OXFORD

Briefings in Bioinformatics, 2024, **25(3)**, bbae224

https://doi.org/10.1093/bib/bbae224

Problem Solving Protocol

Analysis of Emerging Variants of Turkey Reovirus using Machine Learning

Maryam KafiKang[1], Chamudi Abeysiriwardana[1], Vikash K. Singh[2], Chan Young Koh[1], Janet Prichard[3], Sunil K. Mor[2,4,*], Abdeltawab Hendawi[1,5,*]

[1]Computer Science Department, University of Rhode Island, Kingston, 02881, RI, USA
[2]Department of Veterinary Population Medicine, and Veterinary Diagnostic Laboratory, University of Minnesota, Saint Paul, 55108, MN, USA
[3]Department of Information Systems and Analytics, Bryant University, Smithfield, 02917, RI, USA
[4]Department of Veterinary and Biomedical Sciences and Animal Disease Research & Diagnostic Laboratory, South Dakota State University, Brookings, 57007, SD, USA
[5]Faculty of Computers and Artificial Intelligence, Cairo University, Giza, Egypt

*Corresponding authors. Sunil Mor, Email: sunil.mor@sdstate.edu, And Abdeltawab Hendawi, E-mail: hendawi@uri.edu.

Abstract

Avian reoviruses continue to cause disease in turkeys with varied pathogenicity and tissue tropism. Turkey enteric reovirus has been identified as a causative agent of enteritis or inapparent infections in turkeys. The new emerging variants of turkey reovirus, tentatively named turkey arthritis reovirus (TARV) and turkey hepatitis reovirus (THRV), are linked to tenosynovitis/arthritis and hepatitis, respectively. Turkey arthritis and hepatitis reoviruses are causing significant economic losses to the turkey industry. These infections can lead to poor weight gain, uneven growth, poor feed conversion, increased morbidity and mortality and reduced marketability of commercial turkeys. To combat these issues, detecting and classifying the types of reoviruses in turkey populations is essential. This research aims to employ clustering methods, specifically K-means and Hierarchical clustering, to differentiate three types of turkey reoviruses and identify novel emerging variants. Additionally, it focuses on classifying variants of turkey reoviruses by leveraging various machine learning algorithms such as Support Vector Machines, Naive Bayes, Random Forest, Decision Tree, and deep learning algorithms, including convolutional neural networks (CNNs). The experiments use real turkey reovirus sequence data, allowing for robust analysis and evaluation of the proposed methods. The results indicate that machine learning methods achieve an average accuracy of 92%, F1-Macro of 93% and F1-Weighted of 92% scores in classifying reovirus types. In contrast, the CNN model demonstrates an average accuracy of 85%, F1-Macro of 71% and F1-Weighted of 84% scores in the same classification task. The superior performance of the machine learning classifiers provides valuable insights into reovirus evolution and mutation, aiding in detecting emerging variants of pathogenic TARVs and THRVs.

Keywords: *Reovirus analysis; machine learning; reovirus classification; double-stranded RNA*

INTRODUCTION

Avian reoviruses (ARVs) have been linked to various diseases that impact avian species. These species encompass turkeys and chickens[1]. Turkey reoviruses are a subset of ARVs belonging to the genus Orthoreovirus and family Reoviridae. ARVs are non-enveloped double-stranded RNA genomes with icosahedral symmetry and a particle size of 70–80 nm [2, 3]. The viral genome consists of 10 segments divided into three classes, e.g. large (L), medium (M) and small (S), depending on their migration pattern on polyacrylamide gel electrophoresis [3, 4]. The L and M genes are subdivided into three segments each (L_1, L_2, L_3, and M_1, M_2, M_3, respectively), while the S gene has four segments (S_1, S_2, S_3, S_4; 36). The proteins encoded by L, M and S genes are lambda (λ), mu (μ) and sigma (σ), respectively. Three structural proteins λ_A, λ_B and λ_C are encoded by L gene segment, L_1, L_2 and L_3, respectively. M_1 and M_2 segments encode two structural proteins (μ_A and μ_B), while M_3 segment encodes a non-structural protein (μ_{NS}). The three structural proteins (σ_C, σ_A, σ_B) are encoded by S_1, S_2, S_3 segments, respectively, while the non-structural protein σ_{NS} is encoded by the S_4 segment [3, 4].

ARVs, including turkey reoviruses, have been linked to several diseases, including enteritis, hepatitis, myocarditis, respiratory disease and viral/tenosynovitis [5]. The first isolation of turkey reoviruses from lame turkeys was reported in 1980 [6]. In 2011, the virus reemerged, and five turkey arthritis reoviruses (TARVs) were isolated [7]. At that point, turkey reoviruses were divided into two groups depending on the type of disease they produced and the site of virus isolation: TARVs from tendons in lameness cases and turkey enteric reoviruses (TERVs) from intestines in enteric cases [7, 8]. In 2019, another turkey reovirus was isolated from liver samples from several cases of hepatitis in turkey poults and tentatively named turkey hepatitis reovirus (THRV) [9].

Received: October 27, 2023. **Revised:** March 26, 2024. **Accepted:** April 25, 2024

These reovirus diseases can significantly negatively impact the turkey industry as they lead to poor weight gain, uneven growth, poor feed conversion, increased morbidity and mortality, resulting in economic losses and reduced marketability of commercial turkeys [5].

