International Journal of Artificial Intelligence in Medical Issues

Volume 4 Issue 1
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v4i1.438

Research Article

Machine Learning-Based Clustering of Viruses Using Taxonomic and Genomic Features for Health Informatics Applications

Adityo Permana Wibowo [1,*]; Made Leo Radhitya [2]; Edi Faizal [3]; Ika Arfiani [4]
[1] Universitas Teknologi Yogyakarta, Daerah Istimewa Yogyakarta 55285, Indonesia, adityopw@uty.ac.id
[2] Institut Bisnis dan Teknologi Indonesia, Kota Denpasar, Bali 80225, Indonesia, leo_radhitya@instiki.ac.id
[3] Universitas Teknologi Digital Indonesia, Kabupaten Bantul, Daerah Istimewa Yogyakarta 55198, Indonesia, edifaizal@utdi.ac.id
[4] Universitas Ahmad Dahlan, Kota Yogyakarta, Daerah Istimewa Yogyakarta 55166, Indonesia, ika.arfiani@tif.uad.ac.id

Correspondence should be addressed to Adityo Permana Wibowo; adityopw@uty.ac.id

Received 02 January 2026; Revised 06 January 2026; Accepted 25 April 2026; Published 30 May 2026

Abstract:
Viruses remain a major concern in global public health due to their potential to cause outbreaks, epidemics, and pandemics. The rapid organization and analysis of virus-related data are important for supporting computational virology, health informatics, and pandemic preparedness. This study proposes an unsupervised machine learning approach to cluster viruses based on taxonomic and genomic characteristics. The dataset consisted of 70 virus records with attributes including family, genus, genome type, strand type, and envelope status. Since the dataset did not contain predefined epidemiological labels or risk categories, the analysis was designed as an exploratory clustering task rather than a supervised prediction task. Data preprocessing was performed by removing duplicates, handling missing values, standardizing categorical attributes, and transforming selected features using One-Hot Encoding. Three clustering algorithms were evaluated, namely K-Means, Agglomerative Clustering, and DBSCAN. The clustering performance was assessed using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score, while Principal Component Analysis was applied for two-dimensional visualization. The results showed that K-Means with 10 clusters achieved a Silhouette Score of 0.7725 and a Davies-Bouldin Index of 0.8186. Agglomerative Clustering obtained the highest Silhouette Score of 0.7754, while DBSCAN produced fewer clusters with lower overall performance. Several biologically meaningful groups were identified, including clusters representing Flaviviridae, Coronaviridae, Herpesviridae, Poxviridae, and enveloped RNA viruses. However, a large proportion of records contained unknown values, which influenced the formation of a dominant incomplete-data cluster. These findings indicate that taxonomic and genomic features can support machine learning-based virus grouping, although data completeness remains a critical factor. This study provides an initial computational framework for AI-driven viral data exploration and may serve as a foundation for future viral risk stratification using enriched epidemiological and clinical features.

Keywords: Virus Clustering, Machine Learning, Taxonomic Features, Genomic Features, Health Informatics, Computational Virology, Pandemic Preparedness.

Dataset link: -

1. Introduction

Viruses remain one of the most important biological agents in global public health because of their ability to cause outbreaks, epidemics, and pandemics. The COVID-19 pandemic has demonstrated that rapid identification, classification, and monitoring of viral pathogens are essential for strengthening disease surveillance and public health preparedness [1]. In this context, genomic surveillance has become a critical component of modern infectious disease

control, as it enables researchers and health authorities to monitor pathogen evolution, detect emerging variants, and support timely public health responses. The World Health Organization has emphasized the need to strengthen genomic surveillance for pathogens with epidemic and pandemic potential as part of global preparedness strategies.

Virus classification plays an important role in understanding biological relationships among viruses and provides a scientific basis for organizing viral diversity [2]. According to the International Committee on Taxonomy of Viruses, virus classification is based on several characteristics, including genome composition, capsid structure, envelope status, gene expression strategy, host range, pathogenicity, and sequence similarity [3]. These taxonomic and genomic characteristics are useful not only for biological classification but also for computational analysis, particularly in bioinformatics and health informatics applications [4].

The availability of structured biological databases has created new opportunities for applying artificial intelligence and machine learning to virology-related problems [5]. The National Center for Biotechnology Information taxonomy database provides organized taxonomic information for viruses and other organisms, supporting computational analysis and data-driven biological research [6]. In recent years, machine learning methods have been increasingly used for viral classification, especially through genome-based, alignment-free, and feature-based approaches. Previous studies have shown that encoded genomic features and machine learning models can support virus taxonomy classification and reduce the computational burden associated with conventional alignment-based methods [7].

Despite these advances, many viral datasets remain incomplete or are not directly prepared for predictive modelling [8]. In many cases, available datasets contain basic taxonomic and genomic information but lack epidemiological labels such as mortality rate, transmission mode, reproductive number, or risk category. This limitation makes supervised disease severity prediction difficult. However, unsupervised machine learning can still be used to explore hidden patterns among viruses based on their biological and genomic characteristics [9]. Clustering analysis can help identify groups of viruses with similar taxonomic or genomic profiles, which may provide an initial foundation for further risk profiling and pandemic preparedness research [10].

Therefore, this study proposes a machine learning-based exploratory analysis of viruses using taxonomic and genomic features. The dataset used in this study contains viral attributes such as family, genus, genome type, strand type, and envelope status [11], [12]. Several unsupervised learning algorithms, including K-Means, Agglomerative Clustering, and DBSCAN, are applied to group viruses based on feature similarity. Dimensionality reduction using Principal Component Analysis is also used to visualize the clustering structure. The quality of the clustering results is evaluated using internal validation metrics such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score.

The main contribution of this study is the development of an initial machine learning framework for viral grouping based on taxonomic and genomic characteristics. Although the current dataset does not yet include complete epidemiological variables, the proposed framework can serve as a foundation for future viral risk stratification models by integrating additional features such as transmission mode, mortality rate, basic reproduction number, severity index, and risk category. This study is expected to contribute to AI-driven health informatics and computational

virology by demonstrating how machine learning can be used to support the organization, exploration, and interpretation of virus-related datasets [13].

2. Method

Research Design:

This study employed an exploratory machine learning approach to analyze virus data based on taxonomic and genomic characteristics. Since the dataset did not contain predefined epidemiological labels such as mortality rate, transmission mode, reproductive number, severity index, or risk category, the analysis was designed as an unsupervised learning task. The main objective was to identify potential grouping patterns among viruses using clustering algorithms.

The research workflow consisted of several stages, including dataset acquisition, data preprocessing, categorical feature encoding, clustering model development, cluster evaluation, dimensionality reduction, and result interpretation. The overall methodological framework was designed to provide an initial computational basis for virus grouping and future viral risk stratification studies [14].

