International Research Journal of Multidisciplinary Scope (IRJMS), 2025; 6(3):1180-1189

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2025.v06i03.04915

Efficient DNA Sequence Classification through Machine-Learning Techniques

Papri Ghosh[1], Subhram Das[1]*, Debrupa Pal[2], Soumyabrata Saha[3], Suparna Dasgupta[3]

[1]Computer Science and Engineering, Narula Institute of Technology, Kolkata, India, [2]Computer Application, Narula Institute of Technology, Kolkata, India, [3]Information Technology, JIS College of Engineering, Kalyani, India. *Corresponding Author's Email: subhram@gmail.com

Abstract

In the domain of computational biology and biomedical data analysis, classifying DNA sequences is a significant challenge. Identifying and classifying DNA sequences of various species is of utmost importance. Various Machine Learning (ML) techniques have been successfully applied to this task recently. This study introduces a new approach for effectively categorizing valid DNA sequences from unrelated sequences using different ML techniques. The valid datasets were systematically collected from the NCBI database, while the unrelated datasets were generated using random techniques. Various ML techniques were then applied to distinguish between these two categories. It was observed that Gradient Boosting Machine (GBM) performed the best, achieving 0.971 accuracy and a 0.975 F1 score. The outcome of XGBoost is also good that achieving 0.935 Accuracy and 0.93 F1 Score. It is also observed that this method consistently achieves the best execution time when compared to other existing machine learning methods. The results were also verified using a Phylogenetic Tree constructed through Clustal Omega, a well-known traditional alignment-based method for DNA sequence comparison. In both cases, the results were consistent, although Clustal Omega had a much higher execution time compared to the present method. Therefore, the proposed technique significantly enhances the efficiency of DNA sequence classification.

Keywords: Bioinformatics, Clustal Omega, DNA Sequences Comparison, Machine Learning, Phylogenetic Tree.

Introduction

The comparison of DNA sequences plays a crucial role in understanding the genetic relationships among different species (1, 2). Traditional researchers have relied on methods such as sequence alignment and phylogenetic analysis to classify organisms based on their genetic information (3-6). Traditional methods are often slow and resource-intensive, particularly with large-scale datasets (7, 8). The rise of high-throughput sequencing has led to an explosion in genetic data, highlighting the need for faster and more accurate techniques for DNA sequence analysis (9, 10). Machine learning (ML) offers powerful tools for handling large-scale biological data and has the potential to revolutionize the field of genomics (11, 12). By leveraging patterns in DNA sequences, ML algorithms can classify species with high accuracy and predict evolutionary relationships. Recent advancements in ML, including deep learning techniques, have shown promise in addressing the complexities of genomic data, enabling the development of more robust and scalable classification systems. In this study, we explore the application of machine learning methods to differentiate species based on their DNA sequence. To this end, we employ both original DNA sequences from various species and artificially generated unrelated sequences to train and test our models. The use of unrelated sequences serves as a control to assess the robustness and specificity of the ML algorithms in distinguishing genuine biological differences from random sequence variations. We pre-process DNA, extract features, train models, and evaluate using accuracy, precision, recall, F1. DNA sequence comparison plays an essential role in Sequence Data Analysis (SDA), particularly in predicting sequences and exploring evolutionary relationships (13, 14). Applying machine learning (ML) techniques to actual G4 datasets enables the extraction of relevant information from DNA sequences (15). This process is significantly

(Received 11th April 2025; Accepted 29th June 2025; Published 29th July 2025)

enhanced by feature engineering, which allows ML classifiers to predict specific DNA sequence comparisons. For instance, Touati *et al.,* were attentive to the classification of the helitron family utilizing various machine learning algorithms and feature mining from FASTA DNA sequences (16). The study by Hamed *et al.,* presents a DNA sequence categorization model that coalesces machine learning with pattern-matching techniques, showing high accuracy and efficiency, especially using SVM Linear across different pattern lengths (17). Their work underscores the importance of feature engineering in capturing helitron-specific characteristics.