In a survey published in the Proceedings of the 2019 USAHA Annual Meeting, TARV cases increased by 108% (a total of 5.627 million birds) in 2019. They were ranked in the top 10 diseases of concern for the turkey industry [10]. The severity of the financial impact was estimated to be as high as 33.7 million dollars with highly pathogenic strains of TARV [10].

Therefore, early detection and proper classification of pathogenic turkey reoviruses and implementing strategies to mitigate their impact can play a critical role in ensuring the profitability and stability of the turkey farming industry.

Machine learning (ML) and deep learning (DL) techniques have emerged as powerful tools for classifying types of viruses in genomics data [11–15]. Studies have demonstrated that ML and DL models can achieve high accuracy and efficiency when applied to genomics data, making them a valuable tool for identifying similarities and differences between DNA sequences [16–18].

ML algorithms analyze biological data in two main ways: supervised and unsupervised learning. Supervised learning algorithms are trained on labeled data to predict the outcome of new data points. Unsupervised learning algorithms are trained on unlabeled data to find patterns and relationships in the data. ML algorithms can answer various biological questions, such as predicting disease risk, identifying new disease genes, grouping genes with similar expression patterns or identifying groups of patients with similar clinical features [19, 20].

Several noteworthy works have contributed to advancing DNA sequence clustering, classification and disease diagnosis in bioinformatics and computational biology. Wei *et al.* [21] introduced the mBKM (modified Bisecting K-Means algorithm) technique for clustering DNA sequences, accompanied by the innovative DMk (Distance Measure based on k-tuples) sequence similarity metric. Nguyen *et al.* [22] proposed a novel approach to classify DNA sequences utilizing convolutional neural networks (CNNs). Machuve *et al.* [23] contributed to the field by employing DL for diagnosing poultry diseases based on image analysis. In contrast, Mbelwa *et al.* [24] introduced a DL solution for detecting chicken diseases using CNNs. Gunasekaran *et al.* [11] delved into DNA sequence classification for viruses, achieving remarkable accuracy. Whata *et al.* [25] focused on classifying severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) gene sequences and regulatory motif identification using DL. De *et al.* [26] introduced a novel gene sequence representation using Genomic Signal Processing for SARS-CoV-2 virus classification. Cho and Won [27] systematically evaluated feature selection methods and ML classifiers on benchmark cancer datasets, and Eickholt *et al.* [28] presented a boosted ensemble of deep networks for protein disorder prediction from sequences. These works collectively contribute to the expanding landscape of computational biology and bioinformatics, offering valuable insights and innovative methodologies for various applications in the field.

However, previous research has not yet focused on developing ML models specifically for the detection and characterization of emerging variants of ARVs. The objective of this study is to devise a rapid classification method for Turkey Reoviruses, aimed at enhancing the health and productivity of turkey flocks.

MATERIALS AND METHODS

Dataset

It has been well established through numerous biological studies that biological sequences, such as those found in DNA, are not random or unordered strings but instead have a highly structured and organized linear arrangement of more minor elements [29]. The DNA sequence is composed of four distinct deoxyribonucleotides bases—adenine(A), thymine(T), cytosine(C) and guanine(G)—and the specific order or sequence of these bases is what gives rise to the diversity of DNA molecules and the genetic information they carry.

We start with a set of TRV sequences that includes turkey reovirus whole-genome sequences (WGSs), where each sequence is labeled with a reovirus type (TERV, TARV or THRV). The objective of this study is to identify new possible clusters and develop a reliable and efficient ML method for detecting and classifying various types of reoviruses in a given turkey reovirus WGS dataset comprised of 229 whole genomes, which correspond to 2290 sequences (229 samples $*$ 10 segments). However, our study focuses solely on 254 and 257 sequences of σ_C and μ_B, respectively.

The dataset contains a comprehensive WGS profile of four distinct turkey reoviruses variants sequenced from 2007 to 2021. These variants include mild enteritis (TERV) with 45 DNA sequences, tenosynovitis or arthritis (TARV) with 1650 DNA sequences and hepatitis (THRV) with 470 DNA sequences. Some sequences isolated from the spleen or heart that could not be classified as TARV, TERV or THRV are categorized as TRV, totaling 351 DNA sequences. Furthermore, this dataset is presented in 10 different FASTA files, each one providing DNA identifiers and their sequences on specific segments of the turkey reovirus as described above (λ_A, λ_B, λ_C, μ_A, μ_B, μ_C, σ_A, σ_B, σ_C and σ_{NS}) that are evident in Table 1. However, our study focuses solely on a subset of 511 sequences associated with the μ_B and σ_C segments, which are major outer capsid proteins and hence may be helpful in better classification of turkey reoviruses [30–32].

The data provided were aligned and trimmed using TRIMAL and then presented in the FASTA format. This diverse and extensive dataset provides a valuable resource for identifying and understanding the genetic makeup of these variants of turkey reoviruses, which is crucial for developing effective strategies to mitigate the negative effects of these viruses on the turkey farming industry.

Data preprocessing for modeling

DNA data pose unique challenges in processing and analysis, primarily consisting of a sequence of characters (A, T, C, G). It is necessary to convert the DNA sequence into numerical values to use DNA sequence data in many ML or DL models. Various techniques, such as encoding, are available for converting categorical data to numerical form.