Dataset Description:

The dataset used in this study was a structured tabular dataset containing information on viruses. The available attributes represented basic taxonomic and genomic characteristics of each virus. The dataset consisted of 70 virus records and 7 variables. The variables included virus identification, virus name, taxonomic family, genus, genome type, strand type, and envelope status [15].

The description of the dataset attributes is presented in **Table 1**.

Table 1. Dataset Attributes

No.	Attribute	Description
1	`virus_id`	Unique identifier for each virus
2	`name`	Name of the virus
3	`family`	Taxonomic family of the virus
4	`genus`	Taxonomic genus of the virus
5	`genome_type`	Type of viral genome
6	`strand`	Strand characteristic of the viral genome
7	`enveloped`	Envelope status of the virus

The attributes `virus_id` and `name` were used only as identifiers and were not included as input features in the machine learning process. The features used for clustering were `family`, `genus`, `genome_type`, `strand`, and `enveloped`.

Data Preprocessing:

Data preprocessing was conducted to ensure that the dataset was suitable for machine learning analysis. First, duplicate records were checked and removed when found. Column names were standardized by converting them into lowercase format and replacing spaces with underscores. Missing values in categorical attributes were handled by replacing them with the value `Unknown`. This strategy was applied because several virus records contained incomplete taxonomic or genomic information.

All categorical variables were then converted into string format to ensure compatibility with the encoding process. Since the dataset consisted primarily of categorical variables, no numerical normalization was required at this stage. The preprocessing stage aimed to produce a clean and consistent dataset for subsequent feature transformation and clustering analysis.

Feature Selection and Encoding:

The selected input features consisted of taxonomic and genomic attributes, namely `family`, `genus`, `genome_type`, `strand`, and `enveloped`. These features were selected because they describe the biological characteristics of viruses and may represent meaningful similarities among viral entities [16].

Because all selected features were categorical, One-Hot Encoding was applied to transform each categorical value into a numerical representation. This encoding method converts each category into a binary vector, allowing machine learning algorithms to process non-numerical data [17]. The encoded feature matrix was then used as the input for clustering models.

The feature transformation process can be summarized as follows:

Table 2. Feature Preparation Process

Stage	Description
Identifier removal	virus_id and name were excluded from the input features
Feature selection	Taxonomic and genomic attributes were selected
Missing value handling	Missing values were replaced with Unknown
Data type conversion	All categorical variables were converted into string format
Feature encoding	One-Hot Encoding was applied to categorical features

Clustering Algorithms:

Three unsupervised machine learning algorithms were used in this study: K-Means, Agglomerative Clustering, and DBSCAN. These algorithms were selected to compare different clustering mechanisms.

K-Means was used as a centroid-based clustering method that partitions data into a predefined number of clusters. Agglomerative Clustering was used as a hierarchical clustering approach that builds clusters by progressively merging

similar data points. DBSCAN was used as a density-based clustering method capable of identifying dense regions and noise points [18].

The algorithms used in this study are summarized in **Table 3**.

Table 3. Clustering Algorithms Used in This Study

Algorithm	Type	Purpose
K-Means	Centroid-based clustering	Groups viruses based on similarity to cluster centroids
Agglomerative Clustering	Hierarchical clustering	Groups viruses based on hierarchical similarity structure
DBSCAN	Density-based clustering	Identifies dense clusters and potential outlier viruses

K-Means was used as the primary clustering model because it is widely used, computationally efficient, and suitable for exploratory grouping tasks after categorical features have been encoded into numerical form [19].

Determination of the Optimal Number of Clusters:

For the K-Means algorithm, the optimal number of clusters was determined by testing several values of k. The number of clusters was evaluated using internal clustering validation metrics, particularly the Silhouette Score. The value of k with the highest Silhouette Score was selected as the optimal number of clusters.

The range of k values was determined based on the dataset size. Each clustering result was evaluated using three internal validation metrics: Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score.

Cluster Evaluation Metrics:

The clustering performance was evaluated using internal validation metrics [20]. These metrics were selected because the dataset did not contain ground-truth cluster labels.

The first metric was the Silhouette Score, which measures how similar an object is to its own cluster compared with other clusters. A higher Silhouette Score indicates better cluster separation. The second metric was the Davies-Bouldin Index, which measures the average similarity between clusters. A lower Davies-Bouldin Index indicates better clustering quality. The third metric was the Calinski-Harabasz Score, which evaluates the ratio between inter-cluster dispersion and intra-cluster dispersion. A higher Calinski-Harabasz Score indicates more compact and well-separated clusters.

The evaluation metrics are summarized in **Table 4**.

Table 4. Clustering Evaluation Metrics

Metric	Interpretation
Silhouette Score	Higher value indicates better cluster separation
Davies-Bouldin Index	Lower value indicates better clustering structure

Metric	Interpretation
Calinski-Harabasz Score	Higher value indicates better-defined clusters

Dimensionality Reduction and Visualization:

Principal Component Analysis was applied to reduce the high-dimensional encoded feature space into two principal components. This step was conducted to visualize the clustering results in a two-dimensional space. The PCA visualization was used to observe the distribution of viruses across clusters and to support the interpretation of the clustering structure.

Although PCA visualization does not replace quantitative evaluation, it provides an intuitive representation of the grouping pattern produced by the clustering model. The first two principal components were used as the horizontal and vertical axes in the visualization.

Cluster Profiling:

After clustering was completed, each virus was assigned a cluster label. Cluster profiling was then conducted to identify the dominant characteristics within each cluster. For each cluster, the most frequent values of `family`, `genus`, `genome_type`, `strand`, and `enveloped` were extracted. This profiling process was used to interpret the biological characteristics of each cluster [21].

The purpose of cluster profiling was to determine whether viruses within the same cluster shared similar taxonomic and genomic patterns [22]. The resulting profiles were used to support the discussion of how machine learning can organize virus data based on biological similarity.

Experimental Environment:

The data analysis and machine learning workflow were implemented using Python. The main libraries used in this study included Pandas and NumPy for data manipulation, Scikit-learn for preprocessing, clustering, evaluation, and dimensionality reduction, and Matplotlib for data visualization. The output of the analysis included cleaned data, encoded features, clustering evaluation results, cluster profiles, PCA visualization, and exported result files in CSV and Excel formats.

Research Workflow:

The overall research workflow is described as follows. First, the virus dataset was loaded and inspected to identify its structure, variables, missing values, and duplicate records. Second, data preprocessing was performed to standardize column names, handle missing values, remove duplicates, and prepare categorical variables. Third, relevant taxonomic and genomic features were selected and transformed using One-Hot Encoding. Fourth, clustering algorithms were applied to group viruses based on feature similarity. Fifth, the clustering results were evaluated using internal validation metrics. Sixth, PCA was used to visualize the clustering pattern. Finally, cluster profiling was conducted to interpret the dominant characteristics of each virus group.

