hyperplane that maximizes the margin, which is the distance between the hyperplane and the closest positive and negative sample points. This margin maximization enhances the model's generalization ability, resulting in higher prediction accuracy on unseen data.In SVM, the samples closest to the hyperplane are called support vectors. These support vectors play a crucial role in defining the position and orientation of the hyperplane. They are the most critical samples in the training process, as they directly influence the final placement of the separating hyperplane. By optimizing the objective function of SVM, the algorithm can effectively determine the optimal separating hyperplane while maximizing the margin between the two classes.

For the case of linearly separable data, SVM finds the optimal separating hyperplane by solving the following optimization problem, as shown in Equation (2-14)

$$\begin{aligned} &\min_{\omega,b} \tfrac{1}{2}\|\omega\|^2 \\ &\text{s.t. } y_i(\omega^T x_i + b) \geq 1, \ i=1,2,\cdots,m. \end{aligned} \tag{3}$$

Where, ω and b represent the normal vector and the intercept of the hyperplane, x_i denotes the feature vector, y_i is the class label (with values of +1 or -1), and m is the number of samples.

However, in real-world problems, data is often not linearly separable. To address this issue, SVM introduces the concepts of a soft margin and kernel functions. The soft margin allows some data points to be misclassified in order to achieve better generalization performance. The kernel function, on the other hand, maps the original feature space into a higher-dimensional space where the data may become linearly separable.

In this study, the SVM algorithm uses a linear kernel as the kernel function, which is expressed as follows:

$$K(x_i, x_j) = x_i^T x_j \tag{4}$$

SVM is a powerful classifier, suitable for handling both linear and nonlinear problems. By applying the SVM algorithm, we can make accurate classification predictions for COVID-19 virus sequences.

2.5. Performance assessment

In machine learning, performance evaluation metrics are crucial as they help us gain a comprehensive understanding of the strengths and weaknesses of a model. Commonly used evaluation metrics include sensitivity, specificity, accuracy, F1 score, and ROC curve [26-28].

$$\begin{cases} \mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \\ \mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \\ \mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \\ \mathrm{ACC} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FN+FP}} \\ \mathrm{F1} = \frac{2\times\mathrm{Precision}\times\mathrm{Sn}}{\mathrm{Precision+Sn}} \\ \mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP+FN})(\mathrm{TP+FP})(\mathrm{TN+FP})(\mathrm{TN+FN})}} \end{cases} \tag{5}$$

3. Results and Analysis

3.1. Binary classification of viruses

In this study, we randomly divided a total of 398 samples into a training set (280 samples, 70% of the total) and a test set (118 samples, 30% of the total), as shown in Table 1. To evaluate the performance of the classifiers in the binary classification task for COVID-19, we compared the RSCU and K-mer methods. Both classifiers demonstrated high recognition accuracy, exceeding 82% on the test set (Table 2).

Table 2. Comparison of the performance among RSCU and K-mer.

method	categories	accuracy	precision	recall	f1	mcc
RSCU	Alpha	0.823	0.823	0.829	0.824	0.649
	Beta		0.828	0.818	0.822	
K-mer	Alpha	0.859	0.843	0.884	0.861	0.723
	Beta		0.884	0.834	0.856	

Further analysis revealed that, although both classifiers performed well in terms of accuracy, the K-mer method outperformed RSCU. Specifically, the accuracy of the K-mer method was 85.9%, significantly higher than the 82.3% achieved by the RSCU method. This result indicates that the K-mer method is more effective in feature extraction, capturing important patterns in the COVID-19 samples.

Based on the evaluation metrics, including precision, recall, F1 score, and MCC, the K-mer method consistently outperformed the RSCU method across all indicators. The precision of the K-mer method was higher than that of RSCU in classifying both Alpha and Beta samples, with the precision for Beta samples being 88.4% compared to 82.8% for RSCU. In terms of recall, the K-mer method also exhibited superior performance, achieving a recall of 88.4% for Alpha samples, while RSCU achieved 82.9%.