In this work, we apply two encoding methods, k-mer and sequential encoding [11, 33] (S1 in the supplementary). These techniques can effectively convert DNA sequences into numerical representations that can be input for ML or DL models. Subsequently, we analyzed these techniques' impact on classification accuracy.

Clustering methods

We employ two clustering methods, `K-means` and `Hierarchical` clustering, to discover additional clusters within the data. These techniques allowed us to identify novel patterns and associations within the data, potentially leading to the discovery of new viral

Table 1: Turkey reovirus datasets

Reovirus	λ_A	λ_B	λ_C	μ_A	μ_B	μ_C	σ_A	σ_B	σ_C	σ_{NS}
TARV	163	168	165	165	**166**	164	164	166	**163**	166
THRV	47	47	45	48	**48**	46	47	47	**48**	47
TRV	36	38	38	40	**39**	40	40	40	**39**	39
TERV	5	4	5	5	**4**	5	5	4	**4**	4
Sum	251	257	253	258	**257**	256	256	257	**254**	256

subtypes or genetic variations that were not previously known. By expanding the scope of our analysis beyond the known viral types, we can gain a more comprehensive understanding of the underlying genetic structure of the dataset and enhance the depth of our findings. By applying this approach, we can more effectively analyze the sequences in our dataset and draw meaningful conclusions about the performance of different classifier models in identifying and classifying reoviruses. This allowed us to gain valuable insights into the underlying patterns and structures within the data, which helped inform our models' development and refinement. The supplementary materials section explains clustering algorithms (S2 in the supplementary).

Classification methods

This research uses four ML-based classification models (i.e. Decision Tree classifier, Random Forest Classifier, Support Vector Machine (SVM) and Multinomial Naive Bayes Classifier). We also used a DL classification model, CNNs.

This work employs the Random Search (RS) approach as the strategy for hyperparameter optimization to identify the best set of hyperparameters for each ML models under consideration. The RS method randomly selects a specified number of samples within a pre-defined range, which serves as the candidate hyperparameters, and trains them until the defined budget has been exhausted [34]. The hyperparameters that yield the highest performance are then chosen as the optimal hyperparameters. By utilizing RS, we can thoroughly explore the hyperparameter space, potentially leading to better performance than other tuning methods. Additionally, we implemented the K-fold cross-validation method to create a more generalized model. The supplementary materials section provides detailed explanations of the K-fold cross-validation process, Confident Learning (CL) and the specifics of each classification algorithm used in our study (S3 in the supplementary).

Workflow

Figure 1 presents our research workflow. The process begins with the preprocessing of reovirus sequences, incorporating a data cleaning step that is crucial for maintaining data integrity. This step involves verifying that the sequences exclusively comprise the nucleotides A, C, T and G. Subsequently, these preprocessed sequences undergo feature extraction, employing techniques such as K-mer and sequence encoding. This phase is essential for extracting important features and transforming them into a numerical representation suitable for ML analysis. Then, all features are first fed into a clustering algorithm to group the data and find new variants of reoviruses within the dataset. Secondly, all features are fed into a 10-fold cross-validation process to partition the dataset into training and test subsets. The training subset is utilized to train ML algorithms, aiming to identify and categorize reovirus types (TARV, THRV and TRV) within the sequences. The test subset serves to evaluate the performance of the models on unseen data.

Evaluation metrics

The performance of classification models is assessed by employing various classification metrics, including F1-score for each type of virus and overall scores such as accuracy, F1-macro, and F1-weighted.

$$Precision = \frac{TP}{(TP + FP)} \tag{1}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{2}$$

$$F1_{Score} = 2 * \left(\frac{Recall * Precision}{Recall + Precision}\right) \tag{3}$$

$$F1_{Macro} = \frac{\Sigma_{i=1}^{3} F1_{score}}{3} \tag{4}$$

$$F1_{Weighted} = \frac{\Sigma_{i=1}^{3} N_i * F1_i}{\Sigma_{i=1}^{3} N_i} \tag{5}$$

In class i, N_i denotes the number of instances, while *TP* denotes the number of correctly classified positive instances. *FP*, conversely, refers to the number of negative instances misclassified as positive instances, while *FN* represents the number of positive instances misclassified as negative. The F1-score is calculated as the harmonic mean of *Precision* and *Recall*, where *Precision* indicates the proportion of predicted positives that are truly positive, and *Recall* is the proportion of actual positives that are correctly classified. The F1-score is a more reliable measure of incorrectly classified cases than the *Accuracy* metric.

RESULTS OF CLASSIFICATION METHODS

As mentioned earlier, we evaluated various clustering and classification models on the 10 data files provided for this study (3 λ, 3 μ and 4 σ files). However, most of these models did not perform well, likely due to the low sequence diversity in the files. The sequences with the highest potential for clustering and classification were μ_B and σ_C. These results align with previous studies, showing that these proteins are located in the outer capsid, a region with high sequence diversity [30–32, 35].

The following sections detail the results of classification models developed for these two datasets. Notably, the TERV-labeled sequences were limited in both μ_B and σ_C. Consequently, the classifier models could not accurately classify these sequences, and we excluded them from our classification analysis. The results of clustering models are explained in the supplementary material section (S4 in the supplementary).