Furthermore, the research by Ryu *et al.,* introduced Maximal Average Shift (MAS) algorithms to enhance the efficiency of Pattern Scan Order through the use of q-grams (18). The approach in a few research papers demonstrated superior performance compared to traditional methods (17, 19-22). This body of work collectively highlights the key role of sophisticated ML techniques and feature engineering in advancing our understanding of DNA sequences and their functional and evolutionary properties.

The study by Zhang *et al.,* analyses the Chrysanthemum nankingense genome, noting a low presence of tandem repeats (1.02%) and an elevated number of low-density sequences, probably from identical elements (23). These findings offer insights into the genome's evolution and structure, aiding the understanding of diversity in both diploid and polyploid Chrysanthemum species. Furthermore, the research by Hamed *et al.,* introduces a novel DNA sequence classification model that integrates machine learning and pattern-matching, with SVM Linear demonstrating superior accuracy and efficiency (17). The model outperforms existing methods, highlighting the significance of pattern length in DNA sequence classification, and holds potential for applications in drug detection, customized medicine, and disease finding.

The extensive body of research on DNA sequence analysis, including this study, underscores the ongoing efforts and diverse approaches in the field, highlighting the importance of continually refining models for improved accuracy and efficiency across various applications.

Methodology

ML Methodology for FASTA DNA Sequences

In this study, a database was developed using FASTA-formatted DNA sequences (Table 1) to assess the performance of multiple machine learning algorithms. This dataset comprises mitochondrial DNA sequences from 41 mammalian species, which serve as critical genetic markers for species identification, evolutionary analysis, and biodiversity assessment. These sequences are conserved yet exhibit species-specific variations, making them ideal for studying phylogenetic relationships, tracing maternal lineage, and investigating evolutionary divergence among mammals. The aim was to implement automated classification models tailored to specific DNA sequences and construct a biological database for classification tasks. Various models, including K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), Support Vector Machines (SVM), Logistic Regression (LR), Gradient Boosting Machine (GBM), XGBoost, Artificial Neural Network (ANN), and K-means classifier, were applied to the dataset (24-39). Their performance was calculated using accuracy, precision, recall, F1 score, and execution time to recognize the best appropriate model. Efficient DNA sequence comparison plays a vibrant role in molecular biology and genetics, facilitating faster and more accurate sequence analysis for diverse research applications.

Table 1: Information on the 41 Mammalian Genomes

Sl. No.	Accession Number	Description
1	V00662.1	Human
2	D38116.1	Pigmy chimpanzee
3	D38113.1	Common chimpanzee
4	D38114.1	Gorilla
5	X99256.1	Gibbon
6	Y18001.1	Baboon
7	AY863426.1	Vervet monkey

Sl. No.	Accession Number	Description
8	D38115.1	Bornean orangutan
9	NC_2083.1	Sumatran orangutan
10	U20753.1	Cat
11	U96639.2	Dog
12	AJ002189.1	Pig
13	AF010406.1	Sheep
14	AF533441.1	Goat
15	V00654.1	Cow
16	AY488491.1	Buffalo
17	EU442884.2	Wolf
18	EF551003.1	Tiger
19	EF551002.1	Leopard
20	X97336.1	Indian Rhinoceros
21	Y07726.1	White Rhinoceros
22	DQ402478.1	Black Bear
23	AF303110.1	Brown Bear
24	AF303111.1	Polar Bear
25	EF212882.1	Giant Panda
26	AJ001588.1	Rabbit
27	X88898.2	Hedgehog
28	NC_2764.1	MacacaThibet
29	AJ238588.1	Squirrel
30	AJ001562.1	Dormouse
31	X72204.1	Blue whale
32	NC_5268.1	Bowhead Whale
33	NC_7441.1	Chiru
34	NC_8830.1	Common warthog
35	NC_1788.1	Donkey
36	NC_1321.1	Fin Whale
37	NC_5270.1	Gray Whale
38	NC_1640.1	Horse
39	NC_5275.1	Indus River Dolphin
40	NC_006931.1	North Pacific Right Whale
41	NC_010640.1	Taiwan serow

