Comparative Analysis of Supervised Algorithms for Protein Cluster Classification Using k-mer Image Embeddings

Giovanna A. P. Soares[1], Hannah I. da S. Marques[1], Matheus Dalmolin[1,2], Raquel de M. Barbosa[3] and Marcelo A. C. Fernandes[1,2,4]

[1]InovAI Lab, nPITI/IMD, UFRN, 59.078-900, Natal, RN, Brazil

[2]Bioinformatics Multidisciplinary Environment (BioME), IMD, UFRN, Natal, Brazil

[3]Faculty of Pharmacy, Granada University, Granada, Spain

[4]Department of Computer Engineering and Automation, UFRN, Natal, Brazil

giovanna.assuncao.705@ufrn.edu.br, hannahisabele1516@gmail.com

matheusdalmolinrs@gmail.com, rbarbosa@ugr.es

mfernandes@dca.ufrn.br

Abstract. *The rapid expansion of protein sequence databases requires effective computational strategies for accurate classification and functional annotation. This work presents a supervised learning framework for protein cluster classification using vector embeddings derived from k-mer image representations. Four supervised algorithms were systematically compared using five-fold stratified cross-validation: Logistic Regression (L2), Random Forest, k-Nearest Neighbors (kNN), and XGBoost. Embeddings extracted from k-mer images served as input features, enabling alignment-free and scalable classification. On the UniRef100 dataset, all models achieved strong performance. Logistic Regression obtained an accuracy of 98.1% and macro F1-score of 0.981, while Random Forest, kNN, and XGBoost achieved even higher accuracies of 99.7%, 99.4%, and 99.8%, respectively. XGBoost presented the best overall results, with an accuracy of 99.85%, F1-score of 0.9985, AUC of 1.000, and the lowest log loss (0.0071). For UniRef90, a more heterogeneous and challenging dataset, a decrease in accuracy and F1 was observed for all methods. Logistic Regression achieved 93.1% accuracy and F1-score of 0.931, while Random Forest, kNN, and XGBoost obtained accuracies of 99.3%, 98.6%, and 99.6%, respectively. Once again, XGBoost showed the best results for UniRef90, with an accuracy of 99.56%, F1-score of 0.9957, AUC of 0.9999, and log loss of 0.0194. The confusion matrices for both datasets indicate that most protein clusters were correctly classified, with only minor misclassifications among the most challenging classes. These findings demonstrate the effectiveness of embedding-based representations and tree-based ensemble methods for robust and interpretable protein cluster classification, even in more complex and diverse datasets.*

1. Introduction

The continuous expansion of protein sequence databases, driven by advances in high-throughput sequencing technologies, has generated an unprecedented volume of biological data. Efficiently organizing, classifying, and annotating this information is a

fundamental challenge in computational biology and bioinformatics. Accurate protein cluster classification plays a key role in understanding functional relationships, supporting comparative genomics, and guiding experimental research. However, the inherent diversity and complexity of protein families, combined with the presence of evolutionary variants and functionally similar but taxonomically distinct groups, make automated classification a non-trivial task. Addressing these challenges requires robust computational approaches capable of capturing relevant patterns in protein sequences and providing reliable predictions across diverse biological contexts.

Machine learning and natural language processing (NLP) methods have been applied to the prediction of protein properties and to the classification of protein sequences. Recent studies report the use of supervised and deep learning models, including CNNs, GNNs, transformers, and protein-specific NLP models, to extract representations from large-scale datasets and to perform prediction of protein structure, function, and interactions [Wang et al. 2024]. These works note the relevance of dataset quality, feature extraction, and evaluation strategies, as well as the use of data partitioning approaches that address evolutionary similarity. In this context, NLP-based methods that combine n-gram, embedding, and transformer representations with various classification algorithms are reported to improve protein classification when datasets are diverse and data splitting reflects biological structure [Perveen and Weeds 2025, Zhang et al. 2023].

Protein sequence classification is a fundamental task in bioinformatics, as it enables the organization, annotation, and functional prediction of proteins in large biological datasets. Accurate classification facilitates the identification of homologous proteins, supports the inference of structural and functional relationships, and aids in the discovery of new protein families [Mall et al. 2025, Suyunu et al. 2025, Luo and Cai 2024, Murad et al. 2023]. This process is essential for comparative genomics, evolutionary studies, and the development of applications in biotechnology and medicine, including drug discovery and disease diagnosis. As the volume of protein sequence data continues to increase, reliable classification methods are required to extract relevant biological information and to support advances in both basic and applied research [Tasnim et al. 2024, Ahmed et al. 2023, Lilhore et al. 2024, Blum et al. 2024, Balamurugan et al. 2023].