This study evaluates the performance of ML and DL models on the μ_B dataset using two data encoding methods: k-mer and

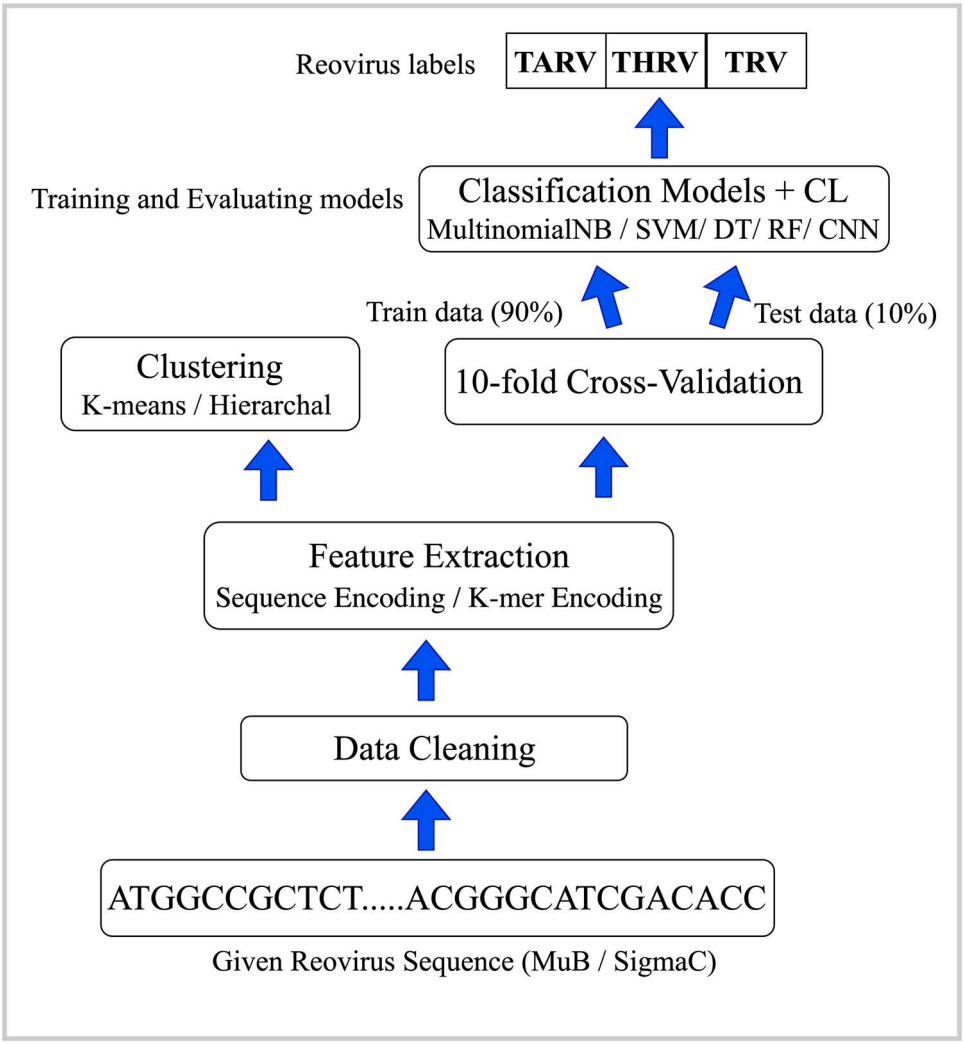

Figure 1. ML workflow for analysis of turkey reovirus sequences.

sequence encoding. The results of these models are presented in Tables 2 and 3.

The outcomes of classifiers utilizing the k-mer encoding method, with and without the CL approach, are discussed in Table 2. Additionally, the results of models employing the sequence encoding method, with and without the CL approach, are presented in Table 3.

Table 2 reveals that MultinomialNB achieved the lowest f1-score (0.67) for TARV reovirus, with the lowest accuracy (0.64) and f1-weighted score (0.63) compared with other methods. On the other hand, SVM achieved the highest f1-score (0.87) among all ML models but struggled to correctly classify THRV and TRV viruses, resulting in a low f1-macro score (0.29) despite the highest accuracy.

Furthermore, DT outperformed SVM and MultinomialNB regarding results, although still falling short of RF. The CL approach's effectiveness in improving RF performance is demonstrated in Table 2. With CL, RF achieved improved f1-scores for all reovirus types (0.86 for TARV, 0.77 for THRV and 0.50 for TRV). This approach also positively impacted overall accuracy and f1-weighted score, raising them to 0.81 and improving the f1-score to 0.71.

Moreover, CNN achieved the best f1-score (0.90) for TARV reovirus and successfully detected THRV and TRV with scores of 0.57 and 0.67, respectively, resulting in the highest f1-macro score (0.71) and accuracy score (0.85) among all methods.

Table 3 presents the results obtained from ML models and the CNN method trained on the μ_B dataset using sequence encoding, with and without utilizing the CL approach.

MultinomialNB achieved the lowest f1-score of 0.70 for TARV reovirus, with accuracy and f1-weighted scores of 0.56 and 0.58, respectively. SVM achieved an f1-score of 0.87 for TARV but had the worst f1-macro of 0.29 due to misclassification of THRV and TRV viruses.