The F1 score, as a comprehensive metric that balances precision and recall, provides a better reflection of classifier performance on imbalanced datasets. For Alpha samples, the F1 score of the K-mer method was 0.861, noticeably higher than the RSCU's 0.824. This demonstrates that the K-mer method is better at balancing precision and recall when dealing with imbalanced data, reducing classification errors.

Finally, the MCC was also used to assess the overall performance of the classifiers. Among all evaluation metrics, the K-mer method achieved an MCC of 0.723, which is considerably higher than the RSCU's 0.649. This further confirms the superiority of the K-mer method for COVID-19 classification tasks.

In summary, the K-mer encoding method has significant advantages in nucleic acid classification tasks based on above analysis and Figure1-4. However, it is important to note that the computational complexity of the K-mer encoding method is relatively high, which may increase processing time. Therefore, in practical applications, it is necessary to balance the trade-off between computational complexity and classification performance.

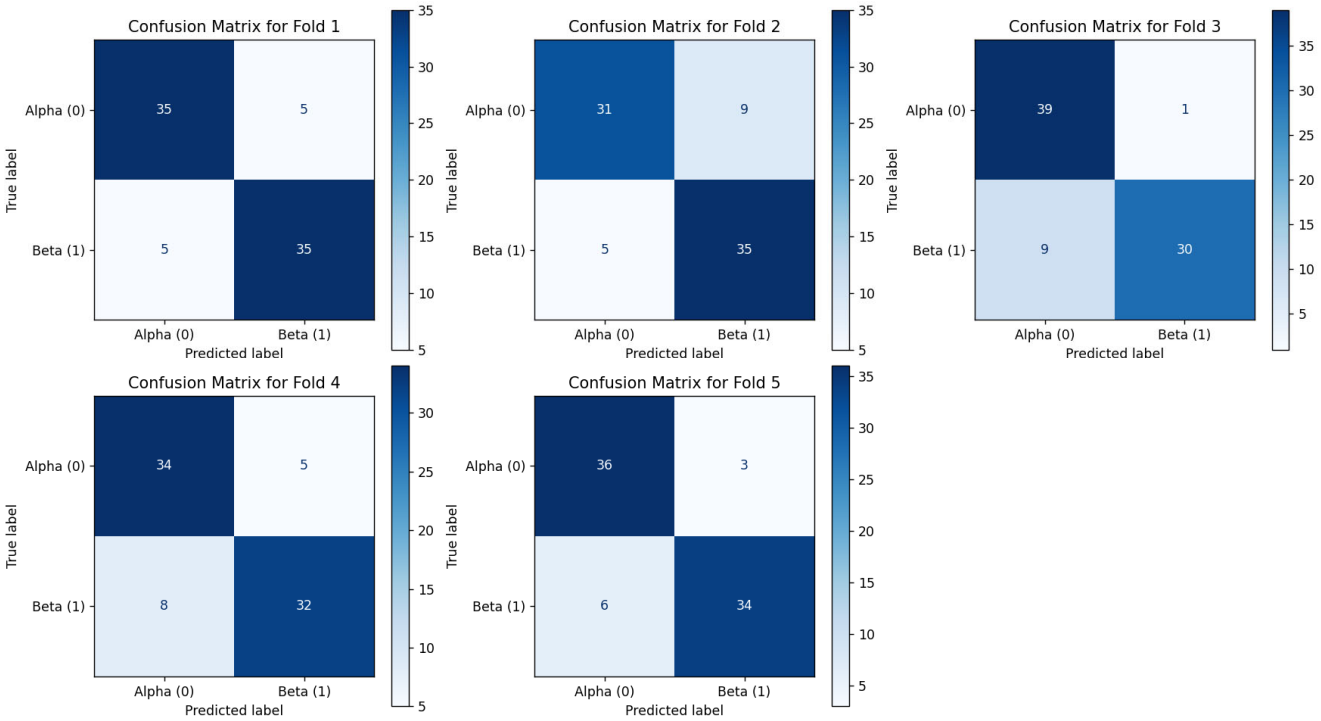

Figure 1. The confusion matrix of each fold in binary classification of K-mer model.

