Pakistan Veterinary Journal

ISSN: 0253-8318 (PRINT), 2074-7764 (ONLINE)
DOI: 10.29261/pakvetj/2026.076

RESEARCH ARTICLE

K-mer-based Machine Learning Method to Classify *Echinococcus granulosus S.L.* Mitochondrial DNA Sequences

Enas Al-khlifeh[1*], Ahmad B. Hassanat[2], Suleyman A. AlShowarah[2,3], Ahmad S. Tarawneh[2], Lujain A Alhasanat[4] and Awni Hammouri[2]

[1]Applied Biology Department, Faculty of Science, Al-Balqa Applied University, Al-Salt, Jordan, [2]Faculty of Information Technology, Mutah University, Al-Karak, Jordan, [3]Software Engineering Department, Faculty of Science and Information Technology, Al-Zaytoonah University of Jordan, Jordan, [4]Faculty of Medicine, Mutah University, Karak, 61710, Jordan
*Corresponding author: Al-khlifeh.en@bau.edu.jo

ARTICLE HISTORY (26-085)

Received: January 25, 2026
Revised: March 13, 2026
Accepted: March 15, 2026
Published online: March 19, 2026

Key words:
E. granulosus
GenBank
k-mer
Machine learning
mtDNA

A B S T R A C T

There are ongoing debates regarding the taxonomy of the *E. granulosus Echinococcus granulosus sensu lato* (*s.l.*) complex, with the species status of genotypes G1/G3 (*E. granulosus s.s.*) and the G6–G10 cluster (*E. canadensis*) being particularly contested. To solve this challenge, we develop a machine learning (ML)-based *k*-mer method to predict genotypes of *E. granulosus s.l.* via the mitochondrial DNA sequence data available in GenBank. We used 7-mer sequences that represent mtDNA to reveal untapped diversity across 948 sequences of seven known genotypes of *E. granulosus s.l.* We evaluated the utility of varying *k*-mer lengths, including fixed vs adaptive lengths, for sequence comparisons. Principal component analysis (PCA) identified the most discriminative patterns, addressed the computational challenges posed by high-dimensional features and fed them into 5 distinct ML algorithms, including both conventional (LR, RF and SVM) and DL (1D-CNN and the LSTM) methods. Moreover, crucial insights into performance implications and efficiency improvements have been emphasized via 10-fold cross-validation. Our method attains approximately 95% accuracy in predicting *E. granulosus s.l.* genotypes. High genetic diversity within G6, G8 and G3 contributes significantly to the controversial taxonomy of the *E. canadensis* cluster and *E. granulosus s.s.*, respectively. Moreover, ML can potentially compete with BLAST (the NCBI sequence alignment tool) when the BLAST-Similarity-KNN classifier is implemented on the *E. granulosus* mtDNA data. This study provides a novel approach for classifying *E. granulosus s.l.* at the species level, thereby supporting disease control activities.

To Cite This Article: Al-khlifeh A, Hassanat AB, AlShowarah SA, Tarawneh AS, Alhasanat LA and Hammouri A, 2026. *K*-mer-based machine learning method to classify *Echinococcus granulosus S.L.* mitochondrial DNA sequences. Pak Vet J, 46(4): 816-829. http://dx.doi.org/10.29261/pakvetj/2026.076

INTRODUCTION

E. granulosus s.l., a member of the Taeniidae family, is a cosmopolitan tapeworm known to cause the zoonotic disease cystic echinococcosis (CE), also known as hydatidosis. *E. granulosus* is distributed worldwide (Alkhlifeh *et al.*, 2023; Borhani *et al.*, 2024). Owing to its significantly high morbidity rate, CE was placed on the World Health Organization (WHO) roadmap for neglected tropical diseases (NTDs) from 2021-2030 (Neglected tropical diseases - GLOBAL).

E. granulosus s.l. is currently recognized as a complex of cryptic species with different levels of selectivity toward intermediate hosts. The term "cryptic" arises from differences in physical and developmental patterns, along with genetic variability within a species (Kinkar *et al.*, 2018a). The key terminologies for *E. granulosus s.l.* species have been established mostly via the mtDNA sequence, specifically the *cox1* partial segment (~ 366bp), in addition to the complete or nearly complete mtDNA.

The most common *E. granulosus* genotype worldwide is *E. granulosus sensu stricto* (*s.s.*) G1–G3. These two genotypes were initially differentiated via the criteria established by Bowles *et al.* (1994) via a small 366bp segment of the *cox1* gene (Bowles *et al.*, 1994). There were two designated places where it varied. However, additional investigations have shown that a longer mtDNA sequence can be used to more accurately distinguish between these two closely related genotypes (Kinkar *et al.*, 2018a). A few studies have indicated a noticeably improved phylogenetic

resolution using larger mtDNA sequences of ~8270 bp (Kinkar *et al.*, 2016; Laurimäe *et al.*, 2023). However, even in these studies, a larger dataset is needed to confirm these distinctions.

E. granulosus s.l. and *E. canadensis* (G6-G10) have unresolved taxonomic issues (Casulli *et al.*, 2022). A disparity in the genetic relatedness between the G6/G7 and G8/G10 genotypes was found in the mtDNA data (see Laurimäe *et al.*, 2023) for revision). Moreover, there was no greater benefit to evidence from nuclear loci. For example, a study that employed two nuclear markers revealed that the G6/G7 and G8/G10 groupings had common alleles (Saarma *et al.*, 2009), whereas research using six nuclear loci indicated that cervids G8 and G10 belong to one lineage, whereas G6/G7 belongs to another (Lymbery *et al.*, 2015; Laurimäe *et al.*, 2023; Malik *et al.*, 2024).

Molecular studies of the *E. granulosus s.l.* coupled mtDNA barcode with population genetic indices, signifies adaptation to a local environment and indicated a parasite's geographic origin (Alvarez Rojas *et al.*, 2014; Al-khlifeh *et al.*, 2024). Nevertheless, the limited coverage of the sites under study (Kinkar *et al.*, 2018b), the low analytical power, variation in sequence length between the same or different investigations, and the lack of genetic resources due to inadequate research funding are among the challenges that still need to be addressed.

A growing number of publicly available mtDNA sequence for *E. granulosus s.l.* provide an opportunity to identify and validate new sequence variations, with GenBank serving as a key internet resource for accessing these data. Nevertheless, the primary challenge at hand remains the high dimensionality of the data, such as sorting through the massive number of known [and unknown] differences among the sequences and using this information to forecast genotype classification. On the other hand, ML and deep learning (DL) are well suited for genotype classification using large amounts of sequence data because their capacity to analyze high-dimensional information efficiently can reveal patterns that traditional experimental methods miss (Deelder *et al.*, 2022; Asif *et al.*, 2024). ML models learn from large datasets to predict or classify, thereby augmenting traditional experimental DNA analysis by handling the scale and complexity of sequence data while working in concert with available molecular biology approaches (Al-khlifeh *et al.*, 2025; Tarawneh *et al.*, 2025).

The proliferation of available sequencing data has coincided with a change in the focus of categorization techniques. The most useful tools in the past were alignment-based techniques. Currently, the most intriguing classification strategy in terms of speed and accuracy is based on sliding window approaches that count *k*-mers. *K*-mer analysis classifies genotypes and species by comparing short, fixed-length DNA subsequences to identify patterns unique to specific genomes or individuals (Bussi *et al.*, 2021). In species classification, *k*-mers reveal evolutionary relationships and microbial composition by showing which *k*-mers are common in *Archaea* but rare in others (Bize *et al.*, 2021) and by identifying SARS-CoV-2 variants (Ali *et al.*, 2021). For genotype classification, *k*-mer analysis reveals genomic variation by identifying changes in *k*-mer presence or frequency (Vinje *et al.*, 2015), indicating single nucleotide variants (SNVs) or larger structural differences such as insertions and deletions (Shajii *et al.*, 2016). However, their application in the helminthology field is limited.

The goal of this study is to revise the classification of *E. granulosus s.l.* via ML and DL techniques and a substantial set of sequences from GenBank by identifying the discriminative *k*-mers.

MATERIALS AND METHODS

Data collection and processing of raw sequencing data: We acquired publicly accessible raw mtDNA sequencing data and genotypes from previously published research on *E. granulosus s.l.* from the GenBank repository, which can be previewed in (https://doi.org/10.6084/m9.figshare.31563277). Using careful criteria, we verified the genotype and species from the information we obtained directly from the sequence "description" and verified these details by reviewing the literature. The sequences used in this study included the *cox1* gene and complete and incomplete mtDNA. We rationalized our choice by pointing out that these are the most common sequence information types for *E. granulosus s.l.* and by evaluating the ability of ML algorithms to handle sequences of varying lengths, identifying patterns and similarities despite size inequalities.

Data partitioning and quality control: We collected 948 unique mitochondrial DNA sequences of different *E. granulosus s.l.* genotypes: G1 (269), G3 (271), G5 (85), G6 (138), G7 (118), G8 (43), and G10 (24). We used Python and genetic libraries, notably Biopython's Entrez utility. Basic sequence analysis, including nucleotide count and GC content, was also performed via the same program.

To make sure that our evaluation was fair and free of data leakage, we put several quality control steps in place. These include:

1) Duplicates removal: We checked each sequence by its base accession number. This ensured that every biological sample appeared only once in our data.

2) Data splitting: Using the well-known stratified random sampling, we divided the dataset into a training set: 80% of the data (758 sequences), and a test set: 20% (190 sequences). This means that we kept the same proportion of each genotype in both sets. See more details in the supplementary (Fig. S1:. 10.6084/m9.figshare.31636462).

3) Data normalization: We converted raw *k*-mer counts into relative frequencies. We did this by dividing each count by the total number of *k*-mers in the sequence. This step automatically handled any differences in sequence length. Then, we applied a MinMaxScaler. This brought all the features into a range between 0 and 1. Such scaling prevents classifiers from being biased by larger numbers in some features. For our cross-validation experiments, we made sure to fit the scaler only on the training folds. We then applied it to the test folds. This extra step is important to avoid any information leakage from the test data.