DT achieved the best f1-scores of 0.93 for TARV and 0.86 for THRV, leading to the highest accuracy of 0.92 and the f1-weighted score of 0.92. RF had the highest f1-macro score of 0.83 for TRV, resulting in the highest f1-macro score of 0.85. CNN achieved f1-scores of 0.89, 0.75 and 0.5 for TARV, THRV and TRV, respectively, outperforming MultinomialNB and SVM, but with lower accuracy, f1-score and f1-weighted compared with DT and RF.

Tables 2 and 3 demonstrate the efficacy of different encoding methods, including k-mer and sequence encoding, along with the CL approach to enhance the performance of ML models. Specifically, the best outcomes were achieved when encoding the μ_B dataset with sequence encoding and utilizing the Decision Tree classifier, with an accuracy of 0.92, f1-macro of 0.83 and f1-weighted of 0.92. This is supported by Figure 2, which clearly

Table 2: Comparison of model performance on μ_B dataset with Kmer encoding method

model	F1-score						Overal scores					
	TARV		THRV		TRV		Accuracy		F1-macro		F1-weighted	
	No CL	With CL	No CL	With CL	No CL	With CL	No CL	With CL	No CL	With CL	No CL	With CL
MultinomialNB	0.64	0.67	0.53	0.5	0.67	0.67	0.6	0.6	0.61	0.61	0.62	0.63
SVM	0.87	-	0	-	0	-	0.77	-	0.29	-	0.67	-
Descision Tree	0.81	0.79	0.57	0.57	0.44	0.5	0.68	0.68	0.61	0.62	0.67	0.67
Random forest	0.84	0.86	0.6	0.77	0.44	0.5	0.72	0.81	0.63	0.71	0.7	0.8
CNN	0.90	-	0.57	-	0.67	-	**0.85**	-	**0.71**	-	**0.84**	-

Table 3: Comparison of model performance on μ_B dataset with sequence encoding method

model	F1-score						Overal scores					
	TARV		THRV		TRV		Accuracy		F1-macro		F1-weighted	
	No CL	With CL	No CL	With CL	No CL	With CL	No CL	With CL	No CL	With CL	No CL	With CL
MultinomialNB	0.7	0.5	0.47	0.48	0.4	0.44	0.56	0.48	0.52	0.47	0.58	0.48
SVM	0.87	-	0	-	0	-	0.77	-	0.29	-	0.67	-
Descision Tree	0.95	0.92	0.86	0.89	0.67	0.5	**0.92**	0.88	**0.83**	0.77	**0.92**	0.89
Random forest	0.93	0.92	0.8	0.8	0.83	0.67	0.88	0.88	0.85	0.8	0.88	0.88
CNN	0.89	-	0.75	-	0.5	-	0.84	-	0.71	-	0.82	-

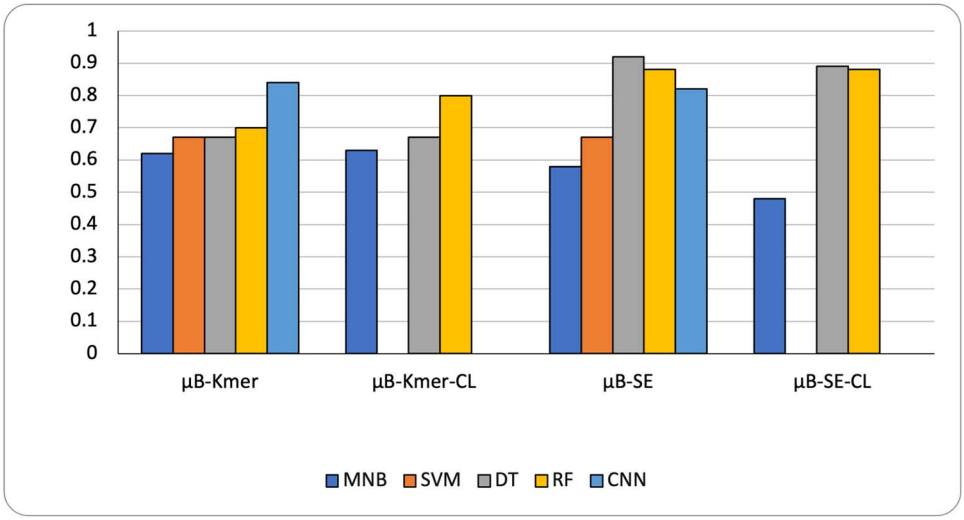

Figure 2. This figure displays F1-weighted scores for ML models trained on the μ_B dataset, comparing k-mer and sequence encoding methods with and without CL.

shows that applying sequence encoding on μ_B and using a random forest classifier produce the best f1-weighted score.

We then evaluated ML and DL models on the σ_C dataset using two data encoding methods: k-mer and sequence encoding. The results are presented in Tables 4 and 5.

Table 5 analyzes the classifier and model performance using k-mer encoding. MultinomialNB achieved the lowest f1-score (0.67) and accuracy (0.64) for TARV reovirus. MultinomialNB without CL had an f1-score of 0.33 for TRV reoviruses, which improved to 0.5 with CL. SVM also struggled to detect THRV and TRV, resulting in the lowest f1-macro score of 0.28.

Applying CL to the DT model significantly improved results, increasing f1-scores for TARV, THRV and TRV. This led to an accuracy score of 0.92, an f1-macro score of 0.83 and an f1-weighted score of 0.92. CL also positively impacted all types of viruses in the RF method, improving accuracy, f1-weighted and f1-macro scores.