The proposed technique involves several key phases. It begins with pre-processing the data of the FASTA DNA sequence, which includes cleaning and filtering to remove noise and irrelevant details. After pre-processing, feature extraction is performed to identify and extract relevant features. Separate features are operated to build a classification model that assesses DNA sequences upon their resemblance to the query patterns. This step involves applying numerous machine learning algorithms, encompassing both supervised and unsupervised approaches, to extend a robust and precise classification framework. Throughout this approach, the integration of ML methodologies enhances the efficacy of searching DNA sequences for a specific pattern, facilitating the extraction of perilous evidence essential for diverse applications in bioinformatics and genome sequence comparison techniques. By optimizing the identification and classification processes, the technique supports advancements in genome research and biotechnological applications, underscoring its significance in contemporary biological data analysis.

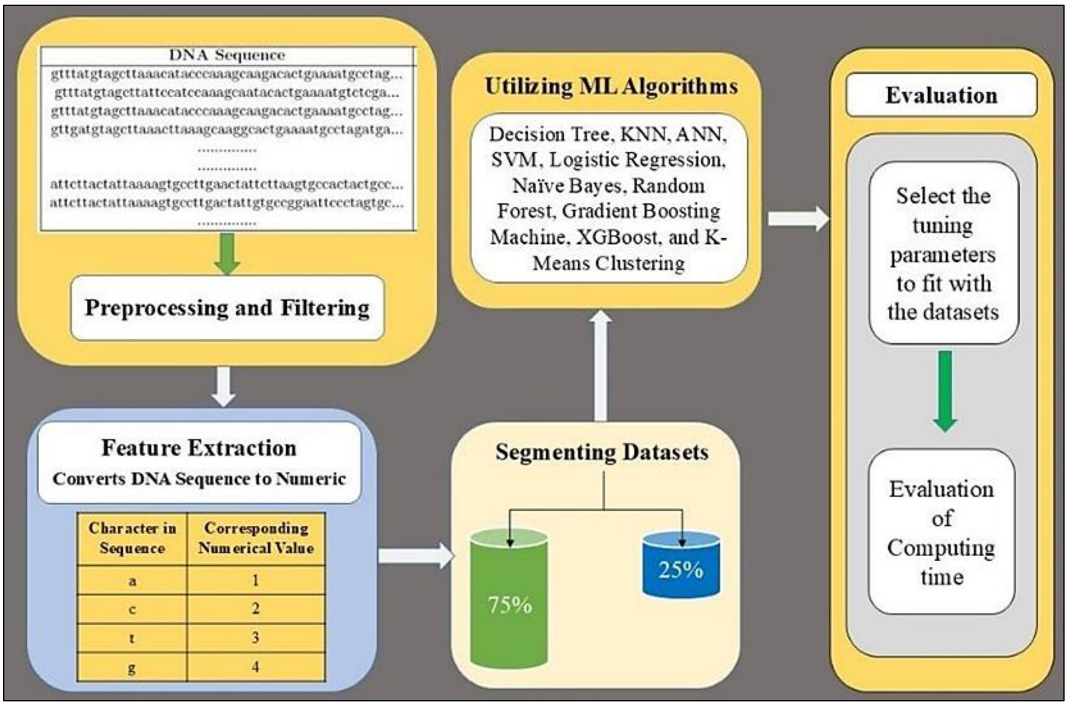
DNA Sequence
gtttatgtagcttaaacatacccaaagcaagacactgaaaatgcctag...
gtttatgtagcttattccatccaaagcaatacactgaaaatgtctcga...
gtttatgtagcttaaacatacccaaagcaagacactgaaaatgcctag...
gttgatgtagcttaaacttaaagcaaggcactgaaaatgcctagatga...
attcttactattaaaagtgccttgaactattcttaagtgccactactgcc...
attcttactattaaaagtgccttgactattgtgccggaattccctagtgc...
Preprocessing and Filtering
Feature Extraction
Converts DNA Sequence to Numeric
Character in Sequence
Corresponding Numerical Value
a
1
c
2
t
3
g
4
Segmenting Datasets
75%
25%
Utilizing ML Algorithms
Decision Tree, KNN, ANN, SVM, Logistic Regression, Naïve Bayes, Random Forest, Gradient Boosting Machine, XGBoost, and K-Means Clustering
Evaluation
Select the tuning parameters to fit with the datasets
Evaluation of Computing time

Figure 1: Structure of the Proposed Work

Figure 1 depicts the comprehensive structure outlining various phases of our ML approach for the DNA sequence comparison study.