Thus, this work proposes a supervised classification framework for protein clusters that uses embeddings generated from k-mer image representations processed by a Vision Transformer (ViT) model. The study systematically compares the performance of Logistic Regression, Random Forest, k-Nearest Neighbors (kNN), and Extreme Gradient Boosting (XGBoost), directly addressing current challenges discussed in the literature [Wang et al. 2024, Perveen and Weeds 2025, Liu 2024], such as the impact of biological diversity, data partitioning, and feature extraction on model performance. The objective is to evaluate the effectiveness of these algorithms in classifying protein clusters with varying degrees of cohesion and heterogeneity, using two representative datasets (UniRef100 and UniRef90). This work aims to investigate whether tree-based ensemble methods, such as XGBoost, can achieve higher accuracy and robustness in heterogeneous scenarios, and to explore how biological context and cluster structure influence classification outcomes through the analysis of confusion matrices.

2. Methods

2.1. Dataset

In this work, two datasets derived from the UniRef database were used: UniRef100 and UniRef90. Both group protein sequences from UniProtKB based on similarity criteria. UniRef100 contains all unique sequences, while UniRef90 groups sequences with at least 90% identity. This difference directly impacts the cohesion of the groups, with UniRef90 being naturally more permissive and, therefore, more heterogeneous. Ten groups (or clusters) were manually selected from each dataset, aiming for functional and taxonomic diversity. Table 1 presents the selected clusters from UniRef100, with information about the cluster identifier, the name of the reference gene or protein, the number of amino acids (AA), the predominant species or group, and the number of sequences per cluster. Similarly, Table 2 summarizes the clusters used from UniRef90, also detailing the diversity of organisms associated with each group.

Table 1. Detailed distribution of UniRef100 groups (*clusters*).

UniRef Cluster ID	Gene/Protein	No. AA (Ref.)	Predominant species or group	No. Sequences
A0A0F7YU55	Matrix protein 1	252	*Influenza A virus*	971
A0A0U0T1S6	PPE family protein	73	*Mycobacterium orygis 112400015*	1
			Mycobacterium tuberculosis	1
			Missing	830
A0A0Z0R4A7	Surface protein G	125	*Escherichia coli*	1
			Leptospira borgpetersenii serovar Ballum	1
			Staphylococcus aureus	6
			Missing	992
A0A1P8KIE2	Genome polyprotein	3391	*Dengue virus*	36
			Dengue virus type 1	3
			Dengue virus type 2	834
A0A3G1IDJ8	Major capsid protein	540	*Norovirus GII*	618
			Norovirus GII.17	241
			Other Norovirus GII.17 (individual isolates)	75
A0A517E530	Gag protein (Fragment)	123	*Human immunodeficiency virus type 1*	799
A0A6M6AQ24	Pol protein (Fragment)	1009	*Human immunodeficiency virus type 1*	887
Q5MXE2	p72 protein (Fragment)	138	*African swine fever virus*	827
Q67953	Large envelope protein	445	*Hepatitis B virus*	994
W6JI24	Cytochrome c oxidase subunit 1	512	Various crustacean species (*Acantholobulus*)	956

In the case of UniRef90, several clusters include multiple organisms or related species (such as different viral serotypes or taxonomic subgroups), which contributes to the increased dispersion observed in subsequent analyses. These characteristics make UniRef90 particularly interesting for evaluating the model's ability to identify grouping patterns even in the presence of greater biological variability.

2.2. Feature Extraction and Representation

Feature extraction was performed by generating vector embeddings from protein sequences using representations derived from k-mer frequency matrices. Each protein sequence was decomposed into k-mers ($k = 3$) using the Seq2MC module, resulting in a frequency matrix that encodes local sequence patterns. This matrix was converted into an image using the MC2Image method, as described in previous studies [Câmara et al. 2022, De Souza et al. 2022, Coutinho et al. 2023]. The image dimensions depend on the value of k, typically producing a square matrix of size $\left\lceil \sqrt{21^k} \right\rceil \times \left\lceil \sqrt{21^k} \right\rceil$, which standardizes the representation for sequences of different

Table 2. Detailed distribution of UniRef90 groups (*clusters*).