Additionally, the CNN method achieved the highest f1-scores for TARV (0.91) and THRV (0.86) among ML models that did not use CL. However, it struggled to classify TRV, yielding a low f1-macro score of 0.59.

Table 4 presents the performance of ML models and the CNN method trained on the σ_C dataset using sequence encoding, with and without the CL approach.

Multinomial NB performed significantly better with sequence encoding than k-mer encoding, achieving f1-scores of 0.87, 0.75 and 0.50 for TARV, THRV and TRV, respectively. This resulted in an accuracy score of 0.8, an f1-macro score of 0.71and an f1-weighted score of 0.8.

Table 4: Comparison of model performance on σ_C dataset with sequence encoding method

model	F1-score						Overal scores					
	TARV		THRV		TRV		Accuracy		F1-macro		F1-weighted	
	No CL	With CL	No CL	With CL	No CL	With CL	No CL	With CL	No CL	With CL	No CL	With CL
MultinomialNB	0.87	0.76	0.75	0.59	0.5	0.5	0.8	0.68	0.71	0.62	0.8	0.69
SVM	0.84	-	0	-	0	-	0.72	-	0.28	-	0.6	-
Descision Tree	0.91	0.94	0.86	0.8	0	0.67	0.88	**0.88**	0.59	**0.8**	0.86	**0.88**
Random forest	0.9	0.89	0.57	0.55	0	0.67	0.8	0.8	0.49	0.7	0.74	0.77
CNN	0.91	-	0.5	-	0.44	-	0.77	-	0.62	-	0.8	-

Table 5: Comparison of model performance on σ_C dataset with Kmer encoding method

model	F1-score						Overal scores					
	TARV		THRV		TRV		Accuracy		F1-macro		F1-weighted	
	No CL	With CL	No CL	With CL	No CL	With CL	No CL	With CL	No CL	With CL	No CL	With CL
MultinomialNB	0.67	0.61	0.7	0.61	0.33	0.5	0.64	0.6	0.57	0.57	0.66	0.6
SVM	0.84	-	0	-	0	-	0.72	-	0.28	-	0.6	-
Descision Tree	0.88	0.97	0.8	0.86	0	0.67	0.84	0.92	0.56	0.83	0.82	0.92
Random forest	0.89	0.94	0.67	0.86	0	1	0.8	**0.92**	0.52	**0.93**	0.75	**0.92**
CNN	0.91	-	0.86	-	0.00	-	0.84	-	0.59	-	0.86	-

However, the utilization of sequence encoding did not significantly impact the performance of the SVM model compared with k-mer encoding. The f1-scores for TARV, THRV and TRV remained the same, as shown in Table 5.

On the other hand, DT with CL showed significant improvements in f1-scores for TARV, THRV and TRV, resulting in the highest accuracy, f1-macro and f1-weighted scores among the tested methods.

The CL method also improved the f1-score of TRV using the RF method but had a negligible impact on TARV and THRV. CNN, like DT, outperformed other methods in terms of TARV but showed inconsistent performance for other reoviruses.

Both Tables 4 and 5 demonstrate that encoding the σ_C dataset using k-mer encoding and applying CL to the Random Forest classifier yielded the highest accuracy, f1-macro and f1-weighted scores. The effectiveness of this approach is further supported by the results shown in Figure 3, highlighting the superiority of using k-mer encoding on σ_C and applying CL with a Random Forest classifier in terms of f1-weighted score.

Furthermore, CL significantly enhances classifier models by effectively managing label noise in datasets. This is evident in Tables 4 and 5, where its application notably improves the outcomes for σ_C datasets using both K-mer and sequence encoding, demonstrating its strength in noise identification and elimination. However, Tables 2 and 3 show its limited impact on μ_B datasets, suggesting fewer noise instances in μ_B compared with σ_C. These findings reveal that CL's effectiveness is more influenced by dataset characteristics than by the choice of feature extraction method.

LIMITATION AND CHALLENGES

In our research journey, we encountered several noteworthy challenges, one of the most prominent being the significant class imbalance within our dataset. Specifically, viral types TARV, THRV, TRV and TERV were unevenly represented in a ratio of 5:3:3:0.5, potentially introducing bias into our model. To tackle this issue, we explored various methods like oversampling, undersampling and Synthetic Minority Oversampling Technique (SMOTE) to balance the classes, as suggested by Blagus and Lusa (2013) [36].

Our study implemented SMOTE exclusively on the training data, aiming to create synthetic samples for minority classes like TRV and THRV to match the TARV class regarding sequence count. However, when evaluating our trained model on the test data, we encountered unsatisfactory results, prompting us to withhold reporting these outcomes.

Additionally, the limited number of sequences available in each dataset presented another significant challenge, potentially affecting the accuracy of our classification model.

CONCLUSION

In this project, we addressed the problem of detecting and classifying different types of turkey reoviruses in a given genome database. We employed various ML and DL algorithms on datasets of two segments (M2- μ_B and S1-σ_C) and evaluated their performance using commonly used metrics. We also used both k-mer and sequence encoding to examine their impact on the performance of the employed ML and DL models.