Data Representation: In order to identify the best feature extraction approach for ML classification, we investigated three different methods for representing DNA sequence data:

***K*-mer frequency representation:** Because our data have a wide range of sequence sizes in the range of [309, 17675], *k*-mer analysis becomes vital as it can identify local sequence patterns and motifs that may be genotype-discriminative without the need for sequence alignment. Here, the sequences are decomposed into overlapping *k*-mers (step=1 nucleotide) of length k, generating 4^k possible *k*-mer combinations. For each sequence, *k*-mer frequencies were calculated and normalized via min–max scaling (0–1 range) to account for sequence length variations. However, the choice of (*k*) in *k*-mer analysis is a critical decision with significant implications for the analysis of biological sequences.

We experimentally determined the optimal value of (*k*) for our data. By 'optimal', we mean the value that provides the richest information while minimizing data sparsity, computational complexity, and memory usage. For example, a small *k*, e.g., (*k*=1) or *k*=2), such as "AA", "AT", "AG", and "AC", are common across many sequences and do not provide much unique information. On the other hand, a large *k* (e.g., (*k*=10) or more) can lead to a high-dimensional feature space, which leads to a sparse data problem, where the number of potential *k*-mers grows exponentially, leading to many *k*-mers that may not appear in the dataset. Additionally, large *k* values result in greater memory usage, as the representation of the data becomes more complex.

For a sequence S of length n, the number of *k*-mers generated is as follows:

$$N_{kmers} = n - k + 1$$

K-mer frequency calculation for a *k*-mer m in sequence S:

$$f(m, S) = \frac{Count(m, S)}{N_{kmers}}$$

where count(m,S) is the occurrence count of *k*-mer m in sequence S.

We also explored fixed-length truncation and adaptive length selection approaches for DL, but these approaches yielded substantially lower performance.

We ran several checks to confirm that our normalization method removed any influence of sequence length, which varied widely from 309 to 17,675bp. First, we trained classifiers using only sequence length as the input feature and found that length alone could not predict genotype. We also calculated the Pearson correlation between each *k*-mer feature and sequence length across the whole dataset. Finally, we evaluated model performance on three separate length groups—short (<1,000bp), medium (1,000–5,000bp), and long (>5,000bp). All these tests confirmed that the model's discriminative power was independent of fragment size, proving our normalization worked. See Results, Section: Sequence length bias control.

Machine Learning Models: To find the best ML model for our data, we evaluated a number of ML methods for each data representation approach:

1- Traditional Machine Learning (*K*-mer Features): These methods were evaluated primarily on the basis of *k*-mer features, as this representation and feature extraction method provides a well-structured and fixed-size feature vector that is suitable for traditional ML approaches (Imam *et al.*, 2021), such as the following:

- **Logistic Regression (LR)**: Linear baseline with L2 regularization.
- **Random forest (RF)**: Ensemble method with 200 estimators adequate for handling high-dimensional *k*-mer features.
- **Support Vector Machine (SVM)**: Linear kernel with probability estimation enabled

2- Deep learning architectures (sequence-based features): DL approaches are well known for their effectiveness, particularly with sequence-based features. For the purpose of this investigation, we utilized two of these approaches:

1D Convolutional Neural Network (1D-CNN):
Architecture:
- Conv1D layers (64 filters, kernel size 3) for LFD
- MaxPooling1D for dimensionality reduction
- Dropout layers (0.3) for regularization
- Dense layers (64 units) for classification
- Softmax output for multiclass prediction

Long short-term memory:
Architecture:
- LSTM layer (64 units) for sequential pattern learning
- Dropout (0.3) for regularization
- Dense output layer

All the neural networks were trained via the Adam optimizer and sparse categorical cross-entropy loss, with early stopping (patience=10) and learning rate reduction (factor=0.5, patience=5) implemented to prevent overfitting and enhance convergence. For hyper parameter tuning and model selection, a 20% validation split from the training data was utilized, whereas the final evaluation was conducted on a held-out test set to ensure unbiased performance estimates.

Having sample sizes ranged from 24 for G10 to 271 for G3, we carried out a sensitivity analysis to address such class imbalance problem in our dataset. We wanted to see if using class weights would improve how well our models handled the smaller genotypes. So, we tested all of the three classifiers again, but this time with inverse-frequency class weights using scikit-learn's balanced mode. This approach assigns higher penalties for mistakes on the minority classes, making the model pay more attention to them. We then compared these results to the default unweighted configuration from our main analysis. This comparison helped us figure out whether poor performance on rare genotypes like G10 and G8 was due to the class imbalance itself or something about the biology of those sequences. See the results, Section: Per-genotype classification performance and error analysis.

Statistical evaluation and confidence intervals: We evaluated model performance using 10-fold stratified cross-validation. To keep everything reproducible, we set a fixed random seed (seed=42) for all runs. For each model, we calculated the mean accuracy and the weighted F1-score across all folds. We also computed 95% confidence intervals for these metrics using the t-distribution with 9 degrees of freedom. We wanted to know if the performance differences between models were statistically significant or not. So, we applied McNemar's test with continuity correction to the

cross-validated predictions for each pair of models. This test helped us determine whether one model truly outperformed another or if the differences were just due to chance.

To further validate our findings, we used SHAP analysis and Logistic Regression coefficients, to see if the most important features identified by our model had real biological meaning. We first identified the top 20 *k*-mers for each genotype based on absolute coefficient values. This gave us 140 *k*-mers in total. We then mapped each of these 7-mers back to the *E. granulosus* G1 reference mitochondrial genome. We used GenBank accession AF297617, which is 13,588 base pairs long. For each *k*-mer, we searched for its exact match in the reference sequence and recorded its position.

Following, we checked where these positions fell within the genome. We compared them against the annotated boundaries for all 12 protein-coding genes, including *cox1* through *nad4L* and *atp6*. We also looked at the two ribosomal RNA genes, *rrnL* and *rrnS*, and the non-coding region. This mapping procedure helped us answer a key question: Were the discriminative patterns that our ML model found located in biologically meaningful regions, or were they just random noise?

Choosing the optimal k for the *k*-mer: To determine the optimal *k*-mer length for *E. granulosus s.l.* genotype classification, we systematically evaluated the performance of logistic regression across k values ranging from 1-10.

Each *k*-mer representation underwent a two-step normalization process: first, raw counts were converted to relative frequencies to normalize for sequence length, followed by min–max scaling to enhance classifier performance. Logistic regression was chosen for this analysis because of its lightweight nature and speed, as it does not rely on complex methods or extensive computational resources, making it particularly suitable for high-dimensional data such as *k*-mers. The total number of possible *k*-mers for nucleotide sequences is $D_k = 4^k$.

The min–max normalized frequency for a single *k*-mer (m) in sequence S is calculated via

$$f_{norm}(m, S) = \frac{f(m, S) - \min(f(m))}{\max(f(m)) - \min(f(m))}$$

where f(m) is the general function that obtains the frequency of any *k*-mer m for all sequences in the dataset.

As shown in Fig. 1, the analysis revealed distinct performance trends as the *k*-mer length increased. For low-complexity *k*-mers (*k*=1-3), the performance was suboptimal (as expected), with accuracies ranging from 40.5% (*k*=1) to 76.8% (*k*=3). The 4-dimensional feature space at *k*=1 was extremely insufficient for capturing discriminative genetic patterns, whereas *k*=2 and *k*=3 showed gradual improvement but still fell short of acceptable thresholds for reliable genotype classification.

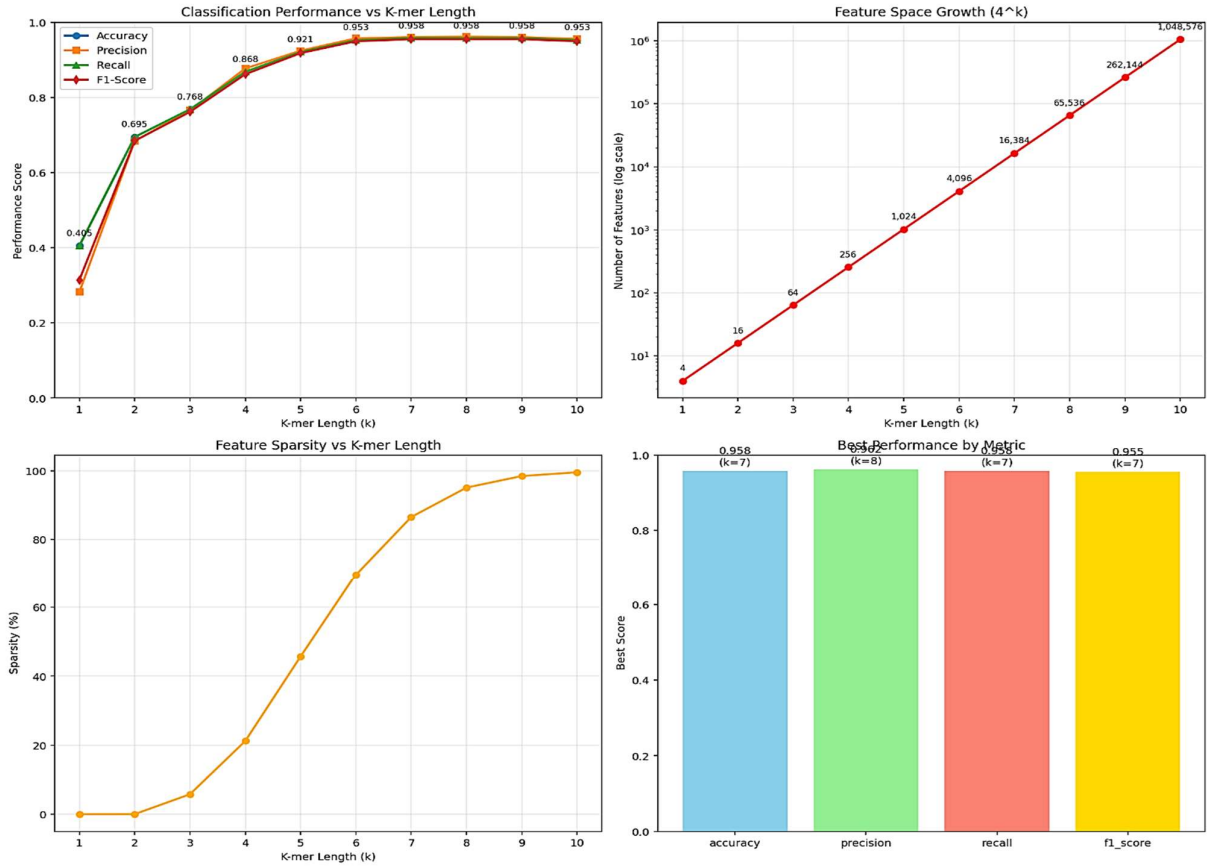


Fig. 1: *E. granulosus s.l.* genotype classification via various k values for *k*-mer analysis. The panels display (top left) classification performance (accuracy, precision, recall, and F1 score) versus *k*-mer length; (top right) growth of the feature space with increasing k; (bottom left) feature sparsity as a function of *k*-mer length; and (bottom right) comparison of performance metrics (accuracy, precision, recall, and F1 score) for the best-performing k values.