Methodological Limitation:

This study has a methodological limitation related to the available dataset attributes. The dataset used in this study contains taxonomic and genomic characteristics but does not include epidemiological variables such as transmission mode, mortality rate, reproductive number, severity index, or risk category. Therefore, this study focuses on exploratory virus clustering rather than supervised disease severity prediction or pandemic risk classification.

Future studies may extend this framework by integrating epidemiological and clinical variables. The addition of variables such as transmission mode, host type, mortality rate, basic reproduction number, and predefined risk category would enable the development of supervised machine learning models for viral risk prediction and pandemic preparedness.

3. Result and Discussion

Dataset Characteristics:

The dataset used in this study consisted of 70 virus records with taxonomic and genomic attributes, including family, genus, genome type, strand type, and envelope status. The exploratory analysis showed that the dataset contained a considerable proportion of incomplete information, represented by the value `Unknown`. This condition was particularly visible in the attribute's `family`, `genus`, `genome_type`, `strand`, and `enveloped`.

The distribution of the `enveloped` attribute showed that 41 records were categorized as `Unknown`, 25 records were identified as enveloped viruses, and only 4 records were identified as non-enveloped viruses. This indicates that most entries did not provide complete envelope status information. Among the known records, enveloped viruses were more dominant than non-enveloped viruses.

Table 5. Distribution of Envelope Status

Envelope Status	Number of Records
Unknown	41
TRUE	25
FALSE	4

As shown in **Figure 1**, the large number of `Unknown` values indicates that envelope status was not available for the majority of virus records. This affects the clustering process because viruses with incomplete information may be grouped together based on shared missing attributes rather than biological similarity [23].

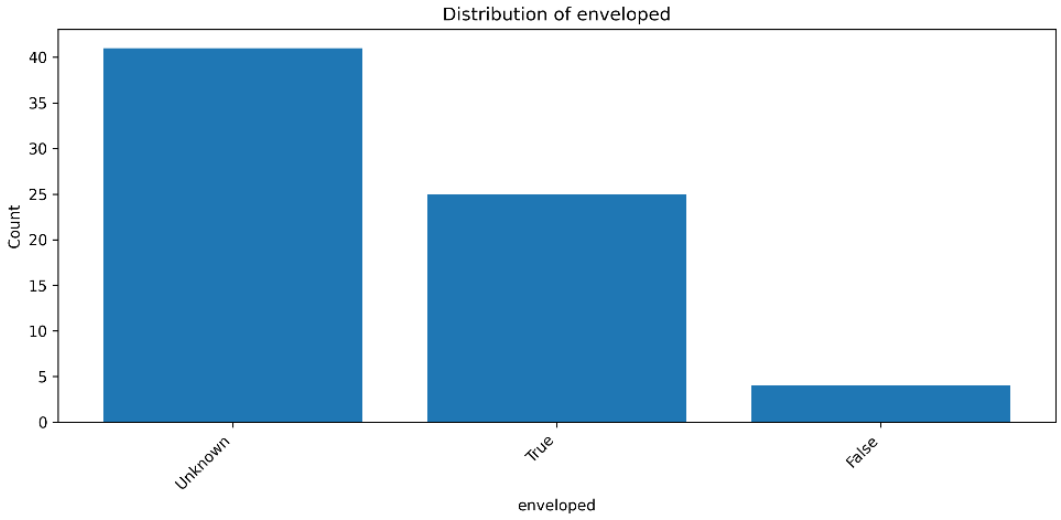

Figure 1. Distribution of envelope status among virus records.

The family-level distribution also showed a strong dominance of `Unknown` values. A total of 41 records did not contain family information. Among known families, Flaviviridae appeared most frequently with 5 records, followed by Coronaviridae with 4 records and Herpesviridae with 3 records. Other families such as Poxviridae, Filoviridae, and Paramyxoviridae appeared in smaller numbers.

Table 6. Distribution of Virus Families

Family	Number of Records
Unknown	41
Flaviviridae	5
Coronaviridae	4
Herpesviridae	3
Poxviridae	2
Filoviridae	2
Paramyxoviridae	2
Other families	1 each

As illustrated in **Figure 2**, the distribution of virus families was highly imbalanced. The dominance of the `Unknown` category suggests that the dataset contains many virus records that are not fully annotated taxonomically. This condition is important because taxonomic features are central to the clustering process.

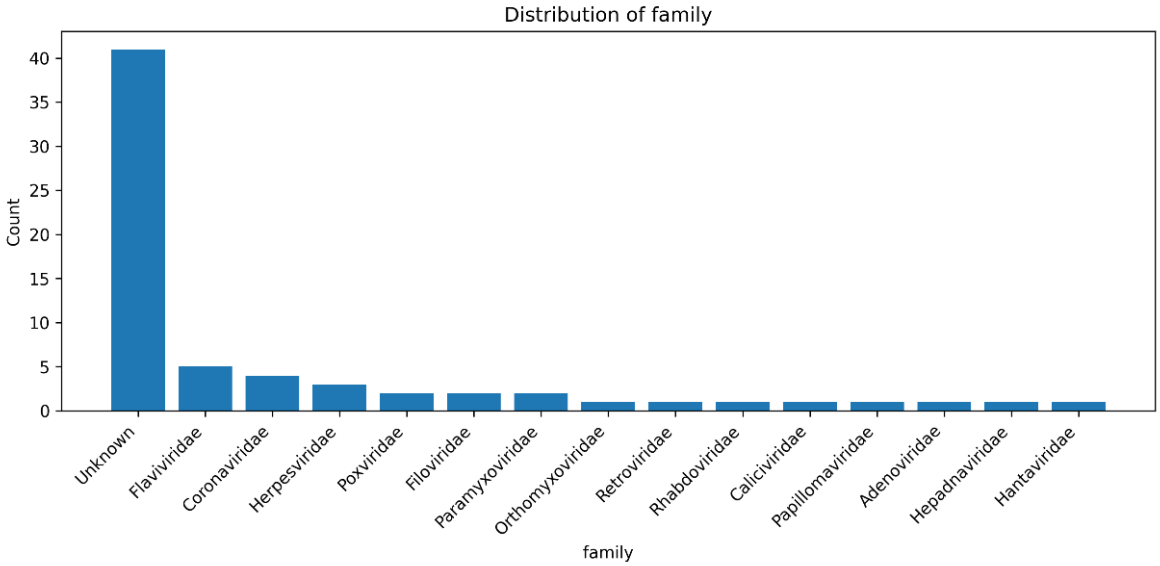

Figure 2. Distribution of virus family categories

The distribution of genome type showed that 41 records were categorized as `Unknown`, 21 records were RNA viruses, and 8 records were DNA viruses. Among the known records, RNA viruses were more frequent than DNA viruses.