UniRef Cluster ID	Gene/Protein	No. AA (Ref.)	Predominant species or group	No. Sequences
A0A0E3X5I8	Capsid protein	540	*Norovirus GII*	277
			Norovirus GII.17	220
			Other Norovirus (individual isolates)	remaining
A0A0U0T1S6	PPE family protein	73	Missing	990
			Mycobacterium orygis 112400015	1
			Mycobacterium tuberculosis	1
A0A0Z0R4A7	Surface protein G	125	Missing	891
			Staphylococcus aureus	6
			Staphylococcus schweitzeri	1
			Escherichia coli	1
			Leptospira borgpetersenii serovar Ballum	1
A0A1D8B9X1	Cytochrome c oxidase subunit 1	511	*Acantholobulus bermudensis*	3
			Other Acantholobulus	remaining
A0A290XZ71	Major capsid protein p72	133	*African swine fever virus*	900
A0A517E5J2	Gag protein (Fragment)	123	*Human immunodeficiency virus type 1*	925
P03141	Large envelope protein	400	*Hepatitis B virus*	895
			Other Hepatitis B virus isolates	remaining
P05777	Matrix protein 1	252	*Influenza A virus*	807
			Other Influenza A virus isolates	remaining
P29990	Genome polyprotein	3391	*Dengue virus type 2*	740
			Dengue virus	132
			Other Dengue virus isolates	remaining
Q8UTD3	Pol protein (Fragment)	1007	*Human immunodeficiency virus type 1*	925

lengths. The k-mer images were processed by a ViT model, resulting in vector embeddings with 768 dimensions. This approach enables the representation of sequences of variable length in a unified latent space, suitable for supervised learning algorithms. The use of image-based models allows parallel computation and use of accelerators, enabling the processing of large datasets. Figure 1 shows the workflow for transforming an amino acid sequence into a vector embedding via k-mer image representation and Vision Transformer.

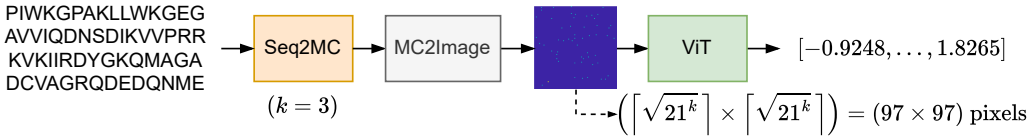

Figure 1. Workflow for transforming an amino acid sequence into a vector embedding through k-mer image representation and processing by a ViT. Example using the HV1 Pol protein (Fragment) from the species *Human immunodeficiency virus type 1*.

2.3. Machine Learning Algorithms and Training Procedures

Four supervised machine learning algorithms were evaluated in this study: Logistic Regression with L2 regularization, Random Forest, kNN, and XGBoost. All models received as input the vector embeddings extracted from the k-mer image representations.

Logistic Regression was applied with ridge (L2) regularization and regularization strength $C = 1$. The Random Forest algorithm was configured with 10 trees (`ntree` = 10) and a minimum node size of 5 (`nodesize` = 5), using the default number of attributes considered at each split (`mtry` = $\sqrt{p}$, where p is the number of features). For kNN, the number of neighbors was set to $k = 5$, using Euclidean distance and uniform

weights. The XGBoost model was trained with 100 trees (`nrounds` = 100), learning rate $\eta = 0.3$, maximum tree depth of 6 (`max_depth` = 6), L2 regularization parameter $\lambda = 1$, subsampling and column sampling set to 1.0, and random seed fixed for reproducibility. The hyperparameter settings for each algorithm are summarized in Table 3.

All models were trained and evaluated using stratified 5-fold cross-validation to ensure that all classes were proportionally represented in each fold. The dataset was divided into five subsets, with each subset used once as a test set and the remaining as the training set. The main evaluation metrics were accuracy, macro-averaged F1-score, log loss, AUC, Kappa, and balanced accuracy. Model training and evaluation were implemented using the `caret` framework in R, with data standardization performed where required by the algorithm.

Hyperparameters were selected to match the settings of previous experiments and allow direct comparison between algorithms. The workflow ensured reproducibility and consistency across all methods.

Table 3. Summary of hyperparameters used for each supervised learning algorithm.

Algorithm	Hyperparameters
Logistic Regression (L2)	Regularization type: L2 (ridge); $C = 1$
Random Forest	Number of trees (`ntree`): 10; Minimum node size (`nodesize`): 5; Attributes per split (`mtry`): $\sqrt{p}$ (default)
kNN	Number of neighbors (k): 5; Distance metric: Euclidean; Weight: Uniform
XGBoost	Number of trees (`nrounds`): 100; Learning rate (η): 0.3; Maximum depth (`max_depth`): 6; L2 regularization (λ): 1; Subsampling: 1.0; Column sampling: 1.0; Seed: Fixed for reproducibility

3. Results

3.1. UniRef100

Figures 2–5 present the confusion matrices obtained for each supervised learning algorithm evaluated on the UniRef100 dataset. Each cell shows the percentage of samples from the true class (row) predicted as each class (column). The confusion matrices for Logistic Regression (Figure 2), Random Forest (Figure 3), kNN (Figure 4), and XGBoost (Figure 5) indicate that most classes are predicted with high accuracy. The main diagonal in all cases concentrates the highest percentages, reflecting correct classification, while off-diagonal values are generally low, indicating few misclassifications. Minor confusions are observed in specific classes, particularly those with greater biological similarity or taxonomic overlap. These results illustrate the ability of the embedding-based approach

and the evaluated algorithms to distinguish between protein clusters with high precision in the UniRef100 dataset.