In contrast, moderate-complexity *k*-mers (*k*=4-6) exhibited substantial performance gains, with accuracy increasing from 86.8% (*k*=4) to 95.3% (*k*=6). The feature space expanded significantly from 256 to 4,096 dimensions, and the sparsity increased from 21.2% to 69.6%. Notably, *k*=6 achieved excellent performance, with 95.3% accuracy and a 95.0% F1 score, while maintaining reasonable computational efficiency.

However, for high-complexity *k*-mers (*k*=7-10), the performance plateaued in the range of 95.3% to 95.8%, with *k*=7, *k*=8, and *k*=9 achieving the highest accuracy of 95.8%. Despite this high accuracy, the feature space expanded exponentially, growing from 16,384 to over 1 million dimensions, with extreme sparsity reaching 99.6% for *k*=10. This indicates potential overfitting and computational inefficiency, suggesting that while higher *k* values can enhance accuracy, they may also complicate model training and application.

Sparsity is a measure of how empty a dataset is, and the sparsity of the *k*-mer representation can be calculated via

$$\text{Sparsity} = \frac{(D_k - N_{observed})}{D_k} \times 100\%$$

where $N_{observed}$ is the number of unique *k*-mers observed in the dataset.

Accordingly, choosing (*k*=6) or (*k*=7) is optimal for our analysis. The feature vector size is significantly reduced with (*k*=6), whereas (*k*=7) offers slightly better results but presents the challenge of a significantly larger feature vector compared with (*k*=6). However, this issue can be addressed through dimensionality reduction techniques, making (*k*=7) an attractive option for further analysis. This selection strikes a balance between biological interpretability, computational efficiency, and classification accuracy, making it suitable for both research applications and potential clinical implementation. The 7-mer approach effectively captures enough sequence context to differentiate between closely related *E. granulosus s.l.* genotypes while minimizing the risk of overfitting associated with higher-order *k*-mers.

Dimensionality reduction and principal component analysis (PCA): We employed PCA to address the computational challenges posed by high-dimensional *k-mer* features (16,384 features for k=7). PCA successfully compresses the feature space to 95 components while retaining 99.0% of the original variance, resulting in a 99.4% reduction in memory usage. The first three components (PC1-PC3) explained 77.7% of the total variation, with G7 showing the greatest separation from the other genotypes. This dimensionality reduction enabled more efficient downstream analysis while preserving biological information.

We took care to prevent information leakage in all our cross-validation experiments involving PCA. For each fold, we fitted PCA only on the training data. We then used that same transformation on the test data. This ensured that no information from the test set influenced the PCA components. Across all folds, we kept 95 principal components, as it preserved 98.99% of the total variance on average.

BLAST comparison: To compare our results with those of the benchmark software commonly used in DNA comparison research, we utilized NCBI BLAST+, which is the most widely adopted tool for this purpose. To ensure a valid comparison, we used the similarity scores calculated via BLAST as a distance metric for a KNN classifier. With the same training and testing datasets, each sequence in the test set is compared via BLAST against all sequences in the training set. The nearest k sequences were then used to determine the genotype of the test sample.

Reproducibility and Implementation: We carried out all our analyses using Python 3.10. For ML, we used scikit-learn. We handled biological sequence data with Biopython. For standard scientific computing tasks, we relied on NumPy, SciPy, and Pandas. We created our figures using Matplotlib and Seaborn. For DL models, specifically the 1D-CNN and LSTM, we used TensorFlow with Keras. In order to make sure that our results were reproducible, we set all random seeds to 42.

For the traditional ML models, we used the following settings. We set Logistic Regression with C=1.0, an L2 penalty, and a maximum of 5000 iterations. We used Random Forest with 200 estimators. For the linear SVM, we used LinearSVC with a maximum of 10,000 iterations. During the revision process, we added several new analyses based on reviewer feedback. These included calculating confidence intervals, checking for length bias, testing class weighting, verifying our PCA approach, and mapping important *k*-mers to genes. We ran all these additional checks using the three traditional ML models. These were the models that reviewers focused on in their comments.

RESULTS

***E. granulosus s.l.* genotype and isolate sequence data:** A comprehensive analysis of the entire dataset revealed high genetic diversity across *E. granulosus s.l.* genotypes, with sequence lengths ranging from 309 to 17,675bp (median: 1,608bp). ANOVA revealed significant variations in sequence length among the genotypes ($F=68.93$, $P<0.001$). G7 had the largest mean sequence length (11,666±4,268 bp), whereas G1 and G5 had shorter, more uniform lengths (1,641±3,261bp and 1,473±2,687bp, respectively). GC content analysis revealed substantial heterogeneity among the genotypes (32.0–37.0%; $F=223.72$, $P<0.001$). All the genotypes presented typical mitochondrial AT-richness (65%).

***K*-mer frequency analysis (*k*=7) results:** Using k=7, we generated 16,384 potential *k*-mers combinations per sequence. The resulting sparse matrix (86.5% zeros) indicated genotype-specific patterns. The analysis revealed a distinct hierarchy of genotype consistency: G7 is a well-defined genetic entity with little internal diversity; G1 and G5 exhibit moderate consistency, indicating stable genotypes with some population structure; and G3, G6, and G8 present traits that may be suggestive of taxonomically unresolved groups that need more research. Genotypes with high internal diversity (G3, G6, and G8) pose more difficulties for ML algorithms and may necessitate alternative feature engineering or hierarchical classification techniques, whereas genotypes with high within-class consistency (G7) are simpler to classify correctly (Fig. 2). Moreover, we provide validation and assess whether sequences allocated to the same genotype display similar *7-mer* patterns across the whole population (Table 1).

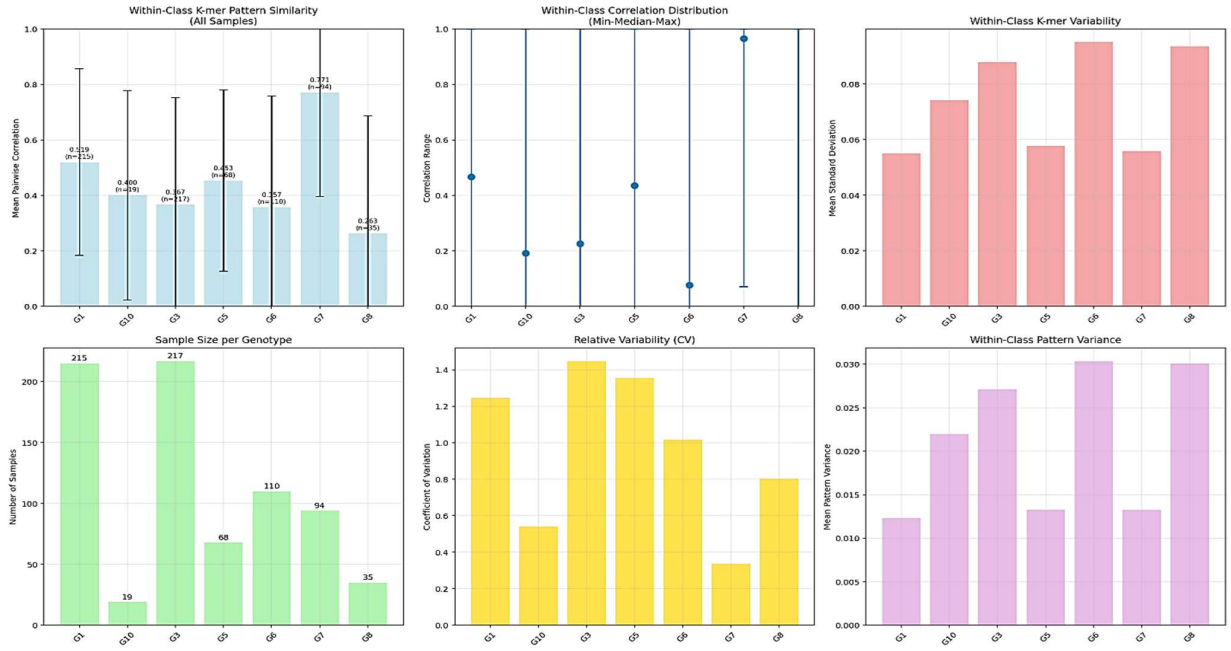


Fig. 2: Analysis of 7-mer features across *E. granulosus s.l.* genotypes. The left column displays (top) the mean 7-mer pattern similarity across all samples, with error bars representing the standard deviation, and (bottom) the within-class correlation distribution (mean–median–max). The right column shows (top) the variability of 7-mer features within each class, (middle) the sample size per genotype, and (bottom) the relative variability (coefficient of variation) of 7-mer features. This visualization highlights the discriminative power and stability of 7-mer patterns among different genotypes.

Table 1: Classification results (weighted average) of *E. granulosus s.l.* genotypes via 7-mer normalized frequency analysis as a feature vector.