Table 7. Distribution of Genome Type

Genome Type	Number of Records
Unknown	41
RNA	21
DNA	8

Figure 3 shows that RNA viruses represented the largest known genome group in the dataset. This is consistent with the inclusion of several medically important RNA viruses, such as coronaviruses, flaviviruses, filoviruses, orthomyxoviruses, and paramyxoviruses. However, the large proportion of `Unknown` values limits the ability to draw broad biological conclusions from the dataset.

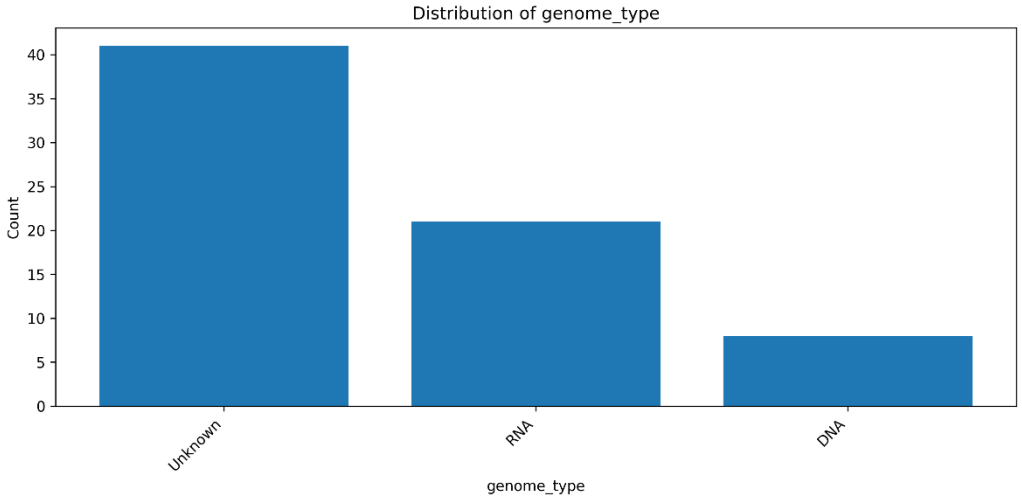

Figure 3. Distribution of genome type among virus records

The genus distribution followed a similar pattern. A total of 41 records were categorized as `Unknown`. Among the known genera, Flavivirus was the most frequent with 5 records, followed by Betacoronavirus with 3 records. Several other genera appeared only once or twice.

Table 8. Distribution of Dominant Virus Genera

Genus	Number of Records
Unknown	41
Flavivirus	5
Betacoronavirus	3
Orthopoxvirus	2
Simplexvirus	2
Other genera	1 each

As shown in **Figure 4**, the known genus categories were highly fragmented. This means that many genera were represented by only one virus record. Such sparsity can affect unsupervised learning because some clusters may represent small, biologically specific groups rather than broad patterns.

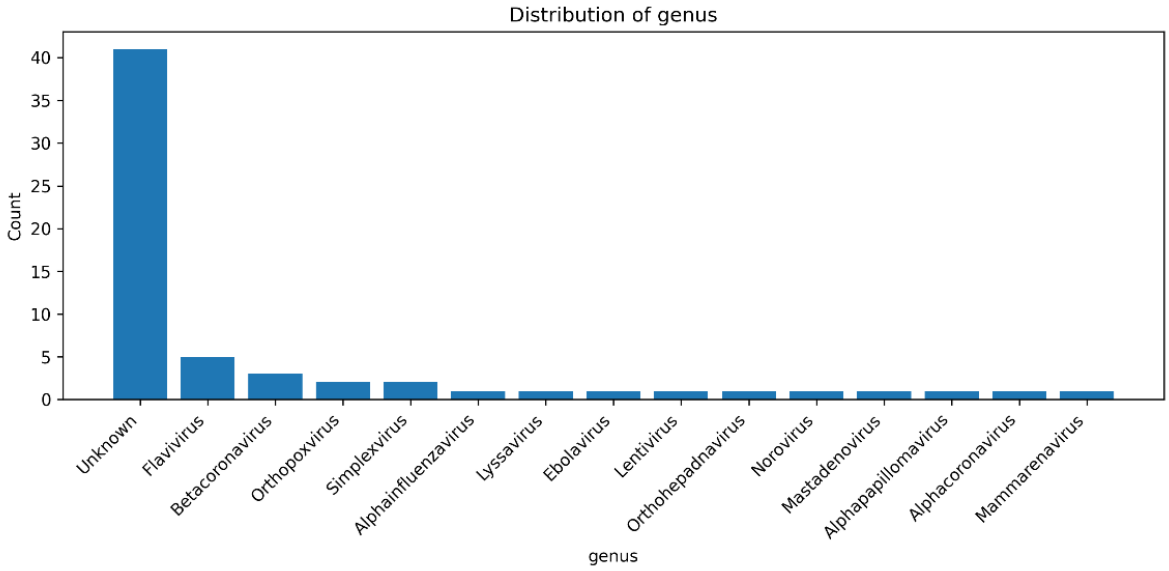

Figure 4. Distribution of virus genus categories

The strand distribution showed that 41 records were labeled as `Unknown`. Among known records, ssRNA(+) was the most frequent strand type with 10 records, followed by ssRNA(-) with 8 records and dsDNA with 7 records. Other strand types, including partially dsDNA, ssRNA-RT, dsRNA, and ssRNA, appeared only once.

Table 9. Distribution of Viral Strand Type

Strand Type	Number of Records
Unknown	41
ssRNA(+)	10
ssRNA(-)	8
dsDNA	7
partially dsDNA	1
ssRNA-RT	1
dsRNA	1
ssRNA	1

Figure 5 indicates that the known strand types were dominated by single-stranded RNA viruses, particularly positive-sense and negative-sense RNA viruses. This finding supports the relevance of genome and strand attributes for grouping viruses based on biological characteristics.