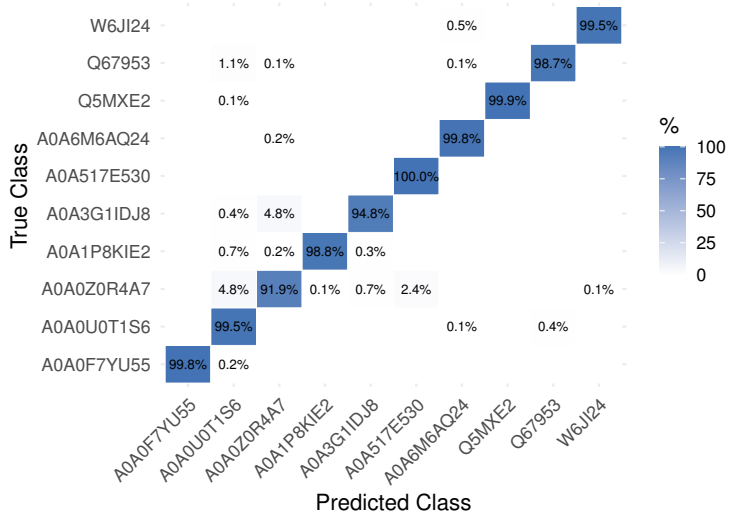

Figure 2. Confusion matrix for Logistic Regression on UniRef100.

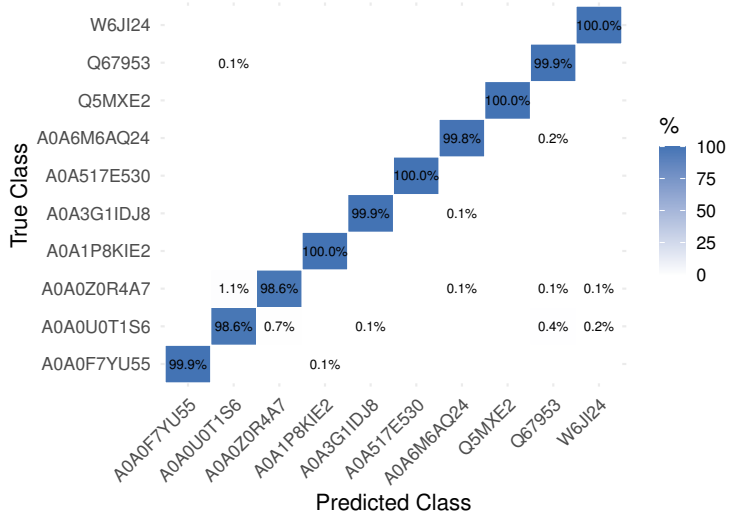

Figure 3. Confusion matrix for Random Forest on UniRef100.

Table 4 summarizes the quantitative performance metrics obtained by each supervised algorithm on the UniRef100 dataset. The reported values correspond to the mean results computed over five-fold cross-validation, considering accuracy, macro-averaged F1-score, log loss, AUC, Kappa, and balanced accuracy. The results allow a direct comparison of the predictive performance and robustness of the evaluated methods using the same feature representation and training protocol.

The results presented in Table 4 indicate that all algorithms achieved high predictive performance on the UniRef100 dataset. XGBoost obtained the highest values across all evaluated metrics, with an accuracy of 99.85%, macro F1-score of 0.9985, AUC of 1.0000, and the lowest log loss (0.0071). Random Forest and kNN also demonstrated

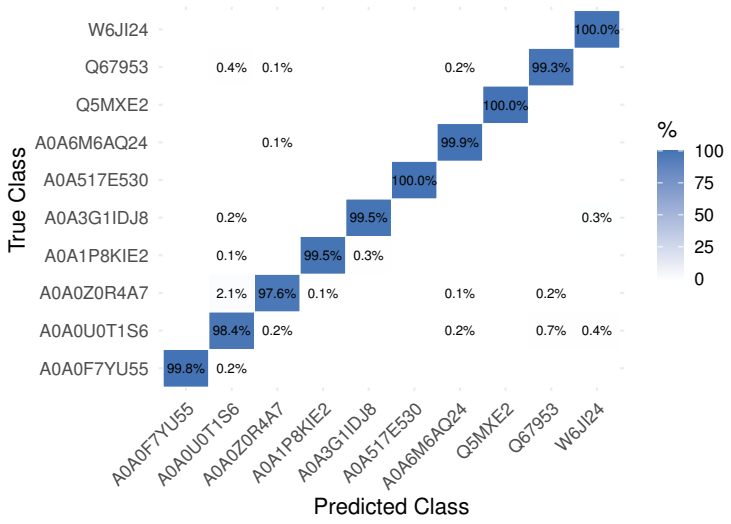

Figure 4. Confusion matrix for kNN on UniRef100.