Dataset

In this study, the dataset was obtained from the National Center for Biotechnology Information (NCBI), consisting of FASTA files that include genomic sequence data. The dataset contains over twelve million characters, mainly compiled of the nucleotides A, C, T, and G. Each FASTA file includes one or more entries, with each entry generally corresponding to a full or partial DNA sequence. The dataset collection process is as follows:

Number of Samples: The dataset includes a total of X genomic sequences (replace X with the actual number of samples).

Sequence Length: The length of each sequence varies, with the longest sequence containing up to 12 million characters.

Features: The dataset is primarily characterized by the nucleotide sequences, which are categorical in nature (A, C, T, G). These sequences serve as the features for downstream analysis.

The categorical nature of genomic sequences presents unique challenges for analysis, particularly due to the high dimensionality and the potential for unbalanced data distribution. An unbalanced dataset issue was identified, where certain sequences or nucleotide combinations were over-represented. To address this, extensive pre-processing was necessary. This involved not only cleaning the data to remove any redundant or irrelevant sequences but also implementing techniques to balance the dataset effectively.

While tools like the `GET_DUMMIES` function in the pandas library can assist in handling categorical data, the contribution of the authors extends beyond basic preprocessing. The preprocessing steps taken by the authors were meticulously tailored to handle the complexities of genomic data, ensuring that the dataset was in optimal condition for analysis. This included custom scripts for sequence alignment, filtering out low-quality sequences, and implementing advanced techniques to manage the dataset's scale and complexity. The authors' contribution lies in developing a comprehensive preprocessing pipeline that addresses the specific challenges posed by such large and intricate genomic datasets, going beyond standard library functions to ensure data quality and relevance.

Pre-Processing of Data

The DNA sequence data obtained from Table 1 is transformed for training via various ML algorithms. Based on the data type and the requirement of the output, the proper existing data transformation technique can be applied. In this study, the DNA sequences are converted from FASTA files to CSV files. The biological DNA sequences are then categorized and label 0 or 1 is assigned. If the DNA sequence is relevant to the existing species given in the Table 1, then it is assigned to 1; otherwise, the unrelated sequences are assigned to 0. A sample dataset is given in Table 2.

Table 2: Sample Data Transformation

DNA Sequence	Label
gtttatgtagcttaaacatacccaaagcaagacactgaaaatgcctag...	1
gtttatgtagcttattccatccaaagcaatacactgaaaatgtctcga...	1
gtttatgtagcttaaacatacccaaagcaagacactgaaaatgcctag...	1
gttgatgtagcttaaacttaaagcaaggcactgaaaatgcctagatga...	1
....	...
....	...
attcttactattaaaagtgccttgaactattcttaagtgccactactgcc...	0
attcttactattaaaagtgccttgactattgtgccggaattccctagtgc...	0
....	...

After the conversion process, it is ensured that all the sequence is assigned to either “0” or “1”. Table 2 provides a sample of the cleaned dataset in CSV format.

Feature Mining

The selection of correct features is crucial for enhancing the categorization accuracy and reducing the preparation time for the model. The ML algorithms cannot be performed on Sequence data in text formats. Therefore, this study needed to transform it into the required format. Hence, the numerical presentation is selected, and the entire DNA sequences are converted into numerical data. Here ‘a’ is denoted as 1, ‘c’ is denoted as 2, ‘t’ is denoted as 3 and ‘g’ is denoted as 4. To accomplish the above, GET_DUMMIES function of the Pandas Library is utilized (40-42). It helps to alter FASTA DNA Sequences to the corresponding numeric variable that encodes categorical evidence. After this conversion, the training and testing via different ML algorithms can be applied.