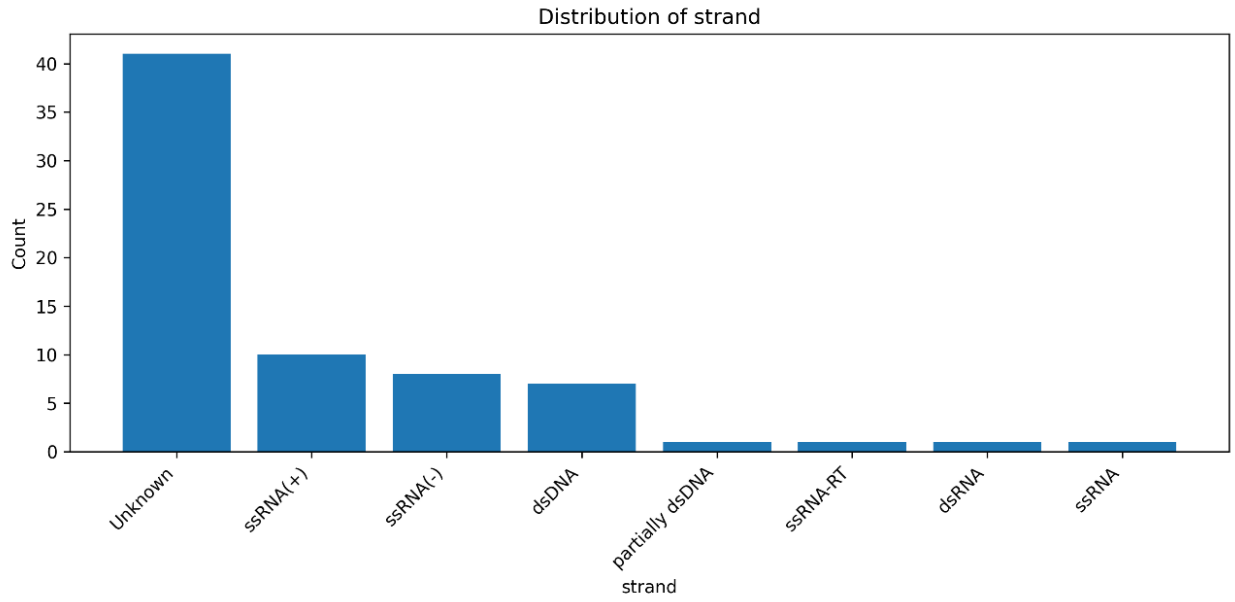

Figure 5. Distribution of viral strand types

Overall, the exploratory analysis revealed two important characteristics of the dataset. First, the dataset contained several biologically meaningful virus groups, especially RNA viruses, DNA viruses, enveloped viruses, and specific virus families. Second, the dataset contained a large number of incomplete annotations, which became a major factor influencing the clustering results.

Determination of the Optimal Number of Clusters:

The K-Means clustering algorithm was evaluated using different numbers of clusters, ranging from 2 to 10. The Silhouette Score was used as the main criterion to determine the most suitable number of clusters. The results showed that the Silhouette Score increased as the number of clusters increased.

Table 10. Evaluation of Different Numbers of Clusters

Number of Clusters	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Score
2	6,713	6,834	837,723
3	6,896	1.2156	649,600
4	6,929	1.3400	576,438
5	7,260	1.1503	520,636
6	7,290	1.1357	488,088
7	7,441	1.0425	484,238
8	7,492	1.0607	476,943

Number of Clusters	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Score
9	7,495	8,854	464,075
10	7,725	8,186	485,840

The highest Silhouette Score was obtained when the number of clusters was set to 10, with a score of 0.7725. This indicates that the data points were relatively well separated under the 10-cluster configuration. The Davies-Bouldin Index at k = 10 was also relatively low, with a value of 0.8186, indicating acceptable cluster separation and compactness.

As shown in Figure 6, the Silhouette Score increased from k = 2 to k = 10. However, it should be noted that k = 10 was the highest value tested in this experiment. Therefore, the result should be interpreted as the best number of clusters within the tested range, not necessarily the absolute optimal number of clusters for all possible values of k.

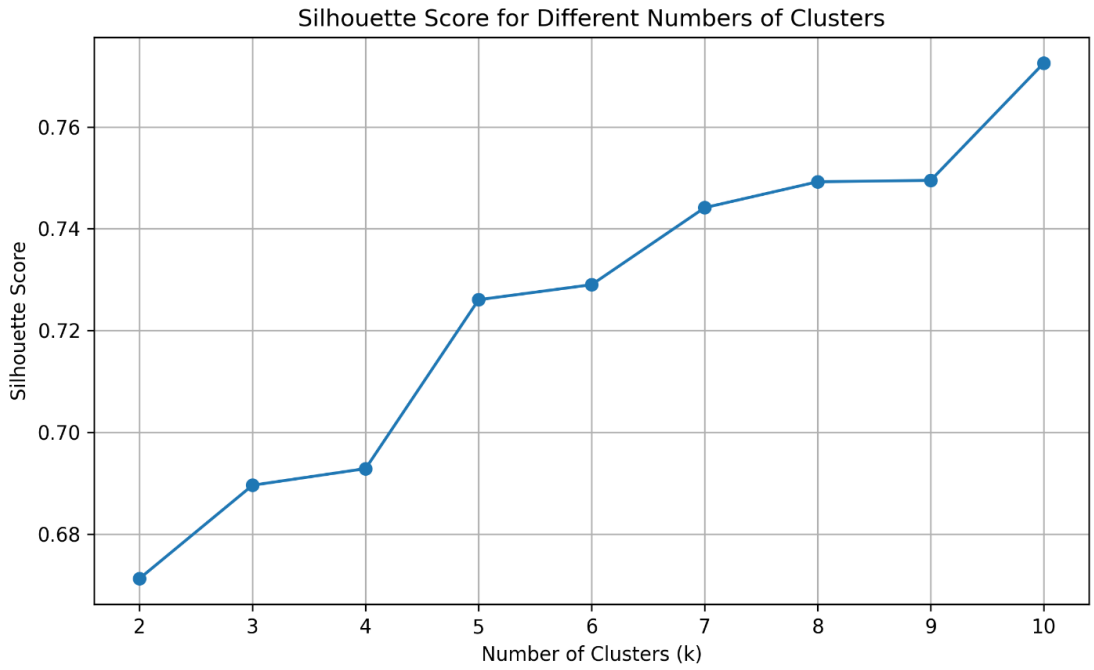

Figure 6. Silhouette Score for different numbers of clusters

The increasing trend suggests that the dataset may contain several small and distinct biological groups. This is reasonable because many viruses in the dataset belong to specific families or genera with unique taxonomic and genomic characteristics. However, the existence of many `Unknown` records also contributed to the clustering structure, particularly by forming a large cluster of incomplete records.

Comparison of Clustering Algorithms:

Three clustering algorithms were evaluated in this study: K-Means, Agglomerative Clustering, and DBSCAN. The performance of these algorithms was assessed using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score.

Table 11. Comparison of Clustering Algorithms

Model	Number of Clusters	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Score
K-Means	10	7,725	8,186	485,840
Agglomerative Clustering	10	7,754	8,933	504,762
DBSCAN	4	7,044	1.1077	406,753

Agglomerative Clustering produced the highest Silhouette Score, with a value of 0.7754, and the highest Calinski-Harabasz Score, with a value of 50.4762. This suggests that the hierarchical clustering approach was slightly better in terms of cluster separation and inter-cluster structure. However, K-Means produced the lowest Davies-Bouldin Index, with a value of 0.8186, indicating more compact clusters compared with the other algorithms.