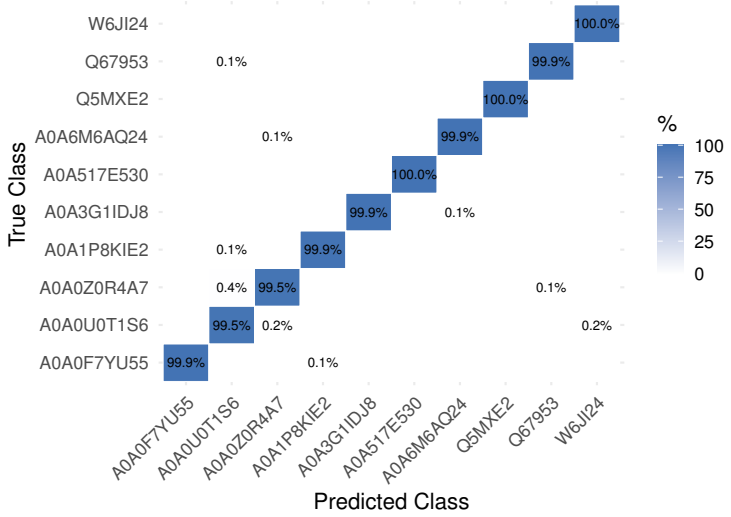

Figure 5. Confusion matrix for XGBoost on UniRef100.

strong results, with accuracies above 99% and balanced accuracy values close to 1. Logistic Regression, while slightly below the other models, maintained accuracy and F1-score values above 98%. The low log loss and high Kappa values observed for all classifiers further support the robustness and reliability of the predictions. These results demonstrate that the embedding-based feature representation, combined with the evaluated supervised learning methods, is effective for protein cluster classification when the underlying groups are cohesive, as in UniRef100.

3.2. UniRef90

Figures 6–9 show the confusion matrices obtained for each supervised learning algorithm on the UniRef90 dataset. As before, each cell displays the percentage of samples from the true class (row) that were predicted as each class (column). In this more heterogeneous dataset, the confusion matrices for Logistic Regression (Figure 6), Random For-

Table 4. Performance of supervised algorithms on the UniRef100 dataset.

Model	Accuracy	$F1_{macro}$	LogLoss	AUC	Kappa	Balanced Accuracy
Logistic Regression	0.9810	0.9813	0.4078	0.9963	0.9789	0.9892
Random Forest	0.9966	0.9966	0.0236	0.9998	0.9962	0.9981
kNN	0.9937	0.9937	0.0659	0.9991	0.9930	0.9964
XGBoost	0.9985	0.9985	0.0071	1.0000	0.9983	0.9991

est (Figure 7), kNN (Figure 8), and XGBoost (Figure 9) reveal an increase in off-diagonal values, reflecting a greater number of misclassifications compared to UniRef100. The main diagonal, however, remains dominant, indicating that most clusters are still accurately classified. Some clusters present more pronounced confusion, especially among groups with higher biological similarity or taxonomic overlap. These results highlight the challenge of protein cluster classification in datasets with greater internal variability and demonstrate the relative performance differences between algorithms under more complex conditions.

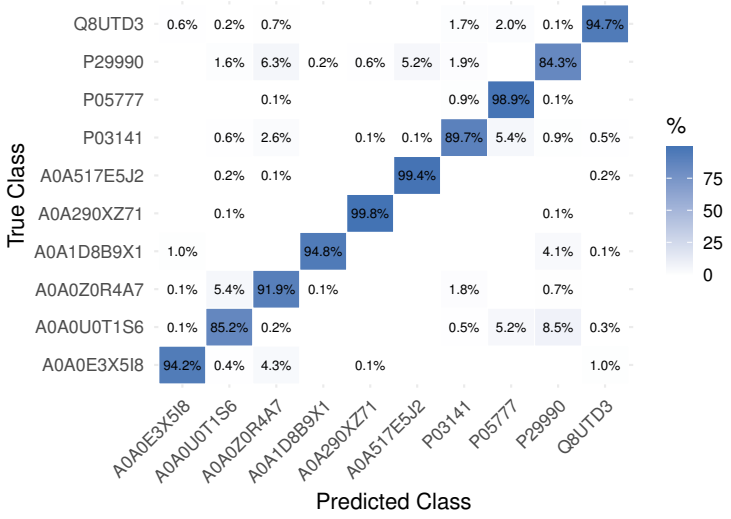

Figure 6. Confusion matrix for Logistic Regression on UniRef90.

