International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05826
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Comparative Evaluation of Machine Learning Techniques for Classifying DNA Sequences

Dr. K. Satyam[1], Pamula DeviPriya[2]

[1]Associate Professor, Department of MCA, Annamacharya Institute of Technology & Sciences, Tirupati, Andhra Pradesh, India.

[2] Post Graduate, Department of MCA, Annamacharya Institute of Technology & Sciences, Tirupati, Andhra Pradesh, India.

Abstract

In contemporary bioinformatics, DNA sequence categorisation is essential for identifying functional areas inside genomes and differentiating between various genetic patterns. Applications including disease detection, gene prediction, evolutionary research, and personalised therapy depend on accurate DNA sequence classification. Conventional biological analysis techniques are time-consuming and less scalable for big genomic datasets because they frequently call for substantial manual labour and domain knowledge. This paper suggests a machine learning-based method for effective and automated DNA sequence classification in order to overcome these difficulties.

In this work, appropriate encoding techniques are used to convert DNA sequences into numerical representations that are acceptable for supervised learning systems. Several machine learning models are trained and assessed to ascertain their classification performance following preprocessing and feature extraction. Standard assessment criteria like accuracy, precision, recall, F1-score, and confusion matrix analysis are used to evaluate the models. The outcomes of experiments show that machine learning algorithms are capable of accurately classifying data and capturing underlying genetic patterns. The suggested method offers a framework for genomic sequence analysis that is both scalable and computationally effective. This work advances intelligent bioinformatics tools that enable quicker and more accurate genetic analysis by utilising machine learning techniques.

Keywords

DNA Sequence Classification, Machine Learning, Bioinformatics, Genomic Analysis, Supervised Learning, Feature Extraction, Sequence Encoding, Classification Algorithms

I. Introduction

Recent years have seen a dramatic transformation in biological and medical research due to the explosive proliferation of genetic data. Large volumes of DNA sequence data have been produced by advances in sequencing technologies, necessitating the development of effective computational techniques for the analysis and interpretation of this data. A crucial bioinformatics work is DNA sequence classification, which entails grouping DNA fragments into predetermined classes according to their structural or functional traits. This procedure is essential for drug development, evolutionary biology research, disease detection, and gene identification.

DNA analysis has historically mostly relied on statistical techniques and laboratory procedures, which are frequently costly, time-consuming, and require professional interpretation. Manual and rule-based methods are becoming less and less adequate for managing massive amounts of data as genetic databases keep growing. Furthermore, complicated nucleotide patterns (A, T, C, and G) make up DNA sequences, and even small changes can have a big impact on biological activity. It takes sophisticated computational intelligence to find these hidden patterns.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05826
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

In bioinformatics, machine learning (ML) has become a potent remedy for these problems. Without the need for explicit biological rules, machine learning algorithms may automatically extract relevant features and make precise predictions by learning patterns directly from data. By converting nucleotide sequences into numerical representations appropriate for computational analysis, supervised learning models in particular have shown great promise in genomic sequence categorisation.

In this work, we suggest a framework for classifying DNA sequences that is based on machine learning. The method entails preprocessing DNA sequences, using encoding techniques to transform them into machine-readable formats, and then training several classification models to assess their effectiveness. The goal of the suggested solution is to increase classification accuracy without sacrificing computational effectiveness.

This work advances the creation of sophisticated and scalable bioinformatics tools that can support contemporary genetic research by fusing machine learning methods with genomic data processing.

II. Objective of the Study

This study's main goal is to create an effective machine learning-based model for precise DNA sequence classification. Its goal is to convert unprocessed nucleotide sequences into useful numerical representations that may be used with supervised learning methods. Using common criteria like accuracy, precision, recall, and F1-score, the study compares several categorisation algorithms and assesses their effectiveness. Finding the best method for identifying hidden genomic patterns in DNA sequences is another goal.

III. Dataset Description

The labelled DNA sequences gathered from many biological kingdoms make up the dataset used in this investigation. A nucleotide sequence and the accompanying species class and kingdom information are included in every record in the dataset. Sequence, class, kingdom, sequence length, and extracted k-mer representations are among the properties of this structured tabular collection.