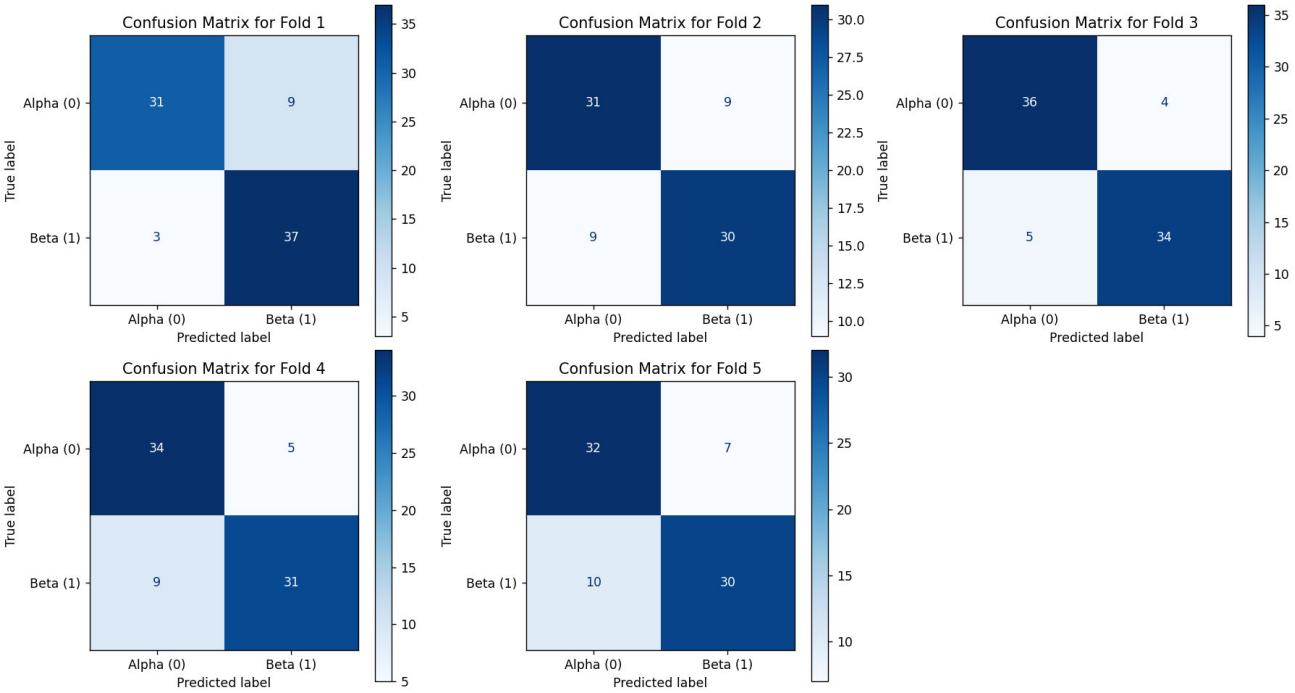

Figure 2. The confusion matrix of each fold in binary classification of RSCU model.