Genotype	N Samples	Mean Correlation	Std Correlation	Median Correlation	Mean Variability	Coefficient Variation
G1	215	0.5195	0.3368	0.467	0.054987	1.244799
G10	19	0.4004	0.3775	0.1912	0.074082	0.539611
G3	217	0.3672	0.384	0.2253	0.087962	1.447673
G5	68	0.4528	0.3276	0.4359	0.057846	1.355544
G6	110	0.3571	0.4019	0.0761	0.095129	1.016294
G7	94	0.7711	0.376	0.9649	0.055794	0.335763
G8	35	0.2626	0.4228	-0.0114	0.093581	0.802674

The value k=7 represents the optimal trade-off between discriminative power (classification accuracy plateaued at k=7) and computational tractability (4^7 = 16,384 features versus 4^8 = 65,536 at k=8). Higher values of k produced no meaningful improvement in accuracy while substantially increasing feature dimensionality and computational cost.

Application of genotype-classification models: We evaluated five ML algorithms for genotype classification using 7-mer frequency features. Table 2 shows the performance with 95% confidence intervals from 10-fold stratified cross-validation. All of the three traditional ML models achieved almost similar and high performance as follows: LR achieved 94.41% average accuracy in the range [92.52–96.29%], RF achieved 94.52% [93.02–96.01%] and SVM achieved 94.30% [92.52–96.09%]. The overlapping confidence intervals indicate no statistically significant differences between these models; this is also confirmed by McNemar's pairwise tests: LR vs RF: p=1.00; LR vs SVM: p=1.00; RF vs SVM: P=0.86; as shown in the supplementary Table S1: 10.6084/m9.figshare.31637011. Such convergence across fundamentally different algorithmic architectures — linear (LR), ensemble (RF), and maximum-margin (SVM) — indicates that the discriminative power resides in the *k*-mer features themselves and not in any particular classifier.

Table 2: The classification performance of machine learning models evaluated by 10-fold stratified cross-validation on 7-mer frequency features

Model*	Mean Accuracy (%)	95% CI	Mean F1
LR	94.41	92.52–96.29	0.9425
RF	94.52	93.02–96.01	0.9435
SVM	94.3	92.52–96.09	0.9419

* The 1D-CNN (95.26%) and LSTM (16%) were evaluated on the held-out test set in the original analysis and are reported in the text; 10-fold CV was applied to the three traditional ML models to enable rigorous statistical comparison.

On the held-out test set (n=190), LR achieved 95.26%, RF 94.74%, and SVM 95.79% accuracy, consistent with the cross-validation estimates. Table 3 presents per-genotype precision, recall, and F1-scores for each model. G5 and G7 were classified with near-perfect accuracy (F1≥0.97 across all models), while G10 was seen as the most challenging genotype (F1=0.67–0.73).

Table 3: Per-genotype F1-scores on the held-out test set (n=190) for the three traditional ML models (unweighted).

Genotype	n_test	LR_P	LR_R	LR_F1	RF_P	RF_R	RF_F1	SVM_P	SVM_R	SVM_F1
G1	54	0.981	0.944	0.962	1	0.926	0.962	1	0.944	0.971
G10	5	0.75	0.6	0.667	0.571	0.8	0.667	0.667	0.8	0.727
G3	54	0.964	0.981	0.972	0.947	1	0.973	0.947	1	0.973
G5	17	1	1	1	1	0.941	0.97	1	1	1
G6	28	0.867	0.929	0.897	0.923	0.857	0.889	0.893	0.893	0.893
G7	24	1	0.958	0.979	1	1	1	1	0.958	0.979
G8	8	0.889	1	0.941	0.8	1	0.889	1	1	1

PCA of 7-mer classification via 10-fold cross-validation: To ensure robust and generalizable results, we employed 10-fold cross-validation. Although the test sample was chosen at random and stratified across the seven classes, k-fold cross-validation is typically recommended to guarantee that the test results are reliable and generalizable (Al-khlifeh and Hassanat, 2024; Al-Khlifeh *et al.*, 2024) . The outcomes of this strategy are presented in Fig. 3, which shows the performance metrics obtained through this rigorous validation technique, along with the comparative analysis for each genotype (Fig. 4).

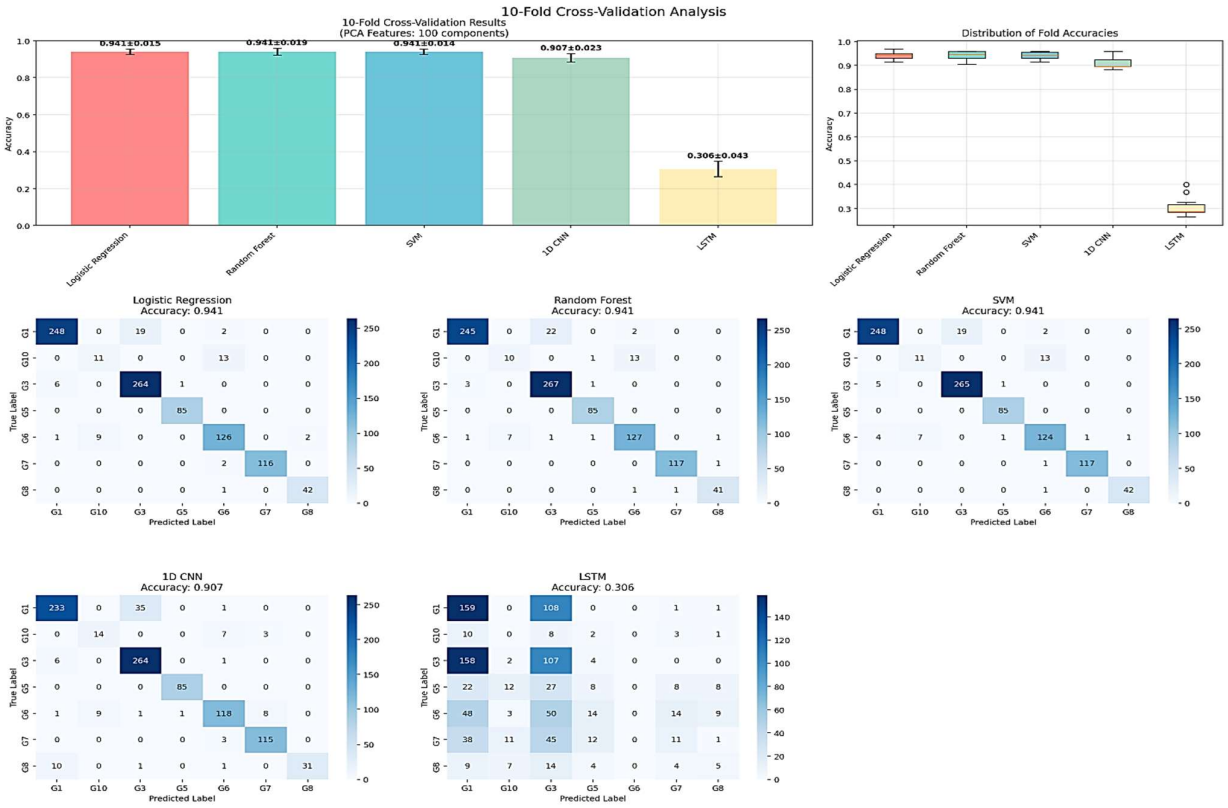


Fig. 3: 10-fold results of *E. granulosus s.l.* Genotype classification via PCA coefficients of the 7-mer.

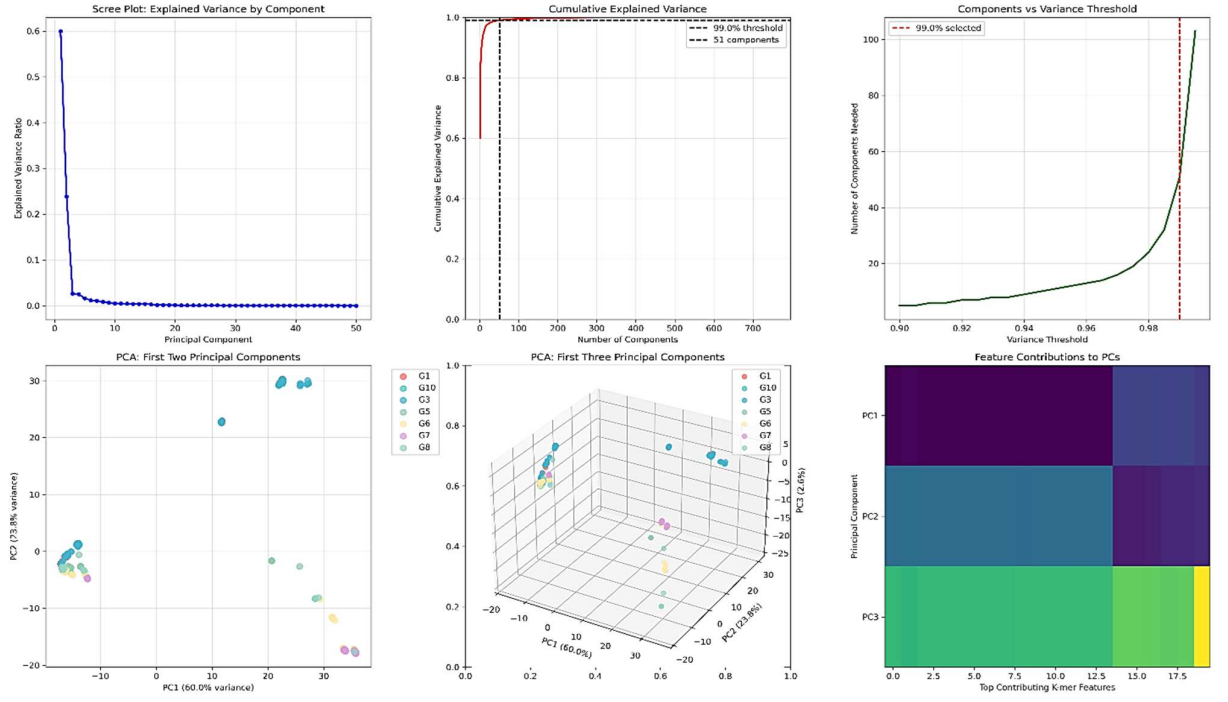


Fig. 4: PCA comprehensive analysis.