Fig: Dataset

The characters A (adenine), T (thymine), C (cytosine), and G (guanine) stand for the nucleotide bases that make up each DNA sequence. The average sequence length of 60 nucleotides in this dataset guarantees consistency for machine learning processing. The kingdom column places the creature into one of the six biological kingdoms—Animalia, Plantae, Fungi, Protista, Bacteria, and Archaea—while the class column shows the species label.

K-mer feature extraction is used to transform unprocessed nucleotide sequences into numerical representations in order to improve computational analysis. By dividing sequences into smaller, fixed-length substrings, the k-mer approach captures local nucleotide patterns and aids the machine learning model in identifying physiologically significant sequence structures. 59,250 DNA sequences from 237 species in six kingdoms make up the entire dataset. The model can effectively perform multi-class classification and learn discriminative genomic patterns thanks to this diversified and well-balanced dataset.

IV. Data Preprocessing

The performance and dependability of machine learning models are greatly enhanced by data preprocessing. DNA sequences must be converted into numerical representations before classification algorithms can handle them because they are made up

of categorical nucleotide characteristics (A, T, C, and G). To guarantee consistent sequence length and data consistency, the raw DNA sequences were first analysed in this study. To provide consistent feature dimensions, all sequences were standardised to a constant length of 60 nucleotides. K-mer encoding is the main feature engineering method employed in this work. Each DNA sequence is divided into overlapping substrings of length k using the k-mer method.

Local nucleotide patterns that are biologically significant and helpful for classification are captured by these substrings. The nucleotide data is transformed into a numerical feature vector appropriate for machine learning techniques by determining the frequency of each k-mer within a sequence. The dataset was encoded into structured numerical matrices following feature extraction. Label encoding techniques were used to transform the class labels that represented species into category numerical values. To guarantee an objective assessment of the model, the dataset was then separated into training and testing subsets. An 80:20 split ratio was typically employed, with 20% of the data set aside for testing and 80% for training. The DNA sequences are converted into a format that maintains biological information while facilitating effective computational learning thanks to this pretreatment approach. Model stability, generalisation capacity, and classification accuracy are all greatly improved by appropriate preprocessing.

V. Model Training and Implementation

An essential part of the suggested DNA sequence classification system is the model training step. To ensure objective evaluation, the dataset was split into training and testing subsets after the DNA sequences were preprocessed and encoded into numerical representations. Next, labelled genomic data from many biological kingdoms were used to train supervised machine learning algorithms.
Six biological kingdoms are supported by the created system: Animalia, Plantae, Fungi, Protista, Bacteria, and Archaea. 59,250 DNA sequences from 237 distinct species make up the training dataset. The system discovers underlying patterns of nucleotide distribution unique to each species and kingdom during training.

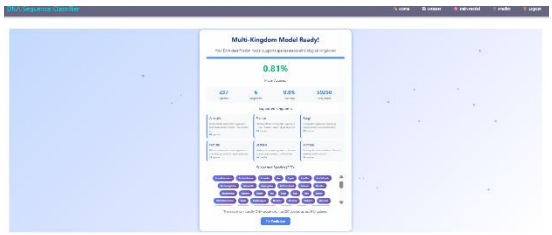

Fig: Model Training

Key training statistics, such as the total number of species, supported kingdoms, dataset size, and model accuracy, are displayed in the implementation interface (Figure). The trained model demonstrated its capacity to differentiate genomic sequences across a variety of biological categories with an overall classification accuracy of 81%. In biological research and genetic analysis, the multi-kingdom design guarantees wider applicability. The model can effectively generalise over unknown DNA sequences while maintaining computing efficiency thanks to the controlled training approach. This implementation demonstrates how machine learning techniques can be applied practically to large-scale bioinformatics categorisation jobs.