DBSCAN generated only 4 clusters and obtained a lower Silhouette Score of 0.7044. Although this score was still acceptable, DBSCAN performed less effectively than K-Means and Agglomerative Clustering in this dataset. This may be caused by the nature of the one-hot encoded categorical data, where density-based separation is less effective than centroid-based or hierarchical grouping. The exponential growth of viral metagenomic data has created an urgent need for accurate and scalable tools for virus discovery, yet the extreme diversity, rapid evolution, and limited reference databases for viruses pose unique computational challenges that traditional sequence comparison methods struggle to address [24]. This systematic review, conducted in accordance with PRISMA 2020, examines current trends and methodological advances in virus discovery tools from 1990 to 2025. As virus discovery is a broad and multi-dimensional topic, this review focuses on the first-line tools used to analyze the results of high-throughput sequencing. The review was conducted using the PubMed database with a snowballing approach, with over 54 key studies selected for the analysis. These studies encompass the following approaches: alignment-based methods, rapid similarity estimation techniques, profile hidden Markov model methods, combination pipelines, k-mer-based approaches, and machine learning-based methods. The transition from alignment-based to machine learning methods has dramatically improved the detection of divergent viruses, yet challenges remain in interpreting model decisions and handling incomplete viral genomes. This review summarizes current knowledge and potential future directions for the development of virus detection capabilities [20].

Based on the overall evaluation, K-Means was selected as the main clustering model for further cluster profiling. Although Agglomerative Clustering produced a slightly higher Silhouette Score, K-Means provided a balanced performance, a lower Davies-Bouldin Index, and a clearer cluster structure for interpretation.

Cluster Profiling:

After applying K-Means with 10 clusters, each virus record was assigned to a cluster. The cluster distribution showed that Cluster 1 contained the largest number of records, with 41 viruses. This cluster was dominated by `Unknown` values across family, genus, genome type, strand, and envelope status. This indicates that the clustering algorithm grouped incomplete records into a single large cluster.

Table 12. Cluster Profile Based on Dominant Attributes

Cluster	Total Viruses	Dominant Family	Dominant Genus	Dominant Genome Type	Dominant Strand	Dominant Enveloped
0	9	Filoviridae	Alphainfluenzavirus	RNA	ssRNA(-)	TRUE
1	41	Unknown	Unknown	Unknown	Unknown	Unknown
2	5	Flaviviridae	Flavivirus	RNA	ssRNA(+)	TRUE
3	2	Adenoviridae	Alphapapillomavirus	DNA	dsDNA	FALSE
4	4	Coronaviridae	Betacoronavirus	RNA	ssRNA(+)	TRUE
5	3	Herpesviridae	Simplexvirus	DNA	dsDNA	TRUE
6	2	Caliciviridae	Norovirus	RNA	dsRNA	FALSE
7	2	Poxviridae	Orthopoxvirus	DNA	dsDNA	TRUE
8	1	Hepadnaviridae	Orthohepadnavirus	DNA	partially dsDNA	TRUE
9	1	Arenaviridae	Mammarenavirus	RNA	ssRNA	TRUE

Cluster 1 was the largest cluster and represented viruses with incomplete taxonomic and genomic information. This cluster included many phage-related and less completely annotated records. The formation of this cluster highlights the influence of missing information on the clustering result. From a data quality perspective, this cluster should not be interpreted as a biologically homogeneous group, but rather as a group formed due to shared missing values.

Cluster 2 represented Flaviviridae viruses, including Zika virus, Dengue virus, Yellow fever virus, West Nile virus, and Japanese encephalitis virus. These viruses shared highly consistent characteristics: they belonged to the family Flaviviridae, genus Flavivirus, had RNA genomes, ssRNA(+) strand type, and were enveloped. This cluster demonstrates that the proposed clustering approach was able to identify biologically coherent virus groups.

Cluster 4 represented coronaviruses, including SARS-CoV-2, MERS-CoV, HCoV-OC43, and HCoV-229E. Most viruses in this cluster belonged to the family Coronaviridae and had RNA genomes with ssRNA(+) strand type. This result indicates that clustering based on taxonomic and genomic attributes can separate coronaviruses into a distinct group.

Cluster 5 consisted of herpesviruses, including Varicella-zoster virus, Herpes simplex virus 1, and Herpes simplex virus 2. These viruses shared DNA genomes, dsDNA strand type, and enveloped status. This cluster also showed strong biological consistency.

Cluster 7 consisted of Smallpox virus and Monkeypox virus, both belonging to the family Poxviridae and genus Orthopoxvirus. These viruses shared DNA genomes, dsDNA strand type, and enveloped status. This result further confirms that the clustering model was able to group viruses with similar taxonomic identity.

Cluster 0 contained several enveloped RNA viruses, especially negative-sense RNA viruses such as Influenza A virus, Ebola virus, Rabies virus, Marburg virus, Hantavirus, Measles virus, Mumps virus, and Respiratory syncytial virus. Although this cluster contained viruses from different families, they shared broader genomic and structural similarities, particularly RNA genome type, negative-sense strand characteristics, and enveloped status.

Clusters 8 and 9 each contained only one virus. Cluster 8 contained Hepatitis B virus, while Cluster 9 contained Lassa virus. These singleton clusters indicate that some viruses had unique combinations of attributes that separated them from the other groups. In clustering analysis, singleton clusters may represent distinct biological profiles or may emerge because of limited sample representation.

PCA Visualization of Virus Clusters:

Principal Component Analysis was used to visualize the clustering result in a two-dimensional space. The PCA visualization showed that several clusters were separated from one another, while some data points overlapped or appeared close together. This pattern reflects the structure of the one-hot encoded categorical features.

As shown in **Figure 7**, many points were concentrated in certain areas because several viruses shared identical or highly similar categorical attributes. In particular, viruses with `Unknown` values tended to overlap because they had the same encoded representation across multiple features. Meanwhile, viruses with specific taxonomic and genomic profiles appeared as smaller separated groups.