Training and Testing

To ensure unbiased model evaluation, the dataset was divided into 75% training and 25% testing subsets using the *train_test_split* function with stratified sampling, maintaining class proportions. The training set was used to build and tune models, while the testing set assessed generalization performance on unseen data.

After pre-processing, multiple classification algorithms were applied to analyse FASTA DNA sequences. Ten machine learning models - KNN, RF, DT, LR, NB, SVM, GBM, XGBoost, K-Means, and ANN - were implemented. Model performance was evaluated using accuracy, precision, recall, F1-score, and classification reports via SKLEARN metrics. Each algorithm offers distinct approaches; for example, KNN classifies based on the majority vote of k nearest neighbors. The prediction is made by:

$$\hat{y} = mode(y_1, y_2, \ldots, y_k) \qquad [1]$$

RF and DT utilize ensemble learning techniques. RF constructs numerous decision trees during training and harvests the methods of the classes for division or the mean prediction for regression:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \qquad [2]$$

where $h_t(x)$ signifies the prediction from the t^{th} tree in the forest. The DT algorithm makes decisions based on feature splits that maximize information gain or minimize impurity, often using criteria like Gini impurity or entropy. LR is a probabilistic model for binary outcomes. The probability that input x belongs to class 1 is modeled as:

$$P(x) = \sigma\,(w^T x + b) = \frac{1}{1+e^{-(w^T x+b)}} \qquad [3]$$

where σ denotes the sigmoid function, ω is the weight vector, and b is the bias. NB applies Bayes’ theorem with intense assumptions between features. The posterior probability for classy given input *x* is given by:

$$P(y|x) = \frac{P(y)P(y)}{P(x)} \qquad [4]$$

SVM excels in high-dimensional spaces by identifying the optimal hyperplane that maximizes the separation margin between classes:

$$maximize \frac{2}{||\omega||}\ subject\ to\ y_i\ (\omega^T x_i + b) \geq 1\ for\ all\ i \qquad [5]$$

GBM and XGBoost, both boosting methods, iteratively refine weak learners by minimizing a differentiable loss function L(y, $F_m(x)$) at each iteration to create a strong predictive model:

$$F_{m+1}(x) = F_m(x) + \gamma_m h_m x \qquad [6]$$

where γ_m is the step size, and $h_m(x)$ is the weak learner at iteration *m*.

K-Means Clustering is employed for partitioning the dataset into *k* distinct clusters, minimizing the sum of squared distances between data points and their respective cluster centroids:

$$minimize \sum_{i=1}^{k} \sum_{x_i \in c_i} ||x_i - \mu_i||^2 \qquad [7]$$

where μ_i is the centroid of cluster C_i.

Finally, ANN, encouraged by the human brain, uses layers of consistent nodes (neurons) to discover complex patterns in data. The output of a neuron is typically given by:

$$a_j = \sigma\left(\sum_{i=1}^{n} \omega_{ij} x_i + b_j\right) \qquad [8]$$

Here, a_i represents the activation of neuron *j*, w_{ij} are the weights, x_i are the input values, b_j is the bias term, and σ denotes the activation function. These diverse techniques ensure a comprehensive approach to analyzing and interpreting data, catering to various aspects of the problem at hand. Model evaluation relies on four key metrics derived from the confusion matrix: accuracy, precision, recall, and F1-score. The confusion matrix summarizes true/false positives and negatives, reflecting actual vs. predicted outcomes. Accuracy measures overall correctness; precision and recall assess positive prediction quality and completeness, respectively; F1-score balances both, aiding in binary classification.