The main contributions of this research are summarized as follows:

- This work is the first, to the best of our knowledge, to apply various ML methods such as SVM, Random Forest Classifier, Multinomial Naive Bayes and Decision Tree Classifier, as well as a DL method known as CNNs to detect and classify different types of reoviruses in turkey genome sequences.
- K-means and Hierarchical clustering methods are adopted to identify new clusters and partition the data into novel groups that exhibit high similarity within their respective clusters.

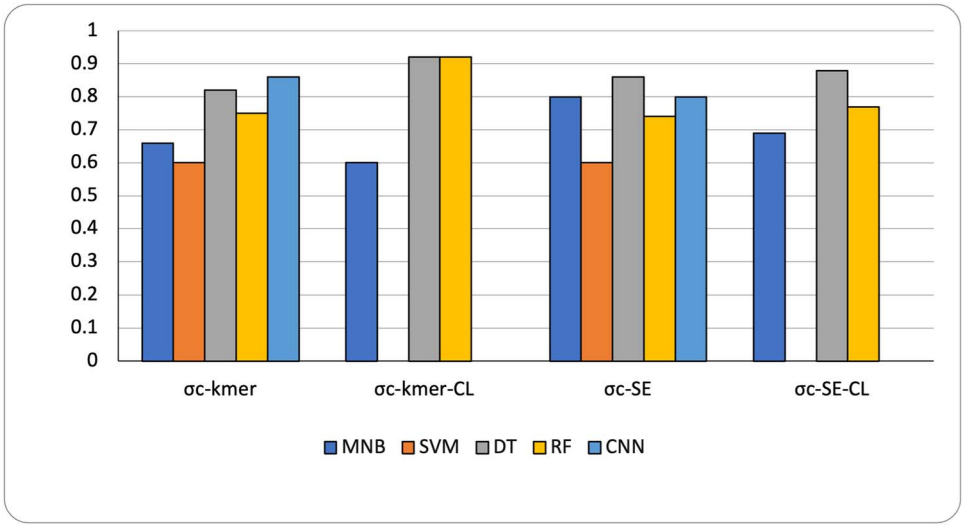

Figure 3. This figure displays F1-weighted scores for ML models trained on the σ_C dataset, comparing k-mer and sequence encoding methods with and without CL.

The results of this study reveal that applying Decision Tree classifiers with sequence encoding for μ_B datasets provides superior performance compared with other methods, achieving classification with accuracy, f1-macro and f1-weighted scores of 0.92%, 0.83% and 0.92%, respectively. Meanwhile, for σ_C, Random Forest with k-mer encoding yielded the best results with accuracy, f1-macro and f1-weighted scores of 0.92%, 0.93% and 0.92%, respectively. Our research clearly illustrates that the effectiveness of ML models is significantly influenced by the specific characteristics of the dataset and the methods used for feature extraction. This understanding is crucial for the development of more accurate and efficient ML applications in biological data analysis.

Furthermore, our study highlights the promising capabilities of both ML and DL algorithms in accurately detecting and classifying turkey reoviruses. These results not only validate the effectiveness of these computational approaches but also open new avenues for advanced research in virology and disease control. The success of these algorithms in our study serves as a stepping stone for further explorations and innovations in the field of viral genomics.

Key Points

- The turkey reoviruses, a subtype of Avian reoviruses, are linked to diverse turkey diseases, causing economic losses and diminishing the marketability of commercial turkeys.
- Our study is the first known application of Machine Learning for detecting and classifying different types of turkey reoviruses.
- We utilize clustering methods to identify novel clusters and unidentified virus variants within turkey genome sequences.

FUNDING

This research received no external funding.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/

AUTHOR CONTRIBUTIONS

All authors have read and agreed to the published version of the manuscript.

CODE AVAILABILITY

All codes and dataset are publicly accessible via the following GitHub repository: https://github.com/abdeltawab/ReoVirus.

REFERENCES

1. Huaguang L, Tang Y, Dunn PA, *et al.* Isolation and molecular characterization of newly emerging avian reovirus variants and novel strains in Pennsylvania, Usa, 2011–2014. *Sci Rep* 2015;**5**(1):14727.
2. Ayalew LE, *et al.* The dynamics of molecular evolution ofemerging avian reoviruses through accumulation of pointmutations and genetic re-assortment. *Virus evolution* 2020;**6**(1):veaa025.
3. Varela R, Benavente J. Protein coding assignment of avian reovirus strain s1133. *J Virol* 1994;**68**(10):6775–7.
4. Martínez-Costas J, Grande A, Varela R, *et al.* Protein architecture of avian reovirus s1133 and identification of the cell attachment protein. *J Virol* 1997;**71**(1):59–64.
5. Jones RC. Avian reovirus infections. *Rev Sci Tech* 2000;**19**(2):614–25.
6. Levisohn S, Gur-Lavie A, Weisman J. Infectious synovitis in turkeys: isolation of tenosynovitis virus-like agent. *Avian Pathol* 1980;**9**(1):1–4.
7. Mor SK, Sharafeldin TA, Porter RE, *et al.* Isolation and characterization of a Turkey arthritis Reovirus. *Avian Dis* 2012;**57**(1):97–103.
8. Sharafeldin TA, Mor SK, Sobhy NM, *et al.* A newly emergent Turkey arthritis reovirus shows dominant enteric tropism and induces significantly elevated innate antiviral and t helper-1 cytokine responses. *PloS One* 2015;**10**(12):1–12.