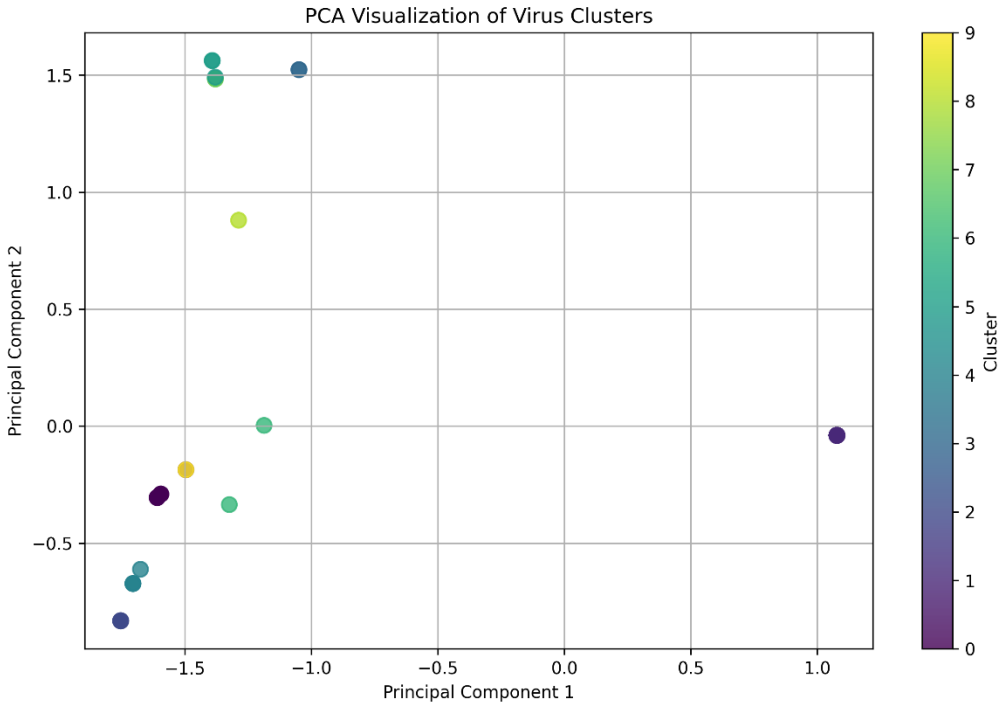

Figure 7. PCA visualization of virus clusters.

The PCA visualization supports the quantitative evaluation results by showing that some clusters were clearly separated. However, because PCA reduces high-dimensional categorical data into only two components, it may not fully represent all dimensions used by the clustering algorithms. Therefore, PCA should be interpreted as a visual aid rather than as the primary evidence of clustering quality.

Discussion:

The results of this study indicate that unsupervised machine learning can be used to explore viral datasets based on taxonomic and genomic characteristics. The clustering process successfully identified several biologically meaningful groups, such as flaviviruses, coronaviruses, herpesviruses, poxviruses, and negative-sense RNA viruses. These results suggest that categorical biological attributes can provide useful signals for computational virus grouping.

The strongest clustering patterns were observed in virus groups with complete and consistent annotations. For example, Flaviviridae viruses were grouped into a coherent cluster because they shared the same family, genus, genome type, strand type, and envelope status. A similar pattern was observed for Coronaviridae, Herpesviridae, and Poxviridae. This finding demonstrates that machine learning can support the organization of virus data when sufficient taxonomic and genomic information is available.

However, the analysis also revealed a major limitation of the dataset. A large proportion of records contained `Unknown` values, resulting in the formation of a large cluster dominated by incomplete data. This condition shows that missing information can strongly influence unsupervised learning outcomes. In this study, 41 out of 70 records

had unknown values in several important attributes. As a result, Cluster 1 was formed primarily because many records shared missing values rather than because they shared confirmed biological similarity.

From a health informatics perspective, this finding is important because data completeness directly affects the reliability of machine learning models. If virus datasets are intended to support pandemic preparedness, disease surveillance, or viral risk stratification, they should include more complete epidemiological and clinical variables. Important attributes such as transmission mode, host type, mortality rate, basic reproduction number, severity index, and predefined risk category would allow future studies to move from exploratory clustering to supervised risk prediction.

The comparison among clustering algorithms showed that K-Means and Agglomerative Clustering performed better than DBSCAN. Agglomerative Clustering obtained the highest Silhouette Score, while K-Means obtained the lowest Davies-Bouldin Index. This indicates that both centroid-based and hierarchical methods are suitable for this type of encoded categorical dataset. DBSCAN produced fewer clusters and lower evaluation scores, suggesting that density-based clustering may be less suitable for this dataset structure.

Overall, the findings demonstrate that the proposed machine learning workflow can provide an initial framework for virus grouping based on biological characteristics. Although the current study does not perform disease severity prediction, it provides a foundation for further research in AI-driven viral risk profiling. With additional epidemiological features, the framework can be extended into a supervised learning model for viral risk classification and pandemic preparedness.

Summary of Key Findings:

The key findings of this study are as follows. First, the dataset was dominated by incomplete annotations, with 41 records categorized as `Unknown` across several attributes. Second, among known records, RNA viruses and enveloped viruses were more frequent than DNA and non-enveloped viruses. Third, K-Means with 10 clusters achieved a Silhouette Score of 0.7725 and a Davies-Bouldin Index of 0.8186. Fourth, Agglomerative Clustering achieved the highest Silhouette Score of 0.7754. Fifth, several biologically meaningful clusters were identified, including clusters representing Flaviviridae, Coronaviridae, Herpesviridae, and Poxviridae. Finally, the large `Unknown` cluster highlighted the importance of data completeness in AI-based virology analysis.

4. Conclusion

This study presented an unsupervised machine learning approach for clustering viruses based on taxonomic and genomic characteristics. The dataset consisted of 70 virus records with attributes including family, genus, genome type, strand type, and envelope status. Since the dataset did not contain predefined epidemiological labels or risk categories, the analysis was designed as an exploratory clustering task rather than a supervised prediction task.

The exploratory analysis showed that the dataset contained a substantial proportion of incomplete annotations, particularly in the `family`, `genus`, `genome_type`, `strand`, and `enveloped` attributes. A total of 41 records

were categorized as `Unknown` across several important features. This condition influenced the clustering results, as many incomplete records were grouped into a large cluster dominated by unknown values. Therefore, the findings highlight the importance of data completeness in developing reliable machine learning models for virology and health informatics applications.

The clustering results demonstrated that machine learning can identify meaningful grouping patterns among viruses when taxonomic and genomic information is available. Several biologically coherent clusters were formed, including groups representing Flaviviridae, Coronaviridae, Herpesviridae, Poxviridae, and several enveloped RNA viruses. K-Means clustering with 10 clusters achieved a Silhouette Score of 0.7725 and a Davies-Bouldin Index of 0.8186, indicating acceptable cluster separation and compactness. Agglomerative Clustering produced a slightly higher Silhouette Score of 0.7754, while DBSCAN produced fewer clusters and lower overall evaluation performance.

Overall, this study demonstrates that unsupervised machine learning can support the organization and exploration of virus-related datasets based on biological similarity. Although the current dataset does not support direct viral risk prediction, the proposed workflow provides an initial computational framework for AI-driven virus grouping. The results may serve as a foundation for future studies in computational virology, health informatics, and pandemic preparedness.

Future work should focus on enriching the dataset with epidemiological and clinical variables, such as transmission mode, host type, mortality rate, basic reproduction number, severity index, and predefined risk category [25]. The inclusion of these variables would enable the development of supervised machine learning models for viral risk classification, disease severity prediction, and pandemic risk modeling. In addition, future research may explore more advanced feature engineering strategies and larger curated datasets to improve the robustness and clinical relevance of AI-based viral analysis [5].

References:

[1] P. J. Walker *et al.*, "Changes to virus taxonomy and to the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2021)," *Arch. Virol.*, vol. 166, no. 9, pp. 2633–2648, Sep. 2021, doi: 10.1007/s00705-021-05156-1.