Results and Discussion

Table 3 presents the results of ten Machine learning algorithms, including their execution times. The metrics provided are accuracy, precision, recall, F1 score, and execution time. The Gradient Boosting Machine (GBM) demonstrates the highest performance, achieving a 0.975 F1 score and a 0.971 accuracy. It indicates superior performance compared to the other algorithms. The outcome of XGBoost is 0.935 accuracy of and 0.93 F1 score. Notably, XGBoost has a shorter execution time of 11.135 seconds, compared to GBM's 11.684 seconds as depicted in Figure 2 and the execution time comparison obtained in different ML Algorithms for DNA Sequence is depicted in Figure 3.

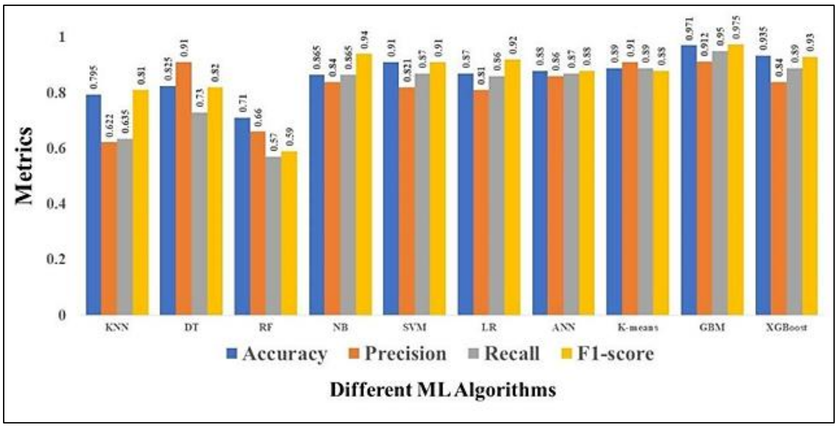

Figure 2: ML Algorithm Metric for a DNA Sequence

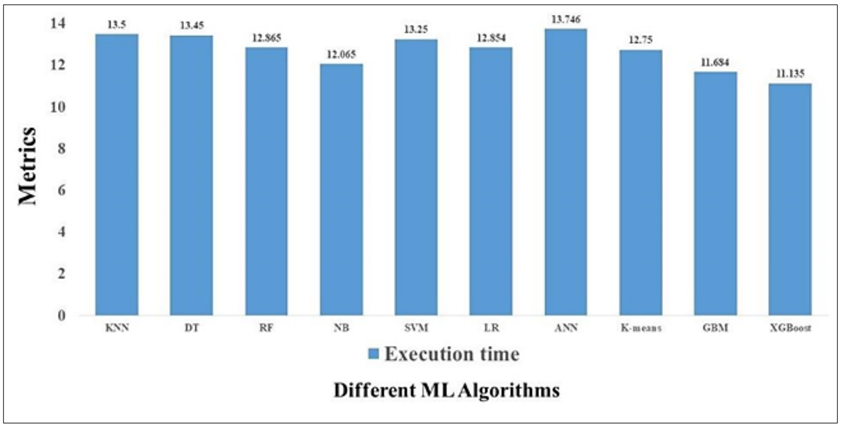

Figure 3: Execution Time in Different ML Algorithms for a DNA Sequence

Table 3: Results of Different ML Algorithm for DNA Sequence Classification

ML Algorithm	Accuracy	Precision	Recall	F1-Score	Execution Time
KNN	0.795	0.622	0.635	0.81	13.5
DT	0.825	0.91	0.73	0.82	13.45
RF	0.71	0.66	0.57	0.59	12.865
NB	0.865	0.84	0.865	0.94	12.065
SVM	0.91	0.821	0.87	0.91	13.25
LR	0.87	0.81	0.86	0.92	12.854
ANN	0.88	0.86	0.87	0.88	13.746
K-means	0.89	0.91	0.89	0.88	12.75
GBM	0.971	0.912	0.95	0.975	11.684
XGBoost	0.935	0.84	0.89	0.93	11.135

This study highlights that while several algorithms perform well, GBM and XGBoost are particularly effective for analyzing FASTA DNA sequences, with GBM achieving the best overall performance and XGBoost offering a competitive balance of accuracy and speed.