Table 5 summarizes the main quantitative results obtained by each supervised algorithm on the UniRef90 dataset. The table reports the mean values of accuracy, macro-averaged F1-score, log loss, AUC, Kappa, and balanced accuracy, calculated over five-fold cross-validation. These metrics allow a direct comparison of the algorithms' predictive performance under conditions of greater internal variability and biological heterogeneity.

The results in Table 5 show a reduction in predictive performance for all algorithms when compared to UniRef100, reflecting the increased complexity and heterogeneity of UniRef90. XGBoost achieved the highest overall performance, with accuracy of 99.56%, macro F1-score of 0.9957, AUC of 0.9999, and the lowest log loss (0.0194). Random Forest and kNN also maintained high accuracy and F1 values, both above 98%, while Logistic Regression exhibited the most pronounced decrease in all metrics, with ac-

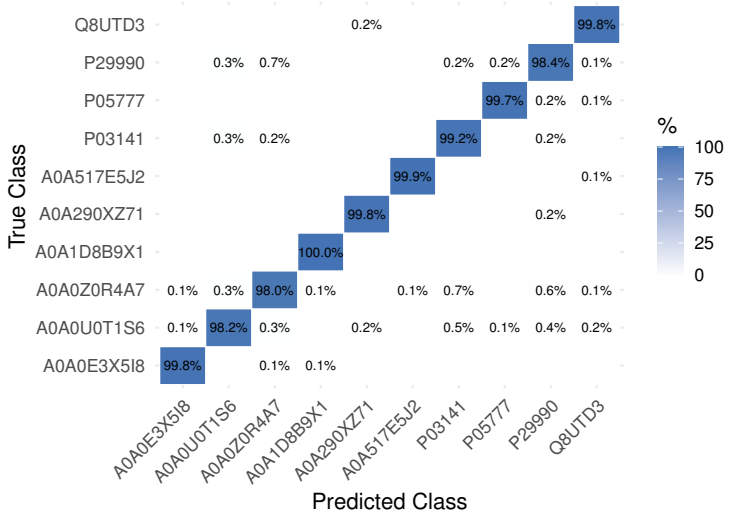

Figure 7. Confusion matrix for Random Forest on UniRef90.

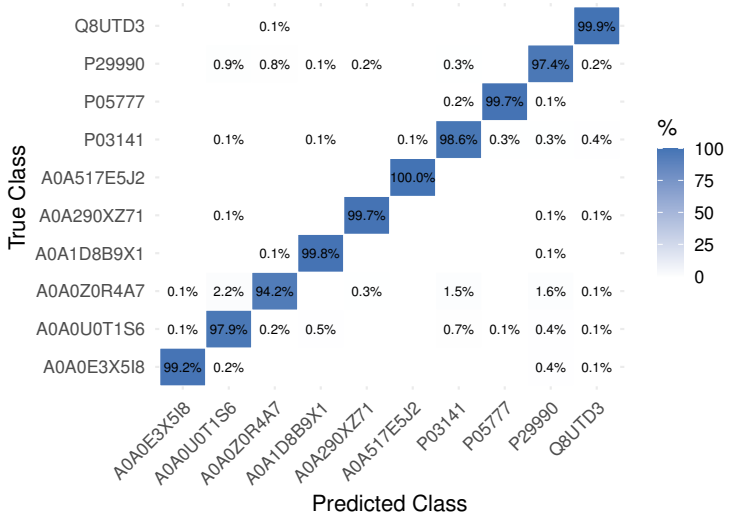

Figure 8. Confusion matrix for kNN on UniRef90.

curacy and F1 around 93%. Despite this reduction, all models demonstrated robustness, as indicated by the high Kappa and balanced accuracy values. The observed differences highlight the challenge of classifying more diverse and permissive clusters, and confirm the advantage of ensemble methods, especially XGBoost, for this type of problem.

4. Analysis and Discussion

The comparative analysis of results between the UniRef100 and UniRef90 datasets demonstrates the influence of cluster composition and biological variability on the performance of supervised classification algorithms.