9. Kumar R, Sharafeldin TA, Sobhy NM, *et al*. Comparative pathogenesis of Turkey reoviruses. *Avian Pathol* 2022;**51**(5):435–44 PMID: 35583932.
10. French David. Incidence and economic impact of reovirus in the poultry industries in the united states. *Avian Diseases* 2022;**66**(4):432–434.
11. Gunasekaran H, Ramalakshmi K, Rex Macedo Arokiaraj A, *et al*. Analysis of dna sequence classification using cnn and hybrid models. *Comput Math Methods Med* 2021;**2021**: 1–12.
12. Li J, Wei L, Zhang X, *et al*. Dismir: deep learning-based noninvasive cancer detection by integrating dna sequence and methylation information of individual cell-free dna reads. *Brief Bioinform* 2021;**22**(6):bbab250.
13. Danilevsky, Polsky AL, Shomron N. Adaptive sequencing using nanopores and deep learning of mitochondrial dna. *Brief Bioinform* 2022;**23**(4):bbac251.
14. Shen, Liu Y, Song J, Yu D-J. Saresnet: self-attention residual network for predicting dna-protein binding. *Brief Bioinform* 2021;**22**(5):bbab101.
15. Aromolaran O, Aromolaran D, Isewon I, Oyelade J. Machine learning approach to gene essentiality prediction: a review. *Brief Bioinform* 2021;**22**(5):bbab128.
16. Zhang, Liu Y, Xu J, *et al*. Leveraging the attention mechanism to improve the identification of dna n6-methyladenine sites. *Brief Bioinform* 2021;**22**(6):bbab351.
17. Gwak H-J, Rho M. Vibe: a hierarchical bert model to identify eukaryotic viruses using metagenome sequencing data. *Brief Bioinform* 2022;**23**(4):bbac204.
18. Sherkatghanad, Abdar M, Charlier J, Makarenkov V. Using traditional machine learning and deep learning methods for on- and off-target prediction in crispr/cas9: a review. *Brief Bioinform* 2023;**24**(3):bbad131.
19. Zou Q, Lin G, Jiang X, *et al*. Sequence clustering in bioinformatics: an empirical study. *Brief Bioinform* 2020;**21**(1):1–10.
20. Hamamoto R, Takasawa K, Shinkai N, *et al*. Analysis of super-enhancer using machine learning and its application to medical biology. *Brief Bioinform* 2023;**24**(3):bbad107.
21. Wei Jiang Q, Wei Y, Wang S. A novel hierarchical clustering algorithm for gene sequences. *BMC Bioinformatics* 2012;**13**(1): 1–15.
22. Nguyen, Tran VA, Ngo DL, *et al*. Dna sequence classification by convolutional neural network. *J Biomed Sci Eng* 2016;**09**(5):280–6.
23. Machuve D, Nwankwo E, Mduma N, Mbelwa J. Poultry diseases diagnostics models using deep learning. *Front Artif Intell* 2022;**5**:733345.
24. Mbelwa Hope, Machuve Dina, Mbelwa Jimmy. Deepconvolutional neural network for chicken diseases detection. **12**(2), 2021.
25. Whata A, Chimedza C. Deep learning for sars cov-2 genome sequences. *IEEE Access* 2021;**9**:59597–611.
26. de Souza LC, Azevedo KS, de Souza JG, *et al*. New proposal of viral genome representation applied in the classification of sars-cov-2 with deep learning. *BMC Bioinformatics* 2023;**24**(1):1–19.
27. Cho, *et al*. Machine learning in dna microarray analysis forcancer classification. In Proceedings of the First APBCon Bioinformatics 2003-Volume 19, pages 189–198, AUS, 2003. Australian Computer Society, Inc.
28. Eickholt J, Cheng J. Dndisorder: predicting protein disorder using boosting and deep networks. *BMC Bioinformatics* 2013;**14**(1):1–10.
29. Yang A, Zhang W, Wang J, *et al*. Review on the application of machine learning algorithms in the sequence data mining of dna. *Front Bioeng Biotechnol* 2020;**8**:1032.
30. Mor SK, Marthaler D, Verma H, *et al*. Phylogenetic analysis, genomic diversity and classification of m class gene segments of Turkey reoviruses. *Vet Microbiol* 2015;**176**(1–2):70–82.
31. Ayalew LE, Ahmed KA, Mekuria ZH, *et al*. The dynamics of molecular evolution of emerging avian reoviruses through accumulation of point mutations and genetic re-assortment. *Virus. Evolution* 2020;**6**(1):veaa025.
32. Kovács E, Varga-Kugler R, Mató T, *et al*. Identification of the main genetic clusters of avian reoviruses from a global strain collection. *Front Vet Sci* 2023;**9**:1094761.
33. Souvorov A, Agarwala R, Lipman DJ. Skesa: strategic k-mer extension for scrupulous assemblies. *Genome Biol* 2018;**19**(1):153.
34. Bergstra James. Bengio Yoshua. Random searchfor hyper-parameter optimization. *Journal of MachineLearning Research* 2012;**13**(10):281–305.
35. Egaña-Labrin S, Hauck R, Figueroa A, *et al*. Genotypic characterization of emerging avian reovirus genetic variants in California. *Sci Rep* 2019;**9**(1):9351.
36. Blagus R, Lusa L. Smote for high-dimensional class-imbalanced data. *BMC Bioinformatics* 2013;**14**:1–16.