The eight other algorithms also demonstrate strong performance. However, upon this study, we can conclude that XGBoost and Gradient Boosting Machine (GBM) are particularly effective for analyzing FASTA DNA sequences. Their superior performance in this context suggests they may be the most suitable choices for such tasks. After applying ten different machine learning algorithms, the outcome of the proposed work is assessed with a few traditional methods depicted in Figure 4. The graphical comparison analysis ensures that the Gradient Boosting Machine (GBM) gains the highest accuracy of 97.1% comparative to the other traditional method.

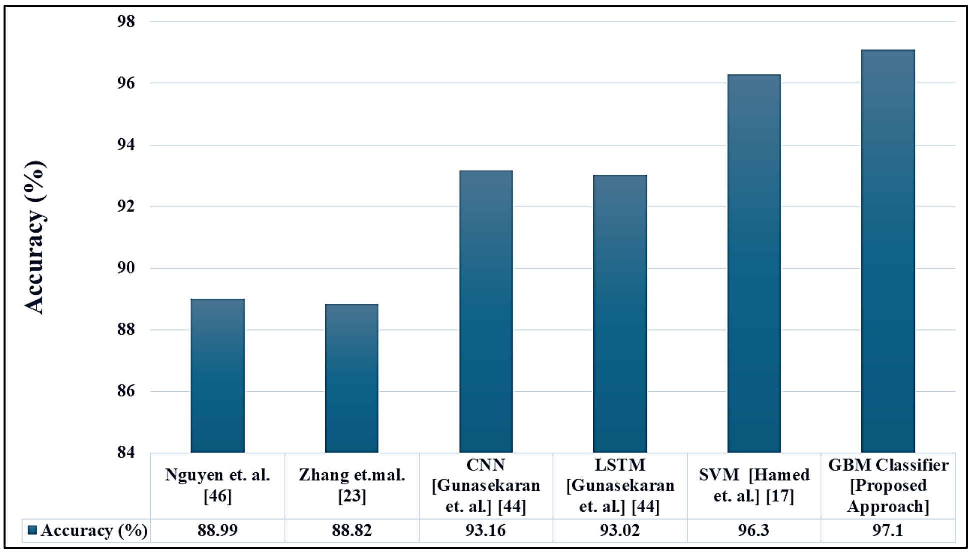

Figure 4: Comparison Analysis of Current Work with Previous Work (17, 43–45)

To justify the results of the intended method, we compared it with Clustal Omega, a well-known alignment-based method for sequence comparison. The same dataset was used to construct the phylogenetic tree with Clustal Omega method, as shown in Figure 5 (46). It is observed that the original species were clustered in the same group, while the sequence of dummy species is separated into different clusters. Therefore, the results of our method are consistent with those obtained from the Clustal Omega method (46).

The clustering pattern observed in the phylogenetic tree reflects known taxonomic relationships among the mammalian species, indicating that the ML-classified sequences correspond to functionally conserved genomic regions. This provides indirect biological validation, supporting the robustness of our classification approach.

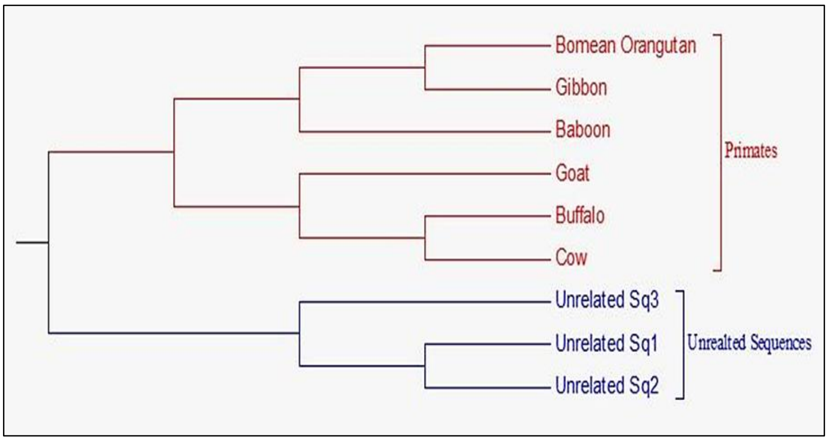

Figure 5: Phylogenetic Tree for FASTA DNA Sequence of the Original Species and Unrelated Species