As shown in Table 3, best performers, SVM, LR, and RF, are extremely consistent across the different folds. The consistency across the 10 runs reflects the ability of a model to perform well regardless of the data used, and it can be calculated via

$$\text{Consistency} = (1 - \text{Performance Range}) \times 100\%$$

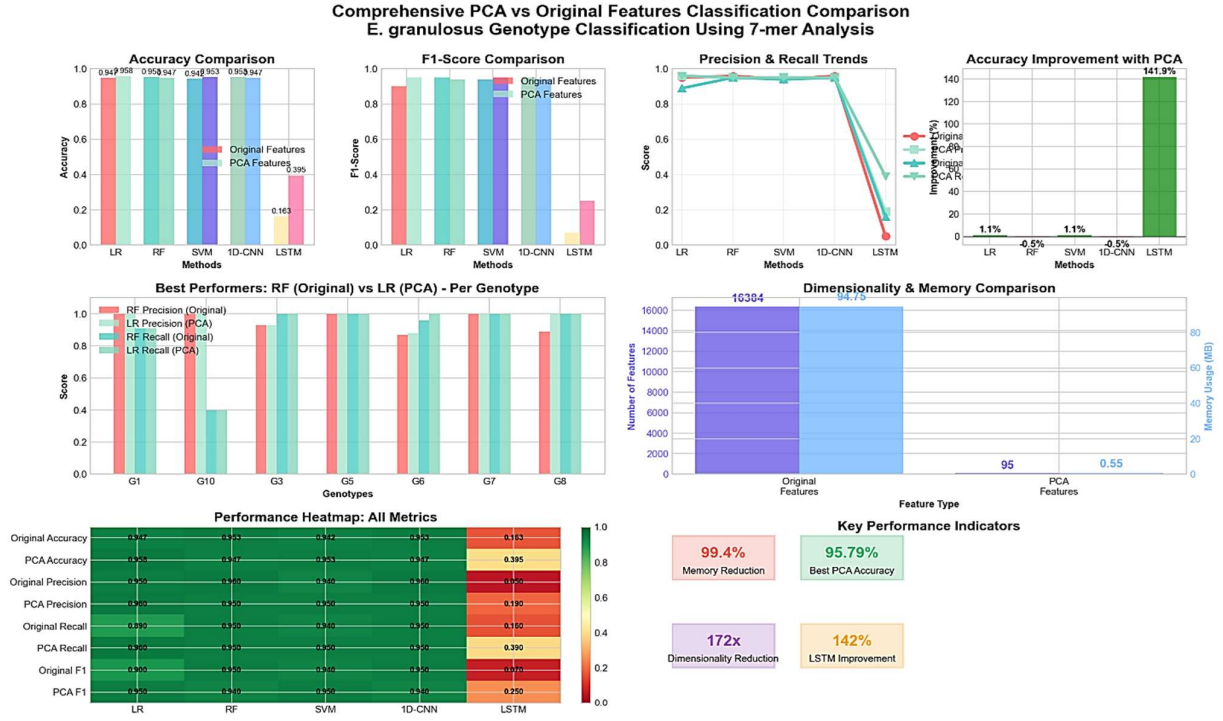
where the Performance Range = (Max Accuracy - Min Accuracy)

The PCA findings for *E. granulosus s.l.* genotype categorization emphasize crucial insights into performance implications and efficiency improvements (Fig. 5). The findings show that PCA transformation has diverse but usually beneficial effects on classification performance across a variety of ML models.

In order to confirm that no information leakage occurred through the PCA process, we repeated the dimensionality reduction with PCA fitted exclusively within each cross-validation fold, in other words, the PCA transformation was learned on training data only and applied to held-out test data. Such leak-free process yielded results virtually identical to the original analysis: LR 94.41% [92.97–95.84%], RF 94.20% [92.86–95.54%], and SVM 94.51% [92.92–96.10%]. The negligible differences (≤0.5 percentage points) confirm that the original PCA did not inflate performance estimates and that the 95 principal components (retaining 98.99% of variance) truly capture the discriminative structure of the *k*-mers feature space.

BLAST comparison results: Machine learning (using *k*-mers) remains superior, with 95.44% accuracy. On the other hand, BLAST-KNN (bit score/percent identity) achieves competitive accuracies of 88.95% and 88.95% (Table 4). While DL methods lag behind at 66–81%, similarly, BLAST-KNN (E value) performs poorly at 26–50%. The bit score and percent identity both achieve identical 88.95% accuracy with *k*=5, and we notice a consistent improvement with increasing k (number of neighbors) values (1→3→5), indicating robust performance across different genotype classes. We also notice the poor performance of the classification when the E value is used as a similarity measure, where the results are dramatically inferior (approximately 25.79–50.00% accuracy). Perhaps the E value transformation creates inappropriate distance scaling, which negatively affects the final results. In addition to length dependency, their logarithmic transformation may not preserve classification-relevant distances well.

Sequence length bias control: Knowing the substantial variation in sequence lengths across the dataset (309–17,675bp), we conducted a series of control experiments to verify that genotype classification was driven by nucleotide composition patterns rather than by differences in sequence length (Fig. 6).



Performance Summary Table

Method	Original Accuracy	PCA Accuracy	Improvement	Status
LR	94.74%	95.79%	+1.11%	Improved
RF	95.26%	94.74%	-0.55%	Decreased
SVM	94.21%	95.26%	+1.11%	Improved
1D-CNN	95.26%	94.74%	-0.55%	Decreased
LSTM	16.32%	39.47%	+141.85%	Improved

Fig. 5: Comprehensive comparison of PCA and original features for *E. granulosus* genotype classification via 7-mer analysis.

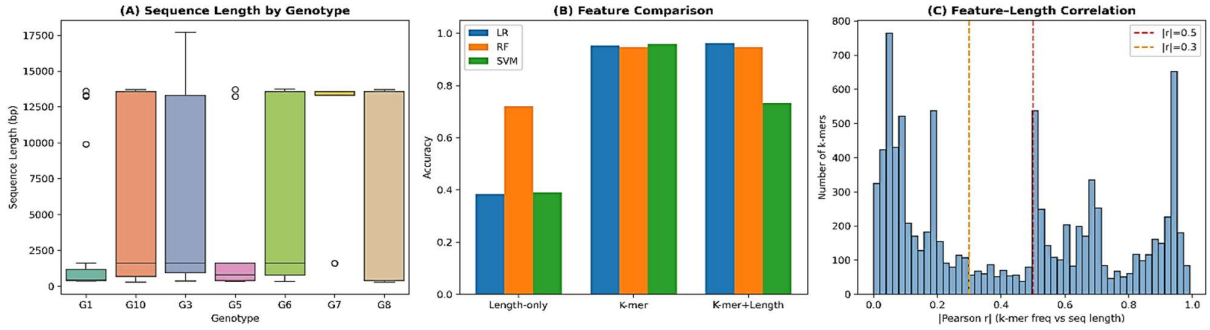


Fig. 6: Sequence length bias control analysis.

Table 4: Classification Results of *E. granulosus s.l.* Genotypes Using the BLAST-Similarity-KNN Classifier on a pure DNA sequence as a feature vector.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Time (s)
BLAST-1NN-bit-score	86.84	89	87	86	248
BLAST-3NN-bit-score	87.37	88	87	87	245.3
BLAST-5NN-bit-score	88.95	89	89	89	242.9
BLAST-1NN-percent-identity	87.37	88	87	87	250.1
BLAST-3NN-percent-identity	87.89	88	88	88	247.3
BLAST-5NN-percent-identity	88.95	90	89	89	248
BLAST-1NN-evalue	27.89	33	28	26	248.9
BLAST-3NN-evalue	25.79	31	26	24	237.9
BLAST-5NN-evalue	50.00	68	50	50	244.7
Best Performer: BLAST-5NN-bit-score (88.95% Accuracy)					
Genotype	Support	Precision (%)	Recall (%)	F1-Score (%)	
G5	17	100	100	100	
G7	24	100	100	100	
G8	8	100	100	100	
G6	28	87	96	92	
G1	54	93	74	82	
G3	54	80	94	86	
G10	5	67	40	50	

When classifiers were trained using only sequence length as the input feature, performance was markedly poor: LR achieved 38.4%, RF 72.1%, and SVM 38.9% accuracy — far below the corresponding *k*-mer-based performance of 95.3, 94.7, and 95.8%, respectively, as seen in Fig. 6a. The RF's elevated length-only accuracy (72.1%) reflects its ability to partially discriminate G7, which has distinctly longer sequences (mean 11,666±4,268bp), but this single-genotype effect was insufficient to approach *k*-mer-based performance. Combining *k-mer* features with sequence length as an additional feature produced negligible improvement (LR:96.3%, RF:94.7%, SVM:73.2%), indicating that *k*-mer features already encode all discriminative information and that length adds no meaningful signal.

Our Pearson correlation analysis between each *k*-mer feature and sequence length revealed that, among 9,333 non-zero-variance *k*-mers, the mean absolute correlation was |r|=0.43, as seen in Fig. 6c. While this moderate correlation is expected — shorter partial sequences (e.g., cox1 fragments) inherently sample different *k*-mer distributions than complete mitochondrial genomes— it does not drive classification.

This was directly indicated by evaluating classification accuracy within length-stratified subsets: accuracy remained high for short sequences (<1,000bp: 97.3%), medium sequences (1,000–5,000bp: 90.3%), and long sequences (>5,000bp: 91.5%), confirming that *k*-mer frequency patterns are discriminative regardless of fragment size.