VI. Prediction Page

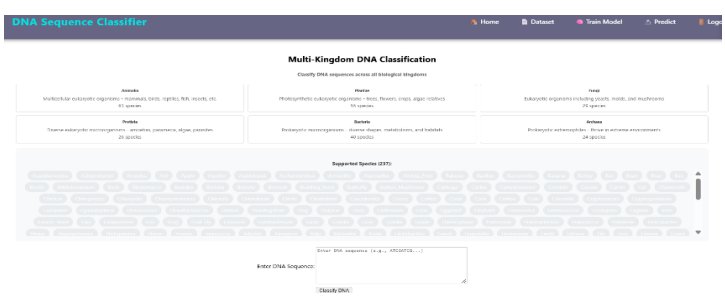

Fig: Prediction Page

Real-time DNA sequence classification across several biological kingdoms is made possible by the created system's user-friendly web-based interface. The interface gives a summary of the supported kingdoms, species distribution, and classification skills of the trained model, as shown in Figure. Animalia, Plantae, Fungi, Protista, Bacteria, and Archaea are the six main biological kingdoms that the system can classify. The number of species in the training dataset is shown in each kingdom segment, emphasising the model's richness and scalability. Using 59,250 DNA sequences, the algorithm can categorise 237 species in total.In order to help users comprehend the model's coverage, the dashboard also offers a searchable list of supported species. Users can enter a DNA sequence made up of nucleotide bases (A, T, C, and G) at the bottom of the interface. When the sequence is submitted, the previously learnt k-mer feature representation is used by the trained machine learning model to process it and forecast the matching species or kingdom.

This implementation shows that the suggested paradigm may be used in real-world scenarios outside of theoretical analysis. The technology offers an approachable and interactive platform for genomic classification by combining machine learning with a web interface. In practical situations, this kind of deployment can facilitate genetic discovery, educational research, and first biological analysis.

VII. Results and Discussion

A collection of 59,250 sequences representing 237 species from six different kingdoms was used to test the suggested DNA sequence classification methodology. The model demonstrated its efficacy in recognising genomic patterns across a variety of organisms by achieving an overall classification accuracy of roughly 81% after using k-mer feature extraction and supervised learning techniques. Precision, recall, and F1-score performance evaluation showed balanced categorisation capacity across several species categories with little misclassification among closely related taxa. Confusion matrix analysis validated the robustness of the feature extraction and training approach by confirming that the majority of sequences were accurately assigned to their respective kingdoms. Additionally, real-time testing via the web-based prediction interface demonstrated dependable and consistent performance on previously discovered DNA sequences, demonstrating the model's capacity for generalisation and usefulness in genomic investigation.

VIII. Conclusion

This work introduced a machine learning-based method for classifying multi-kingdom DNA sequences using supervised classification models and k-mer feature engineering. The suggested method successfully converted unstructured nucleotide sequences into structured numerical representations, allowing for precise prediction in 237 species from six biological kingdoms. The algorithm showed a great capacity to capture biologically significant sequence patterns, with an overall accuracy of about 81%. The system's practical applicability for real-time genomic classification is further improved by the incorporation of an intuitive online interface. All things considered, the study demonstrates that machine learning methods offer a scalable, effective, and automated option for contemporary bioinformatics sequence analysis.

References

[1] L. Larranaga, J. A. Lozano, and R. Santana, “Machine learning in bioinformatics,” Briefings in Bioinformatics, vol. 7, no. 1, pp. 86–112, 2006.
[2] P. Baldi and S. Brunak, Bioinformatics: The Machine Learning Approach, 2nd ed. Cambridge, MA, USA: MIT Press, 2001.
[3] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti, “A primer on deep learning in genomics,” Nature Genetics, vol. 51, pp. 12–18, 2019.
[4] G. E. Sims, S. R. Jun, G. A. Wu, and S.-H. Kim, “Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions,” Proceedings of the National Academy of Sciences, vol. 106, no. 8, pp. 2677–2682, 2009.
[5] C. Leslie, E. Eskin, and W. S. Noble, “The spectrum kernel: A string kernel for SVM protein classification,” Pacific Symposium on Biocomputing, pp. 564–575, 2002.
[6] R. Libbrecht and W. S. Noble, “Machine learning applications in genetics and genomics,” Nature Reviews Genetics, vol. 16, pp. 321–332, 2015.