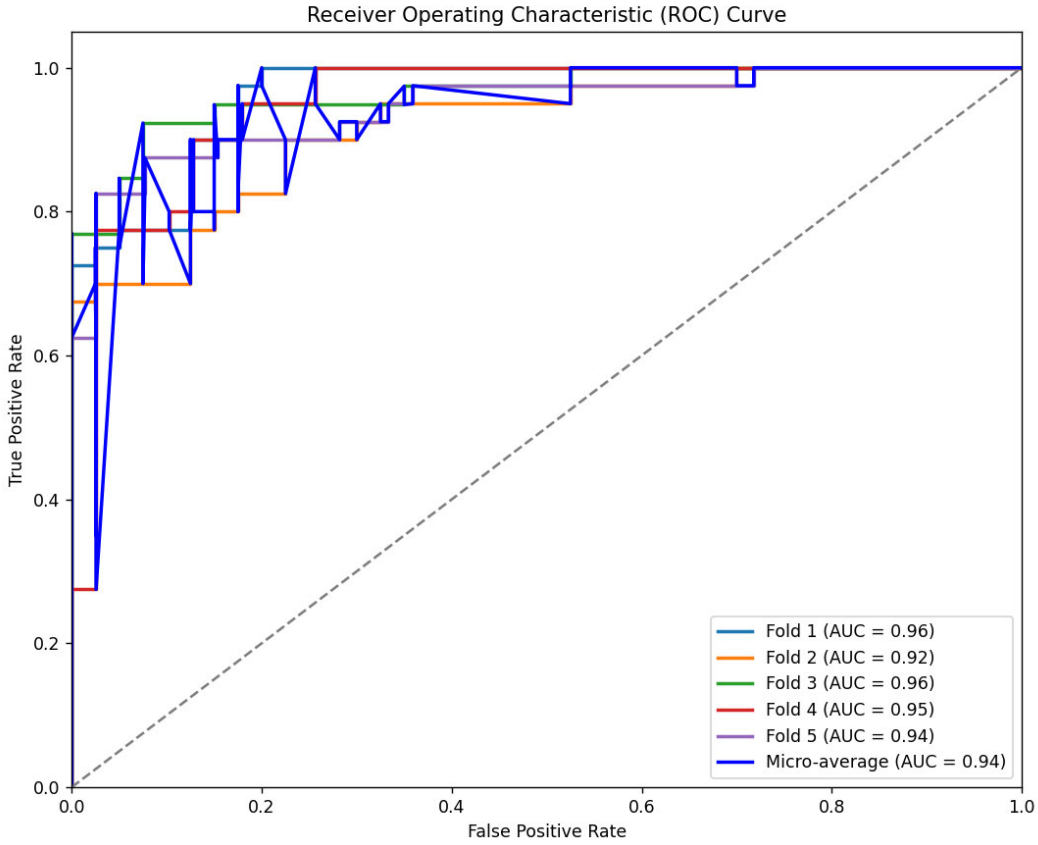

Figure 3. The ROC in binary classification of K-mer model.

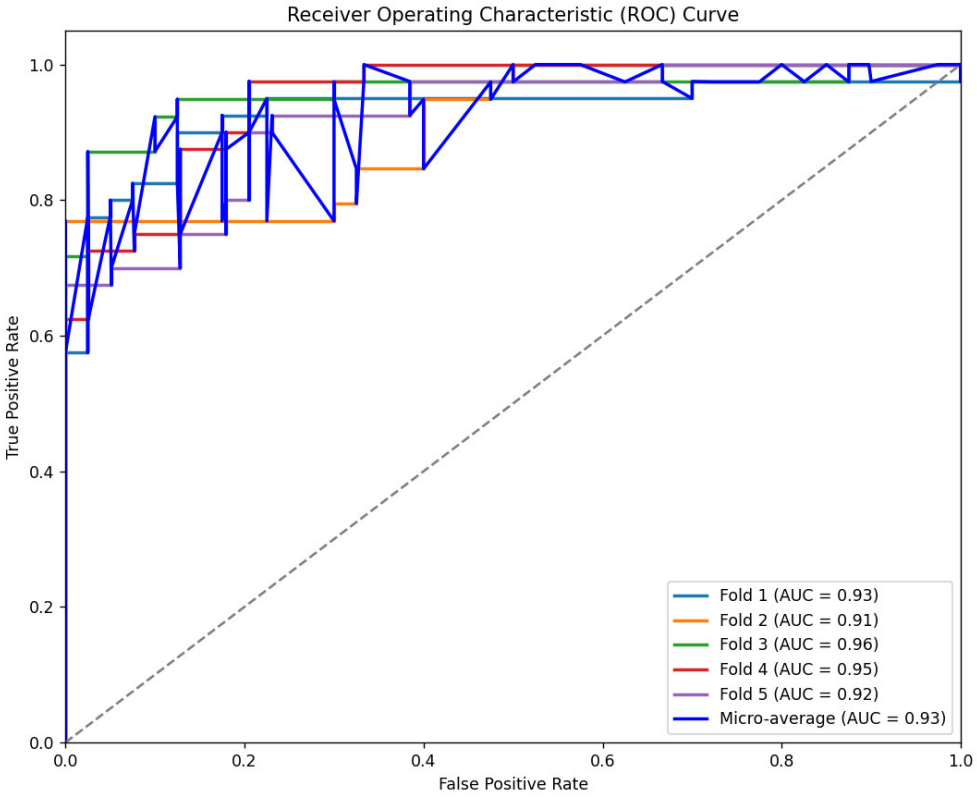

Figure 4. The ROC in binary classification of RSCU model.