Per-genotype classification performance and error analysis: To understand variation in classification difficulty across genotypes, within-class *k*-mer diversity and confusion patterns were analyzed using 10-fold cross-validated predictions (Fig. 7). The Genotype classification performance was strongly associated with both sample size and within-class molecular diversity. G7 (n=118) exhibited the lowest within-class cosine distance (0.23), indicating highly homogeneous *k*-mer profiles, and was classified with near-perfect accuracy (RF F1=0.992). Similarly, G5 (n=85, diversity=0.51) achieved RF F1=0.994, and G1 (n=269, diversity=0.43) achieved RF F1=0.952. In contrast, G10 represented the most challenging genotype, achieving RF F1=0.524 in 10-fold CV. This reduced performance is attributable to two compounding factors: (i) the smallest sample size in the dataset (n=24), providing limited training examples, and (ii) high within-class *k*-mer diversity (cosine distance=0.60), indicating substantial molecular heterogeneity. Confusion matrix analysis revealed that G10 misclassifications were overwhelmingly directed toward G6 (13 of 24 G10 samples misclassified as G6), consistent with the known phylogenetic proximity of these genotypes within the E. Canadensis complex (G6/G7/G8/G10). Reciprocally, G6 (RF F1=0.911) exhibited misclassification primarily toward G10 (7 cases) and G8 (3 cases), further confirming that classification errors track phylogenetic relationships rather than occurring randomly.

G8, despite having the highest within-class diversity (cosine distance=0.69) and a small sample size (n=43), achieved RF F1=0.923, suggesting that it possesses sufficiently distinct *k*-mer signatures to remain separable from other genotypes.

A sensitivity analysis comparing unweighted and balanced class-weighted models (Supplementary Table S2: 10.6084/m9.figshare.31637011 and Fig. S2: 10.6084/m9.figshare.31636462, revealed that applying inverse-frequency class weights produced minimal changes to overall accuracy (≤1 percentage point) and per-genotype F1-scores. For G10, LR showed a modest improvement (F1: 0.667→0.714), while RF showed a decline (F1:

0.667→0.545), and SVM remained unchanged (F1: 0.727). These results indicate that the classification difficulty for G10 is driven primarily by biological factors (small population representation and phylogenetic proximity to G6) rather than by algorithmic sensitivity to class imbalance.

AI explained and SHAP analysis: In order to validate the biological relevance of the most discriminative features identified, the top 20 *k*-mers per genotype (140 total, ranked by absolute Logistic Regression coefficient) were mapped to the annotated *E. granulosus* G1 reference mitochondrial genome (AF297617, 13,588bp) (Fig. 8, and the supplementary Table S3: 10.6084/m9.figshare. 31637011.

The discriminative 7-mers were distributed across all mitochondrial gene regions, with the highest concentrations in cytochrome b (cytb, 27 *k*-mers) and cytochrome c oxidase subunit I (cox1, 25 *k*-mers) (Fig. 8). This enrichment is biologically expected, as cytb and cox1 are the two most commonly used molecular markers for *Echinococcus* genotyping and are known to harbor the highest density of genotype-specific single nucleotide polymorphisms. Substantial representation was also seen in the NADH dehydrogenase genes (nad5: 17, nad1: 16, nad4: 12, nad3: 12), ATP synthase 6 (atp6: 11), and the ribosomal RNA genes (rrnL: 12, rrnS: 7). Even the non-coding region (NCR: 4 *k*-mers) contributed discriminative features, confirming that genotype-level variation is distributed genome-wide.

Several genotype-specific motif patterns provided biological insight. The 7-mer GATGTGT exhibited strong but opposing coefficients for G1 (−0.574) and G3 (+0.585), consistent with the known close evolutionary relationship of these genotypes within *E. granulosus s.s*, where a small number of point mutations in conserved regions drive divergence. Similarly, CCGGTGT showed opposite associations for G6 (−0.622) and G8 (+0.326), reflecting divergence within the *E. canadensis* complex. The *k*-mer GTTCTAT (coefficient +0.592 for G1) and TGATGTG (+0.589 for G3) mapped to the cox1 gene region, further corroborating that the classifier exploits the same barcoding regions that parasitologists use for conventional genotyping. These results indicate that the *k*-mer-based ML approach does not rely on arbitrary or biologically meaningless sequence patterns but instead captures genuine molecular variation concentrated in genes with known diagnostic and evolutionary significance.

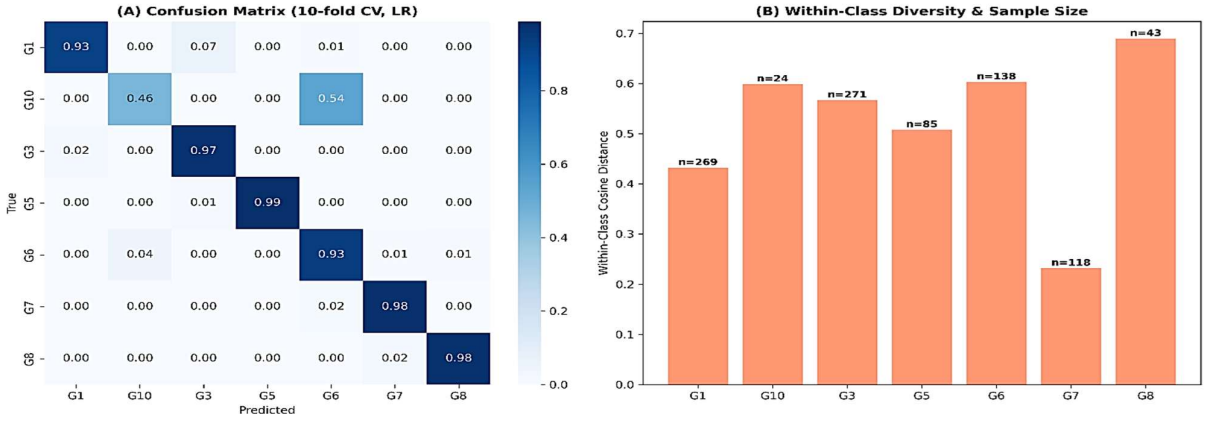


Fig. 7: Within-class *k*-mer diversity and confusion matrix predictions using 10-fold cross-validation.

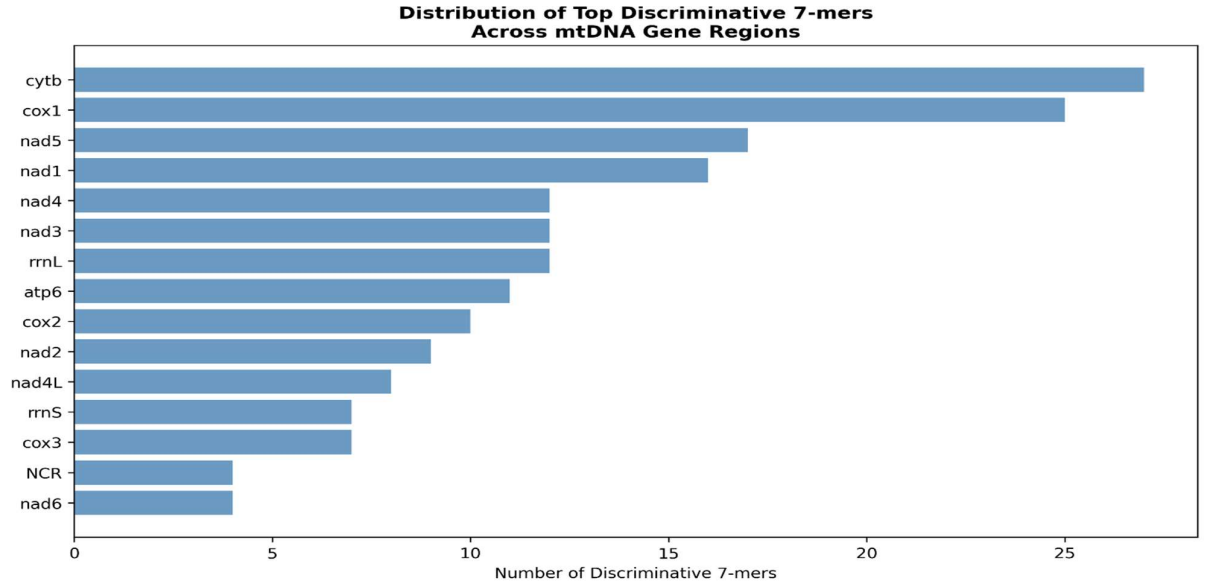


Fig. 8: The distribution of top discriminative 7-mers across mitochondrial gene regions of the *E. granulosus* reference genome (AF297617).

DISCUSSION

The "paradox" of tapeworm classification, primarily at the species level, is a significant challenge because of their intricate life cycles and the small number of observable morphological traits, especially during the larval stages (Mlowe *et al.*, 2022). This work is the first to streamline the application of ML to classify genotypes and species of *E. granulosus s.l.*, a tapeworm primarily defined as a cryptic species complex (Nakao *et al.*, 2010). Previous molecular studies emphasize the need to revise the classification of *E. granulosus s.l.* (Alvarez Rojas *et al.*, 2014; Peytavin de Garam *et al.*, 2025). This revision is necessary because *E. granulosus s.l.* contains multiple genotypes that are potentially genetically distinct enough to be considered separate species (Lymbery *et al.*, 2015). However, the utilization of mtDNA markers could not provide an efficient genotypic distinction, particularly in the case of the *E. granulosus s.s.* and *E. canadensis* clusters.

The robustness of the *k*-mer classification approach was validated through several complementary analyses. First, a sequence length control experiment indicated that length alone is a poor predictor of genotype (38–72% accuracy vs. 95–96% with *k*-mers), and classification remained accurate within narrow length strata (90–97%), confirming that discriminative power derives from nucleotide composition rather than fragment size. Second, PCA dimensionality reduction performed strictly within cross-validation folds yielded results identical to the global PCA analysis (94.2–94.5%), ruling out information leakage. Third, all three traditional ML models (LR, RF, SVM) achieved statistically indistinguishable performance (McNemar's P>0.85), indicating that the classification success is attributable to the informativeness of *k*-mer features rather than to any specific algorithmic advantage.