For UniRef100, all evaluated methods achieved high accuracy and F1-score, with XGBoost, Random Forest, and kNN reaching values above 99% for both metrics, and Logistic Regression slightly lower, but still above 98%. These results reflect the structural cohesion of the clusters in UniRef100, which consist of unique sequences or closely

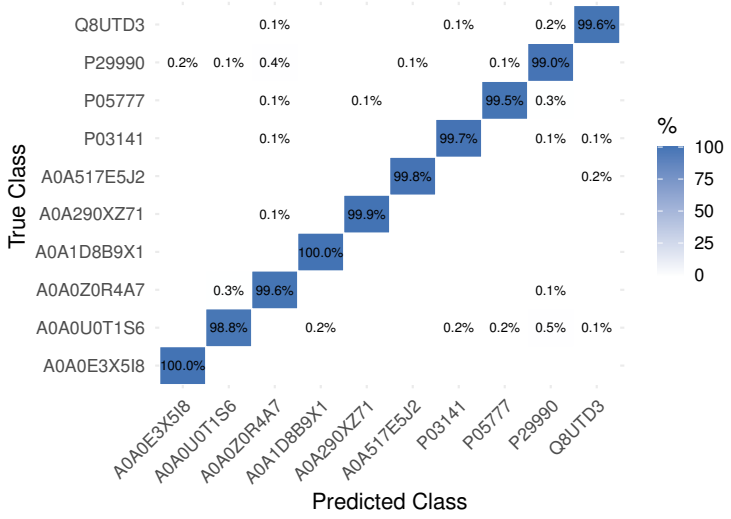

Figure 9. Confusion matrix for XGBoost on UniRef90.

Table 5. Performance of supervised algorithms on the UniRef90 dataset.

Model	Accuracy	$F1_{macro}$	LogLoss	AUC	Kappa	Balanced Accuracy
Logistic Regression	0.9308	0.9310	0.5714	0.9934	0.9231	0.9616
Random Forest	0.9927	0.9928	0.0732	0.9993	0.9919	0.9960
kNN	0.9861	0.9862	0.1578	0.9978	0.9845	0.9923
XGBoost	0.9956	0.9957	0.0194	0.9999	0.9952	0.9976

related proteins, as detailed in Table 1. Most clusters in this set correspond to a single species or viral subtype, such as the ”Matrix protein 1” group (A0A0F7YU55, Influenza A virus) or the ”p72 protein (Fragment)” (Q5MXE2, African swine fever virus), leading to less ambiguity in the classification process.

In contrast, the performance on UniRef90 reveals a consistent decrease in all metrics for every algorithm, most notably for Logistic Regression, which drops to around 93% accuracy and F1-score. Even for the tree-based methods, which remain robust, accuracy and F1-score are reduced compared to UniRef100. This decrease can be attributed to the greater heterogeneity of the clusters in UniRef90, as shown in Table 2. Many clusters in this dataset aggregate proteins from multiple organisms or variants, including different viral serotypes (e.g., ”Capsid protein” group, A0A0E3X5I8, containing various Norovirus types), or bacterial groups with significant taxonomic overlap. These mixed or ambiguous groups result in increased off-diagonal values in the confusion matrices, reflecting a higher rate of misclassification, particularly among clusters with biological or functional similarity.

The confusion matrices for both datasets further illustrate these differences. For UniRef100, correct predictions are concentrated along the main diagonal, with only minor off-diagonal values. In UniRef90, while the diagonal remains dominant, more significant off-diagonal values are observed, especially between clusters associated with similar or related organisms. These findings indicate that the quality of protein cluster classification

using embedding-based representations and supervised learning is strongly influenced by the internal structure and biological diversity of the clusters. Cohesive and taxonomically homogeneous groups, as in UniRef100, favor higher classification accuracy, while more permissive clusters with multiple organisms or functional variants, as in UniRef90, present additional challenges. The results also confirm the superior performance and robustness of ensemble methods, particularly XGBoost, in scenarios with increased data complexity.

5. Conclusions

This work evaluated supervised machine learning algorithms for protein cluster classification using embeddings derived from k-mer image representations processed by a ViT model. The comparative analysis using two reference datasets, UniRef100 and UniRef90, demonstrated that the quality and structure of the clusters strongly influence classification performance. In UniRef100, where clusters are more cohesive and taxonomically homogeneous, all evaluated algorithms achieved high accuracy, with tree-based ensemble methods, especially XGBoost, reaching the highest performance levels. In contrast, the more heterogeneous and permissive clusters in UniRef90 led to an increase in classification errors, particularly for Logistic Regression, although ensemble methods continued to provide robust results. The findings highlight the effectiveness of embedding-based representations combined with scalable machine learning approaches for automated protein sequence analysis. The results also emphasize the importance of considering biological diversity and group composition when designing classification tasks and benchmarking computational methods. Future work may explore the integration of additional biological information, the use of larger and more complex datasets, and the evaluation of new deep learning architectures to further advance protein classification in diverse bioinformatics contexts.