3.2. KRSCU

The RSCU-based encoding method mainly focuses on compositional features, while the K-mer encoding method emphasizes positional features within the sequence. Given the respective advantages and limitations of these two encoding methods, we propose an innovative multi-feature fusion technique called KRSCU, which aims to improve classification performance and enhance the model's generalization ability. This method combines the compositional features of RSCU encoding with the positional information of K-mer encoding to create a composite feature representation, which effectively captures both the global compositional characteristics and local structural patterns of the sequence.

The specific formula for calculating KRSCU is as follows:

$$\vec{C} = [\vec{R}; \vec{K}] \tag{6}$$

Where $\vec{C}$ represent the feature vector of KRSCU, where$\vec{R}$and $\vec{K}$ represent the feature vector of RSCU and K-mer, respectively.

Table 3. Comparison of the performance among KRSCU, RSCU and K-mer.

method	categories	Accuracy	Precision	Recall	F1	MCC
RSCU	Alpha	0.823	0.823	0.829	0.824	0.649
	Beta		0.828	0.818	0.822	
K-mer	Alpha	0.859	0.843	0.884	0.861	0.723
	Beta		0.884	0.834	0.856	
KRSCU	Alpha	0.876	0.889	0.859	0.873	0.753
	Beta		0.865	0.894	0.879	

The KRSCU method achieved the highest accuracy of 0.876 in both the Alpha and Beta categories, outperforming both the K-mer method (Alpha: 0.859, Beta: 0.859) and the RSCU method (Alpha: 0.823, Beta: 0.823). These results indicate that KRSCU performs most consistently and accurately across the classification task.

In the Alpha category, KRSCU achieved the highest precision of 0.889, which was higher than the K-mer method (Alpha: 0.843) and RSCU (Alpha: 0.823). In the Beta category, K-mer exhibited the highest precision (0.884), slightly outperforming RSCU (0.828), but still lower than KRSCU (0.865). These results suggest that KRSCU is the most precise in predicting positive samples, particularly in the Alpha category.

In the Beta category, KRSCU achieved the highest recall of 0.894, significantly outperforming the other two methods. In contrast, K-mer demonstrated the highest recall in the Alpha category (0.884), compared to RSCU (0.829) and KRSCU (0.859), indicating that K-mer is better at detecting positive samples, particularly in the Alpha category.

In the Alpha category, K-mer achieved the highest F1 score (0.861), outperforming RSCU (0.824) and KRSCU (0.873). However, in the Beta category, KRSCU achieved the highest

F1 score (0.879), slightly better than K-mer (0.856) and RSCU (0.822), suggesting that KRSCU provides the best balance between precision and recall in the Beta category.

The KRSCU method consistently achieved the highest MCC of 0.753 in both the Alpha and Beta categories, outperforming both K-mer (0.723) and RSCU (Alpha: 0.649, Beta: 0.649). This indicates that KRSCU offers the most robust and well-rounded performance across all metrics.

In summary, the KRSCU method outperformed the K-mer and RSCU methodsin terms of accuracy, precision, recall, F1 score and MCC, the confusion matrix and ROC (Figure 1-6), with particularly strong performance in the Beta category. While the K-mer method performed better in recall and F1 score for Alpha samples, KRSCU showed superior overall performance. These results suggest that KRSCU is the most promising method for classification tasks, particularly in scenarios with imbalanced or complex datasets.