In this study, G7 differed from all the other genotypes, demonstrating its unique molecular makeup. On the other hand, the closest pair of genotypes were G6 and G8, along with a high level of genetic variation within each genotype. SHAP analysis revealed that those genotypes also presented some 7-mer overlap. Nonetheless, G8 indicated perfect classification (100% F1 score), whereas G6 had good performance (>90% F1 score) with high recall (96%), indicating substantial sequence conservation. This finding is very significant because there has long been disagreement among the investigators who have examined the genotypes of the *E. canadensis* cluster (G6–G8 and G10) via both nuclear and mitochondrial markers (Nikmanesh *et al.*, 2014; Yanagida *et al.*, 2017; Laurimäe *et al.*, 2019; Ohiolei *et al.*, 2019; Laurimäe *et al.,* 2023b). In contrast to research that assumes that G6/G7 and G8/G10 are two separate species, the results of this study show that these four genotypes are well differentiated and consistent with the hypothesis of a coherent species entity, though ML classification alone cannot resolve species boundaries. This could be because more sequences were included, enabling a more thorough comparison of the data.

The controversial taxonomy of the *E. canadensis* cluster is mostly due to the high genetic variation within the G6 and G8 genotypes, as revealed by the results of the 7-mer features. Nucleotide diversity suggests that the population has recently expanded from a smaller, genetically homogeneous base. Host interaction and biogeography research presume a connection between genotyping patterns within an *E. canadensis* cluster and evolutionary processes such as speciation and migration, allowing populations to adjust to shifting environmental conditions. For example, a study revealed that the African/Middle Eastern subcluster has a high frequency of G6, whereas the European isolates have a high frequency of G7 (Addy *et al.*, 2017; Casulli *et al.*, 2022).

Among the seven genotypes examined, G10 presented the greatest classification challenge (RF F1=0.524 in 10-fold CV). This is attributable to two factors: its limited representation in the dataset (n=24, comprising only 2.5% of samples) and its high within-class *k*-mer diversity (cosine distance=0.60), reflecting substantial molecular heterogeneity among the available sequences. Notably, G10 misclassifications were directed almost exclusively toward G6 (13/24 cases), mirroring the established phylogenetic proximity of these genotypes within the *E. canadensis* complex (Nakao *et al.*, 2007; Romig *et al.*, 2015). The G6–G8–G10 confusion pattern saw in our classifier is consistent with the ongoing taxonomic debate regarding species boundaries within *E. canadensis*, where molecular, morphological, and host-range evidence provide sometimes-contradictory signals. Conversely, G7 achieved near-perfect classification (RF F1=0.992) with the lowest within-class diversity (cosine distance=0.23), suggesting that this genotype possesses a highly conserved mitochondrial signature that is readily distinguished by *k*-mer analysis. A sensitivity analysis with balanced class weights showed minimal changes (≤1 percentage point in overall accuracy), confirming that the difficulty with G10 is driven by biological factors rather than class imbalance. Increasing the representation of undersampled genotypes (particularly G10 and G8) in future studies would be the most effective route to improving classification performance for these challenging groups.

We emphasize that ML classification of molecular sequences provides statistical evidence for genotype distinctiveness but should not be interpreted as definitive taxonomic resolution. Integrative approaches combining molecular, morphological, ecological, and host-specificity data remain essential for formal taxonomic decisions within the *E. granulosus s.l.* complex.

Another significant finding of this study is the relatively high degree of internal diversity in the G3 sequences. The performance metrics of the ML model reveal that G3 has high recall (94%) but lower precision. This result is further explained by the SHAP analysis. Previous studies have confirmed that G3s exhibit significant internal genetic diversity. While the G3 genotype is a distinct lineage within *E. granulosus*, molecular and population genetic analyses revealed a variety of haplotypes and substructures within G3 itself. For example, a 2022 study on *E. granulosus* in cattle in Pakistan revealed 11 different haplotypes for the G3 genotype on the basis of the partial nad5 gene alone (Mehmood *et al.*, 2022). A study in Levant using *cox1* genes identified multiple haplotypes within G3 (Al-khlifeh, Alshammari, and Alnasarat, 2024). Additional evidence from research in countries such as Pakistan and India, where G3 is highly prevalent, shows that it is genetically diverse and appears to be actively diverging and expanding

(Mehmood *et al.*, 2020; Alvi *et al.*, 2023). In some European countries, such as Italy and France, studies have also identified substantial haplotype diversity within their G3 populations, although it may be more localized (Mehmood *et al.*, 2021). SHAP analysis revealed that G1 and G3 also share 7-mers that are distinct from G5-G8 and G10. This result is not surprising, as G1 and G3 are often closely related but distinct lineages. Sympatric presence in a mixture of different host animals potentially contributes to complex transmission dynamics and genetic mixing. Notably, G3 often circulates in the same areas as other genotypes, most notably G1 (sheep strain) (Mehmood *et al.*, 2021).

The biological mapping of the top discriminative *k*-mers to the *E. granulosus* reference mitochondrial genome provides important validation that the classifier exploits genuine molecular variation rather than computational artifacts. The concentration of discriminative motifs in cytb (27 *k*-mers) and cox1 (25 *k*-mers) is particularly significant, as these genes serve as the primary molecular markers in *Echinococcus* taxonomy and diagnostics and are known to harbour the highest density of genotype-specific polymorphisms (Bowles *et al.*, 1994; Nakao *et al.*, 2010). Moreover, the opposing coefficient patterns seen for G1 versus G3 (GATGTGT: −0.574 vs. +0.585) and G6 versus G8 (CCGGTGT: −0.622 vs. +0.326) mirror the subtle nucleotide divergences within *E. granulosus s.s.* and *E. canadensis*, respectively. The distribution of discriminative features across all 12 protein-coding genes, both rRNAs, and the non-coding region validates the use of whole-mitogenome *k*-mer analysis over single-gene approaches and suggests that genotype-level molecular divergence is a genome-wide phenomenon (Santucciu *et al.*, 2023). The results of the analysis of the G5 sequences revealed good separation in the sequence space in the network of the seven tested genotypes, distinct BLAST signatures, and genetic differentiation, which led to perfect classification (100% F1 score). This outcome can actually be interpreted as additional evidence of the ML model's dependability and suitability for the *E granulosus s.l.* classification task, since we evaluate performance on independent species with established classifications to ensure that the model can correctly spot patterns in data that it has not seen before. However, it should be noted that the BLAST-KNN framework represents a simplified proxy for the nearest-neighbour classification strategy commonly used in diagnostic laboratories for rapid genotype assignment, rather than a full phylogenetic analysis incorporating evolutionary models and tree topology. The comparison thus indicates that *k*-mer-based ML can match or exceed the practical accuracy of similarity-based sequence classification, without claiming superiority over rigorous phylogenetic methods.

The results reveal that classical ML algorithms outperform DL methods. Deep learning models require massive amounts of data to be trained effectively and to avoid overfitting, but this is not always available in genotyping studies, especially for parasite populations. On the other hand, the RF algorithm can generalize well from limited data, making it more effective. Moreover, in the context of the high genetic diversity within the *E. granulosus s.l.*, a small, unrepresentative dataset can cause a DL model to "memorize" the training data without learning generalizable patterns, leading to poor performance on new data. RFs have consistently been shown to outperform DL models on structured, tabular datasets, often with more computationally efficient performance. Therefore, for *E. granulosus s.l.* genotyping tasks where the most informative *k*-mers are already known, classical ML can be a more direct and effective approach.

An additional finding of this study is the notable variability in the performance of DL models. The 1D-CNN outperforms the LSTM in this experiment because CNNs excel at local pattern recognition, making them especially useful for collecting local sequence patterns. They also show parameter efficiency while having more parameters, resulting in a more stable training process with a constant learning curve. However, the sequence length of 1,608 nucleotides may pose challenges for the LSTM, potentially making it less effective for this particular dataset. The LSTM achieved only 16% accuracy (near chance level for seven classes), attributable to the fundamental mismatch between the LSTM architecture — which is designed to model sequential dependencies —and *k*-mer frequency vectors, which are inherently order less representations in which positional information from the original sequence is absent. This result underscores that *k*-mer frequency representations, while highly effective for methods that operate on feature distributions (LR, RF, SVM, 1D-CNN with convolutional filters), are unsuitable for recurrent architectures that require meaningful input ordering.

The SHAP results provide a global picture of the best sequence motifs that help classify each genotype by zooming in on the significance of sequence features (AL-zyadat *et al.*, 2022). From a biological standpoint, the discovery of these conserved motifs and their role in successful genotype categorization by ML aligns with earlier findings indicating that the content of the mtDNA gene is conserved throughout metazoans (Shtolz and Mishmar, 2023). However, their nucleotide sequences might differ among and even within species.

Limitations of the *k*-mer-based ML approach could be centered on the choice of *k*-mer length (k), where a small k lacks specificity and a large k suffers from sparsity and computational cost, leading to overfitting. However, in this study, we adjust the *k*-mer length to balance the need for unique, informative sequences with the complexity and potential sparsity of the data. Other limitations include the issues of data requirements because the underlying methodology and model performance improve as more data become accessible. In this study, we include only sequences that describe the genotype name in their GenBank identifier; even though we have stringent sequence selection criteria, we cannot rule out the possibility of misclassified sequences in a sizable database such as GenBank.

Conclusions: *K*-mer analysis is effective at identifying genetic variation of *E. granulosus s.l.*, whereas ML algorithms can be trained on these *k*-mer profiles to classify genotypes rapidly and accurately. This integration is particularly useful for large-scale epidemiological studies that track transmission dynamics and identify infection sources. Training on sequences from genetic regions with known variation, such as the EG95 gene family members

(encodes a protective antigen used in vaccines against hydatid disease), could be a significant future prospect of this study.

Conflicts of interest: The authors declare that they have no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethics statement: This study does not involve human participants or their personal information and therefore does not require ethical approval. The research is based on publicly available datasets that do not present any ethical risks.