[2] Y.-M. Chen *et al.*, "RNA viromes from terrestrial sites across China expand environmental viral diversity," *Nat. Microbiol.*, vol. 7, no. 8, pp. 1312–1323, Jul. 2022, doi: 10.1038/s41564-022-01180-2.

[3] P. J. Walker *et al.*, "Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of Viruses (2022)," *Arch. Virol.*, vol. 167, no. 11, pp. 2429–2440, Nov. 2022, doi: 10.1007/s00705-022-05516-5.

[4] M. Krupovic *et al.*, "Bacterial Viruses Subcommittee and Archaeal Viruses Subcommittee of the ICTV: update of taxonomy changes in 2021," *Arch. Virol.*, vol. 166, no. 11, pp. 3239–3244, Nov. 2021, doi: 10.1007/s00705-021-05205-9.

[5] S. Nayfach *et al.*, “Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome,” *Nat. Microbiol.*, vol. 6, no. 7, pp. 960–970, Jun. 2021, doi: 10.1038/s41564-021-00928-6.

[6] P. Simmonds *et al.*, “Changes to virus taxonomy and the ICTV Statutes ratified by the International Committee on Taxonomy of Viruses (2024),” *Arch. Virol.*, vol. 169, no. 11, p. 236, Nov. 2024, doi: 10.1007/s00705-024-06143-y.

[7] D. Turner *et al.*, “Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee,” *Arch. Virol.*, vol. 168, no. 2, p. 74, Feb. 2023, doi: 10.1007/s00705-022-05694-2.

[8] B. E. Dutilh *et al.*, “Perspective on taxonomic classification of uncultivated viruses,” *Curr. Opin. Virol.*, vol. 51, pp. 207–215, Dec. 2021, doi: 10.1016/j.coviro.2021.10.011.

[9] E. V. Koonin, J. H. Kuhn, V. V. Dolja, and M. Krupovic, “Megataxonomy and global ecology of the virosphere,” *ISME J.*, vol. 18, no. 1, p. wrad042, Jan. 2024, doi: 10.1093/ismejo/wrad042.

[10] A. Zielezinski *et al.*, “Ultrafast and accurate sequence alignment and clustering of viral genomes,” *Nat. Methods*, vol. 22, no. 6, pp. 1191–1194, Jun. 2025, doi: 10.1038/s41592-025-02701-7.

[11] T. S. Postler *et al.*, “Renaming of the genus Flavivirus to Orthoflavivirus and extension of binomial species names within the family Flaviviridae,” *Arch. Virol.*, vol. 168, no. 9, pp. 224, s00705-023-05835–1, Sep. 2023, doi: 10.1007/s00705-023-05835-1.

[12] J. H. Kuhn *et al.*, “2022 taxonomic update of phylum Negarnaviricota (Riboviria: Orthornavirae), including the large orders Bunyavirales and Mononegavirales,” *Arch. Virol.*, vol. 167, no. 12, pp. 2857–2906, Dec. 2022, doi: 10.1007/s00705-022-05546-z.

[13] A. P. Camargo *et al.*, “IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata,” *Nucleic Acids Res.*, vol. 51, no. D1, pp. D733–D743, Jan. 2023, doi: 10.1093/nar/gkac1037.

[14] A. P. Camargo *et al.*, “Identification of mobile genetic elements with geNomad,” *Nat. Biotechnol.*, vol. 42, no. 8, pp. 1303–1312, Aug. 2024, doi: 10.1038/s41587-023-01953-y.

[15] K. Zheng *et al.*, “VITAP: a high precision tool for DNA and RNA viral classification based on meta-omic data,” *Nat. Commun.*, vol. 16, no. 1, p. 2226, Mar. 2025, doi: 10.1038/s41467-025-57500-7.

[16] J.-Z. Jiang *et al.*, “Virus classification for viral genomic fragments using PhaGCN2,” *Brief. Bioinform.*, vol. 24, no. 1, p. bbac505, Jan. 2023, doi: 10.1093/bib/bbac505.

[17] J. Shang, J. Jiang, and Y. Sun, “Bacteriophage classification for assembled contigs using graph convolutional network,” *Bioinformatics*, vol. 37, no. Supplement_1, pp. i25–i33, Aug. 2021, doi: 10.1093/bioinformatics/btab293.

[18] Y. Zhu, G. Chen, and Y. Sun, “VirTAXA: enhancing RNA virus taxonomic classification with remote homology search and tree-based validation,” *Bioinformatics*, vol. 40, no. 10, p. btae575, Oct. 2024, doi: 10.1093/bioinformatics/btae575.

[19] C. Peng, J. Shang, J. Guan, D. Wang, and Y. Sun, “ViraLM: empowering virus discovery through the genome foundation model,” *Bioinformatics*, vol. 40, no. 12, p. btae704, Nov. 2024, doi: 10.1093/bioinformatics/btae704.

[20] F. Alipour, C. Holmes, Y. Y. Lu, K. A. Hill, and L. Kari, “Leveraging machine learning for taxonomic classification of emerging astroviruses,” *Front. Mol. Biosci.*, vol. 10, p. 1305506, Jan. 2024, doi: 10.3389/fmolb.2023.1305506.

[21] G. Chen, J. Jiang, and Y. Sun, “RNAVirHost: a machine learning–based method for predicting hosts of RNA viruses through viral genomes,” *GigaScience*, vol. 13, p. giae059, Jan. 2024, doi: 10.1093/gigascience/giae059.

[22] K. S. Azevedo, L. C. De Souza, M. G. F. Coutinho, R. De M. Barbosa, and M. A. C. Fernandes, “Deepvirusclassifier: a deep learning tool for classifying SARS-CoV-2 based on viral subtypes within the coronaviridae family,” *BMC Bioinformatics*, vol. 25, no. 1, p. 231, Jul. 2024, doi: 10.1186/s12859-024-05754-1.

[23] J. Shang, C. Peng, H. Liao, X. Tang, and Y. Sun, “PhaBOX: a web server for identifying and characterizing phage contigs in metagenomic data,” *Bioinforma. Adv.*, vol. 3, no. 1, p. vbad101, Jan. 2023, doi: 10.1093/bioadv/vbad101.

[24] B. Hegarty *et al.*, “Benchmarking informatics approaches for virus discovery: caution is needed when combining *in silico* identification methods,” *mSystems*, vol. 9, no. 3, pp. e01105-23, Mar. 2024, doi: 10.1128/msystems.01105-23.

[25] J. Galeeva, P. Kuzmichenko, A. Manolov, A. Lukashev, and E. Ilina, “Bioinformatics Tools and Approaches for Virus Discovery in Genomic Data: A Systematic Review,” *Viruses*, vol. 17, no. 12, p. 1538, Nov. 2025, doi: 10.3390/v17121538.