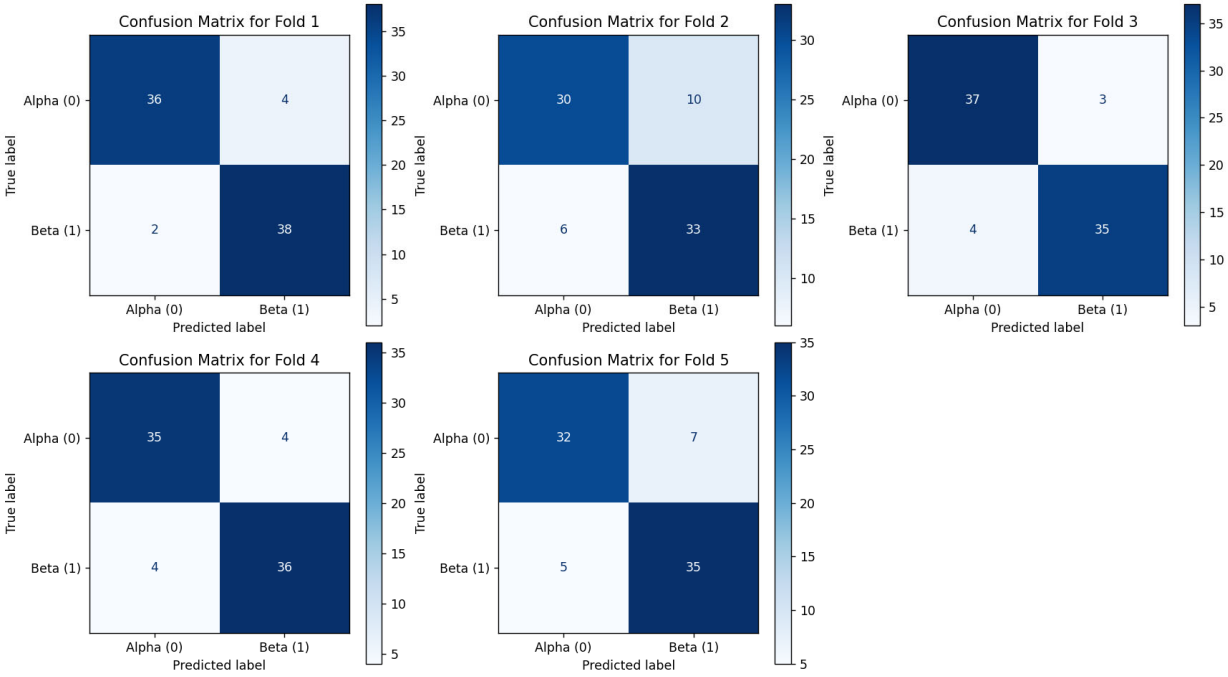

Figure 5. The confusion matrix of each fold in binary classification of KRSCU model.