Data availability: Genotype data associated with the paper can be accessed via: https://doi.org/10.6084/m9.figshare.31563277
Supplementary figures can be accessed via: 10.6084/m9.figshare.31636462
Supplementary Tables (S1, S2 and S3) can be accessed via: 10.6084/m9.figshare.31637011

Funding statement: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contributions: Enas M Al-khlifeh: Conceptualization, data acquisition, interpretation of data for the work, writing of the original manuscript. Ahmad B. Hassanat: Methodology, data analysis, visualization and writing of the original manuscript. Lujain A Alhasanat: Methodology and critical revision of the manuscript. Ahmad S. Tarawneh, Suleyman A. AlShowarah and Awni Hammouri: Data analysis and critical revision of the manuscript.

Acknowledgment: Not applicable.

REFERENCES

Addy F, Wassermann M, Kagendo D, *et al*., 2017. Genetic differentiation of the G6/7 cluster of *Echinococcus canadensis* based on mitochondrial marker genes. International Journal of Parasitology 47:923-931.

Ali S, Sahoo B, Ullah N, *et al*., 2021. A k-mer based approach for SARS-CoV-2 variant identification. In: Bioinformatics Research and Applications (Wei Y, Li M, Skums P, *et al*., eds). Springer International Publishing, Cham, pp:153-164.

Al-khlifeh EM and Hassanat AB, 2024. Predicting the distribution patterns of antibiotic-resistant microorganisms in the context of Jordanian cases using machine learning techniques. Journal of Applied Pharmaceutical Science 14:174-183.

Alkhlifeh E, Saidat N, Khleifat K, *et al*., 2023. Phytochemical profile and in vitro protoscolicidal effects of *Juniperus phoenicea* L., *Calotropis procera* (Aiton) Dryand and *Artemisia judaica* L. against *Echinococcus granulosus* cysts. Journal of Pharmacy and Pharmacognosy Research 11:635-650.

Al-khlifeh E, Alshammari A and Alnasarat H, 2024. High incidence of G1 genotype found in the Levant revealed by sequence-based association analysis of *Echinococcus granulosus* (sensu stricto). Pakistan Veterinary Journal.

Al-khlifeh E, Tarawneh AS, Almohammadi K, *et al*., 2025. Decision tree-based learning and laboratory data mining: an efficient approach to amebiasis testing. Parasites and Vectors 18:33.

Al-Khlifeh EM, Alkhazi IS, Alrowaily MA, *et al*., 2024. Extended spectrum beta-lactamase bacteria and multidrug resistance in Jordan are predicted using a new machine-learning system. Infection and Drug Resistance 17:3225-3240.

Alvarez Rojas CA, Romig T and Lightowlers MW, 2014. *Echinococcus granulosus* sensu lato genotypes infecting humans – review of current knowledge. International Journal of Parasitology 44:9-18.

Alvi MA, Ali RMA, Li L, *et al*., 2023. Phylogeny and population structure of *Echinococcus granulosus* (sensu stricto) based on full-length cytb-nad2-atp6 mitochondrial genes – first report from Sialkot District of Pakistan. Molecular and Biochemical Parasitology 253:111542.

Al-zyadat W, Muhairat M, Alhroob A, *et al*., 2022. A recruitment big data approach to interplay of the target drugs. International Journal of Advances in Soft Computing and Its Applications 14:2-13.

Asif HM, Khan SH, Alahmadi TJ, *et al*., 2024. Malaria parasitic detection using a new deep boosted and ensemble learning framework. Complex and Intelligent Systems 10:4835-4851.

Bize A, Midoux C, Mariadassou M, *et al*., 2021. Exploring short k-mer profiles in cells and mobile elements from Archaea highlights the major influence of both the ecological niche and evolutionary history. BMC Genomics 22:186.

Borhani M, Fathi S, Harandi MF, *et al*., 2024. *Echinococcus granulosus* sensu lato control measures: a specific focus on vaccines for both definitive and intermediate hosts. Parasites and Vectors 17:533.

Bowles J, Blair D and McManus DP, 1994. Molecular genetic characterization of the cervid strain ('northern form') of *Echinococcus granulosus*. Parasitology 109:215-221.

Bussi Y, Kapon R and Reich Z, 2021. Large-scale k-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy. PLoS ONE 16:e0258693.

Casulli A, Massolo A, Saarma U, *et al*., 2022. Species and genotypes belonging to *Echinococcus granulosus* sensu lato complex causing human cystic echinococcosis in Europe (2000–2021): a systematic review. Parasites and Vectors 15:109.

Deelder W, Manko E, Phelan JE, *et al*., 2022. Geographical classification of malaria parasites through applying machine learning to whole genome sequence data. Scientific Reports 12:21150.

Imam AT, Alhroob A and Alzyadat WJ, 2021. SVM machine learning classifier to automate the extraction of SRS elements. International Journal of Advanced Computer Science and Applications 12.

Kinkar L, Laurimäe T, Simsek S, *et al*., 2016. High-resolution phylogeography of zoonotic tapeworm *Echinococcus granulosus* sensu stricto genotype G1 with emphasis on its distribution in Turkey, Italy and Spain. Parasitology 143:1790-1801.

Kinkar L, Laurimäe T, Acosta-Jamett G, *et al*., 2018a. Distinguishing *Echinococcus granulosus* sensu stricto genotypes G1 and G3 with confidence: a practical guide. Infection, Genetics and Evolution 64:178-184.

Kinkar L, Laurimäe T, Acosta-Jamett G, *et al*., 2018b. Global phylogeography and genetic diversity of the zoonotic tapeworm *Echinococcus granulosus* sensu stricto genotype G1. International Journal for Parasitology 48:729-742.

Laurimäe T, Kinkar L, Romig T, *et al*., 2019. Analysis of nad2 and nad5 enables reliable identification of genotypes G6 and G7 within the species complex *Echinococcus granulosus* sensu lato. Infection, Genetics and Evolution 74:103941.

Laurimäe T, Kinkar L, Moks E, *et al*., 2023. Exploring the genetic diversity of genotypes G8 and G10 of the *Echinococcus canadensis* cluster in Europe based on complete mitochondrial genomes (13,550–13,552 bp). Parasitology 150:631-637.

Lymbery AJ, Jenkins EJ, Schurer JM, *et al*., 2015. *Echinococcus canadensis*, *E. borealis* and *E. intermedius*: what's in a name? Trends in Parasitology 31:23-29.

Malik M, Subhani M, Alvi A, *et al*., 2024. High genetic variability in full-length cox2 and nad6 genes of *Echinococcus granulosus* sensu stricto and *Echinococcus ortleppi* recovered from cattle. Pakistan Veterinary Journal 44:148-154.

Mehmood N, Muqaddas H, Arshad M, *et al*., 2020. Comprehensive study based on mtDNA signature (nad1) providing insights on *Echinococcus granulosus* sensu stricto genotypes from Pakistan and potential role of buffalo-dog cycle. Infection, Genetics and Evolution 81:104271.

Mehmood N, Dessì G, Ahmed F, *et al*., 2021. Genetic diversity and transmission patterns of *Echinococcus granulosus* sensu stricto among domestic ungulates of Sardinia, Italy. Parasitology Research 120:2533-2542.

Mehmood N, Muqaddas H, Ullah MI, *et al*., 2022. Genetic structure and phylogeography of *Echinococcus granulosus* sensu stricto genotypes G1 and G3 in Pakistan and other regions of the world based on nad5 gene. Infection, Genetics and Evolution 98:105223.

Mlowe F, Karimuribo E, Mkupasi E, *et al*., 2022. Challenges in the diagnosis of *Taenia solium* cysticercosis and taeniosis in medical and veterinary

settings in selected regions of Tanzania: a cross-sectional study. Veterinary Medicine International 2022:7472051.

Nakao M, Li T, Han X, *et al*., 2010. Genetic polymorphisms of *Echinococcus* tapeworms in China as determined by mitochondrial and nuclear DNA sequences. International Journal of Parasitology 40:379-385.

Neglected tropical diseases – GLOBAL, 2025. Available at: https://www.who.int/health-topics/neglected-tropical-diseases (accessed September 24, 2025).

Nikmanesh B, Mirhendi H, Ghalavand Z, *et al*., 2014. Genotyping of *Echinococcus granulosus* isolates from human clinical samples based on sequencing of mitochondrial genes in Iran, Tehran. Iranian Journal of Parasitology 9:20-27.

Ohiolei JA, Xia CY, Li L, *et al*., 2019. Genetic variation of *Echinococcus* spp. in yaks and sheep in the Tibet Autonomous Region of China based on mitochondrial DNA. Parasites and Vectors 12:608.

Peytavin de Garam C, Boué F, Knapp J, *et al*., 2025. Study of genetic diversity of *Echinococcus granulosus* sensu stricto in France based on full cox1 gene. Infection, Genetics and Evolution 131:105757.

Saarma U, Jõgisalu I, Moks E, *et al*., 2009. A novel phylogeny for the genus *Echinococcus*, based on nuclear data, challenges relationships based on mitochondrial evidence. Parasitology 136:317-328.

Santucciu C, Bonelli P, Peruzzu A, *et al*., 2023. Genetic characterization of *Echinococcus granulosus* sensu stricto isolated from human cysts from Sardinia, Italy. Diseases 11:91.

Shajii A, Yorukoglu D, William Yu Y, *et al*., 2016. Fast genotyping of known SNPs through approximate k-mer matching. Bioinformatics 32:i538-i544.

Shtolz N and Mishmar D, 2023. The metazoan landscape of mitochondrial DNA gene order and content is shaped by selection and affects mitochondrial transcription. Communications Biology 6:93.

Tarawneh AS, Omari AKA, Al-khlifeh EM, *et al*., 2025. Non-invasive cancer detection using blood test and predictive modeling approach. Advances and Applications in Bioinformatics and Chemistry 17:159-178.

Vinje H, Liland KH, Almøy T, *et al*., 2015. Comparing k-mer based methods for improved classification of 16S sequences. BMC Bioinformatics 16:205.

Yanagida T, Lavikainen A, Hoberg EP, *et al*., 2017. Specific status of *Echinococcus canadensis* (Cestoda: Taeniidae) inferred from nuclear and mitochondrial gene sequences. International Journal of Parasitology 47:971-979.