Acknowledgments

The authors would like to thank the National Council for Scientific and Technological Development (CNPq) and the Coordination for the Improvement of Higher Education Personnel (CAPES) for their support and funding. This work was funded by CNPq, the Brazilian Unified Health System (SUS), and the Ministry of Health (MS), through project no. 444306/2023-4 – *PharmaGenoNet: An Integrated Platform for Population Pharmacogenomics and Deep Learning for Predicting Drug-Target Interactions*.

References

Ahmed, B., Haque, M. A., Iquebal, M. A., Jaiswal, S., Angadi, U., Kumar, D., and Rai, A. (2023). Deepaprot: Deep learning based abiotic stress protein sequence classification and identification tool in cereals. *Frontiers in plant science*, 13:1008756.

Balamurugan, R., Mohite, S., and Raja, S. (2023). Protein sequence classification using bidirectional encoder representations from transformers (bert) approach. *SN Computer Science*, 4(5):481.

Blum, M., Andreeva, A., Florentino, L., Chuguransky, S., Grego, T., Hobbs, E., Pinto, B., Orr, A., Paysan-Lafosse, T., Ponamareva, I., Salazar, G., Bordin, N., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D., Letunic, I., Llinares-López, F., Marchler-Bauer,

A., Meng-Papaxanthos, L., Mi, H., Natale, D., Orengo, C., Pandurangan, A., Piovesan, D., Rivoire, C., Sigrist, C. A., Thanki, N., Thibaud-Nissen, F., Thomas, P., Tosatto, S. E., Wu, C., and Bateman, A. (2024). Interpro: the protein sequence classification resource in 2025. *Nucleic Acids Research*, 53(D1):D444–D456.

Coutinho, M. G. F., Câmara, G. B. M., Barbosa, R. d. M., and Fernandes, M. A. C. (2023). Sars-cov-2 virus classification based on stacked sparse autoencoder. *Computational and Structural Biotechnology Journal*, 21:284–298.

Câmara, G. B. M., Coutinho, M. G. F., Silva, L. M. D. d., Gadelha, W. V. d. N., Torquato, M. F., Barbosa, R. d. M., and Fernandes, M. A. C. (2022). Convolutional neural network applied to sars-cov-2 sequence classification. *Sensors*, 22(15):5730.

De Souza, J. G., Fernandes, M. A., and de Melo Barbosa, R. (2022). A novel deep neural network technique for drug–target interaction. *Pharmaceutics*, 14(3):625.

Lilhore, U. K., Simiaya, S., Alhussein, M., Faujdar, N., Dalal, S., and Aurangzeb, K. (2024). Optimizing protein sequence classification: integrating deep learning models with bayesian optimization for enhanced biological analysis. *BMC Medical Informatics and Decision Making*, 24(1):236.

Liu, G. (2024). Hybrid random forest and support vector machine model for protein sequence classification. In *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, pages 1120–1124.

Luo, Y. and Cai, J. (2024). Deep learning in proteomics informatics: Applications, challenges, and future directions. *arXiv preprint arXiv:2412.17349*.

Mall, R., Kaushik, R., Martinez, Z. A., Thomson, M. W., and Castiglione, F. (2025). Benchmarking protein language models for protein crystallization. *Scientific Reports*, 15(1):2381.

Murad, T., Ali, S., Chourasia, P., Mansoor, H., and Patterson, M. (2023). Circular arc length-based kernel matrix for protein sequence classification. In *2023 IEEE International Conference on Big Data (BigData)*, pages 1429–1437.

Perveen, H. and Weeds, J. (2025). Protein sequence classification using natural language processing techniques. *Discover Artificial Intelligence*, 5(1):1–25.

Suyunu, B., Dolu, Ö., and Özgür, A. (2025). evobpe: Evolutionary protein sequence tokenization. *arXiv preprint arXiv:2503.08838*.

Tasnim, F., Habiba, S. U., Mahmud, T., Nahar, L., Hossain, M. S., and Andersson, K. (2024). Protein sequence classification through deep learning and encoding strategies. *Procedia Computer Science*, 238:876–881.

Wang, Y., Zhang, Y., Zhan, X., He, Y., Yang, Y., Cheng, L., and Alghazzawi, D. (2024). Machine learning for predicting protein properties: A comprehensive review. *Neurocomputing*, 597:128103.

Zhang, M., Wan, F., and Liu, T. (2023). Drugfinder: Druggable protein identification model based on pre-trained models and evolutionary information. *Algorithms*, 16(6).