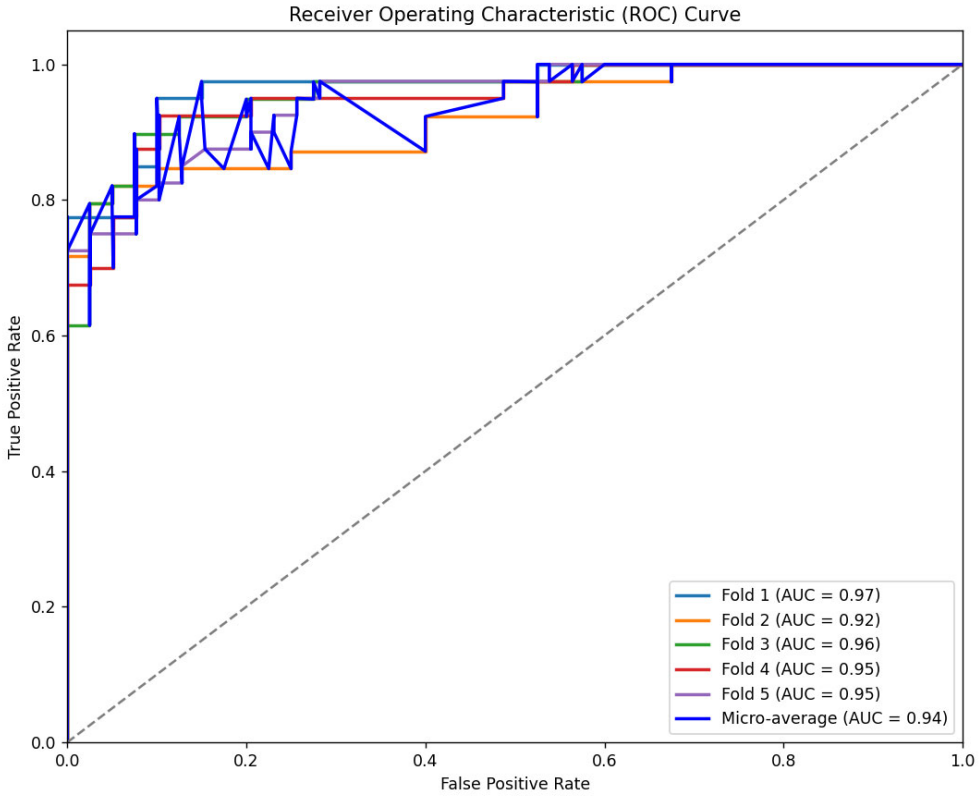

Figure 6. The ROC in binary classification of RSCU model.

4. Conclusions

In this study, we employed three feature extraction methods based onSVM—RSCU, K-mer, and KRSCU—to conduct a detailed evaluation of the classification capabilities for two SARS-CoV-2 subtypes. Through comparative experiments, the results show that the KRSCU method significantly outperforms the individual RSCU and K-mer encoding methods in classification performance. Specifically, the KRSCU method, by combining the compositional features of RSCU encoding with the positional information of K-mer encoding, provides a more comprehensive sequence representation for classification tasks, thereby enhancing the model's ability to distinguish between different subtypes and improving robustness. Compared to the RSCU method, KRSCU effectively overcomes the limitations of using only compositional features, while the K-mer method excels in capturing local sequence patterns but neglects global compositional characteristics. By integrating these two features, the KRSCU method considers both global and local information from the sequence, improving classification accuracy and demonstrating superior adaptability to highly variable data. Our experimental results validate the advantages of the KRSCU method in SARS-CoV-2 subtype classification, offering a more effective feature extraction solution for future genomic classification tasks of similar viruses.

Future research could further explore how to translate the structural information within viral genomes into numerical features that can be utilized by machine learning models, thereby enhancing the accuracy and complexity of classification models. This could be achieved by developing more refined feature extraction techniques, particularly those targeting secondary structures, tertiary structures, and other potential structural elements within sequences, such as the stability of RNA secondary structures or protein-protein interaction networks. The full utilization of this structural information will help capture the deeper relationships between viral mutations and their biological functions, thus improving the performance of classification and prediction models.

References

[1] Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., ... & Shi, Z. L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. nature, 579(7798), 270-273.

[2] Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., ... & Montefiori, D. C. (2020). Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell, 182(4), 812-827.

[3] Zhang, J., & Liu, B. (2019). A review on the recent developments of sequence-based protein feature extraction methods. Current Bioinformatics, 14(3), 190-199.

[4] Bzhalava, Z., Tampuu, A., Bała, P., Vicente, R., & Dillner, J. (2018). Machine Learning for detection of viral sequences in human metagenomic datasets. BMC bioinformatics, 19, 1-11.

[5] Song, W., Ji, C., Chen, Z., Cai, H., Wu, X., Shi, C., & Wang, S. (2022). Comparative analysis the complete chloroplast genomes of nine Musa species: genomic features, comparative analysis, and phylogenetic implications. Frontiers in Plant Science, 13, 832884.

[6] Oh, J. W., & Beer, M. A. (2024). Gapped-kmer sequence modeling robustly identifies regulatory vocabularies and distal enhancers conserved between evolutionarily distant mammals. Nature communications, 15(1), 6464.

[7] Kaniwa, F. (2018). A kmer-based parallel algorithm for pattern searching in DNA sequences on shared-memory model (Doctoral dissertation, Botswana International University of Science & Technology (Botswana)).

[8] Sharp, P. M., & Li, W. H. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic acids research, 15(3), 1281-1295.

[9] Novoa, E. M., & de Pouplana, L. R. (2012). Speeding with control: codon usage, tRNAs, and ribosomes. Trends in Genetics, 28(11), 574-581.

[10] Khandia, R., Gurjar, P., Kamal, M. A., & Greig, N. H. (2024). Relative synonymous codon usage and codon pair analysis of depression associated genes. Scientific Reports, 14(1), 3502

[11] Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic acids research, 18(20), 6097-6100.

[12] He, C., Washburn, J. D., Schleif, N., Hao, Y., Kaeppler, H., Kaeppler, S. M., ... & Liu, S. (2024). Trait association and prediction through integrative k-mer analysis. The Plant Journal, 120(2), 833-850.

[13] Moeckel, C., Mareboina, M., Konnaris, M. A., Chan, C. S., Mouratidis, I., Montgomery, A., ... & Georgakopoulos-Soares, I. (2024). A survey of k-mer methods and applications in bioinformatics. Computational and Structural Biotechnology Journal.

[14] Van Etten, J., Stephens, T. G., & Bhattacharya, D. (2023). A k-mer-based approach for phylogenetic classification of taxa in environmental genomic data. Systematic biology, 72(5), 1101-1118.

[15] Tahara, S., Tsuchiya, T., Matsumoto, H., & Ozaki, H. (2023). Transcription factor-binding k-mer analysis clarifies the cell type dependency of binding specificities and cis-regulatory SNPs in humans. BMC genomics, 24(1), 597.

[16] Gupta, S., & Shankar, R. (2024). Comprehensive analysis of computational approaches in plant transcription factors binding regions discovery. Heliyon, 10(20).

[17] Nguyen, E., Poli, M., Faizi, M., Thomas, A., Wornow, M., Birch-Sykes, C., ... & Baccus, S. (2024). Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. Advances in neural information processing systems, 36.

[18] Sievers, A., Bosiek, K., Bisch, M., Dreessen, C., Riedel, J., Froß, P., ... & Hildenbrand, G. (2017). K-mer content, correlation, and position analysis of genome DNA sequences for the identification of function and evolutionary features. Genes, 8(4), 122.

[19] Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., ... & Neher, R. A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics, 34(23), 4121-4123.

[20] Endrullat, C., Glökler, J., Franke, P., & Frohme, M. (2016). Standardization and quality management in next-generation sequencing. Applied & translational genomics, 10, 2-9.

[21] Giovanetti, M., Slavov, S. N., Fonseca, V., Wilkinson, E., Tegally, H., Patané, J. S. L., ... & Covas, D. T. (2022). Genomic epidemiology of the SARS-CoV-2 epidemic in Brazil. Nature Microbiology, 7(9), 1490-1500.

[22] Beerenwinkel, N., Günthard, H. F., Roth, V., & Metzner, K. J. (2012). Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. Frontiers in microbiology, 3, 329.

[23] Cortes, C. (1995). Support-Vector Networks. Machine Learning.

[24] Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC bioinformatics, 7, 1-8.

[25] Griffel, L. M., Delparte, D., & Edwards, J. (2018). Using Support Vector Machines classification to differentiate spectral signatures of potato plants infected with Potato Virus Y. Computers and electronics in agriculture, 153, 318-324.

[26] Busia, A., Dahl, G. E., Fannjiang, C., Alexander, D. H., Dorfman, E., Poplin, R., ... & DePristo, M. (2018). A deep learning approach to pattern recognition for short DNA sequences. BioRxiv, 353474.

[27] Gunasekaran, H., Ramalakshmi, K., Rex Macedo Arokiaraj, A., Deepa Kanmani, S., Venkatesan, C., & Suresh Gnana Dhas, C. (2021). Analysis of DNA sequence classification using CNN and hybrid models. Computational and Mathematical Methods in Medicine, 2021(1), 1835056.

[28] Mandal, I. (2015). A novel approach for predicting DNA splice junctions using hybrid machine learning algorithms. Soft Computing, 19, 3431-3444.