Conclusion

The study for the classification of FASTA DNA sequence represents a significant advancement with potential practical applications in the medical sector, various discoveries of drugs, and the diagnosis of disease. This study outlines several key steps: constructing a dataset from the FASTA DNA files, thereafter altering it into CSV format, importing and pre-processing the collected data, altering the inputs in text format into arithmetical data, and then keeping fit through diverse classification methods. Evaluation using various ML algorithms reveals that the Gradient Boosting Machine (GBM) accomplishes the highest 0.971 accuracy and 0.975 F1 score among the tested methods, highlighting its superior performance in FASTA DNA sequence classification. Moreover, the XGBoost achieves a decent result of 0.935 accuracy and 0.93 F1 score. Furthermore, the results of the proposed method are also validated using Clustal Omega.

Assessing performance on diverse FASTA DNA sequences, including those from various animals, would assist in identifying the boundaries and opportunities for refinement. This comprehensive evaluation could guide enhancements in feature extraction methods and algorithmic design tailored to specific biological contexts. Overall, while the current study underscores the effectiveness of the SVM linear classifier in DNA sequence classification, ongoing research in these directions promises to unlock further advancements in accuracy, efficiency, and application versatility. Future extensions of this work may involve functional enrichment analyses such as Gene Ontology (GO) or pathway analysis on gene-annotated subsets of the classified sequences, which could further improve the translational relevance and biological interpretability of the results.

Abbreviations

ANN: Artificial Neural Network, DNA: Deoxyribonucleic Acid, DT: Decision Tree, FASTA: Fast-All (a text-based format for DNA and protein sequences), GBM: Gradient Boosting Machine, KNN: K-Nearest Neighbours, LR: Logistic Regression, NB: Naive Bayes, NCBI: National Centre for Biotechnology Information, RF: Random Forest, SVM: Support Vector Machine.

Acknowledgment

We would like to express our heartfelt gratitude to Narula Institute of Technology, Kolkata, India, and JIS College of Engineering, Kalyani, India, for their invaluable support and the enriching research environment that significantly contributed to the successful completion of this work.

Author Contributions

Papri Ghosh: Conceptualization, Overall Supervision, Debrupa Pal: Conceptualization, Subhram Das: Methodology, Overall Supervision, Soumyabrata Saha: Data Curation and Validation, Suparna Dasgupta: Data Curation and Validation.

Conflict of Interest

The authors of this work state that they have no conflicts of interest about its publication.

Ethics Approval

Not applicable.

Funding

This study was not funded by any academic institution.

References

1. Das S, Palit S, Mahalanabish AR, Choudhury NR. A new way to find similarity/dissimilarity of DNA sequences on the basis of dinucleotides representation. In: Computational Advancement in Communication Circuits and Systems: Proceedings of ICCACCS 2014. Springer. 2015 Jan 1:151-160. https://doi.org/10.1007/978-81-322-2274-3_19
2. Kennedy SR, Prost S, Overcast I, Rominger AJ, Gillespie RG Krehenwinkel H. High-throughput sequencing for community analysis: the promise of DNA barcoding to uncover diversity, relatedness, abundances and interactions in spider communities. Dev Genes Evol. 2020;230(2):185–201.
3. Das S, Deb T, Dey N, Ashour AS, Bhattacharya D, Tibarewala D. Optimal choice of k-mer in composition vector method for genome sequence comparison. Genomics. 2018;110(5):263–73.
4. Pal D, Dey S, Ghosh P, Das S, Maji B. A new method for protein sequence comparison using chaos game representation. In: International Conference on Data Analytics & Management. Springer. 2023 Dec 29:389–397. https://doi.org/10.1007/978-981-99-6553-3_3
5. Das S, Das A, Mondal B, Dey N, Bhattacharya DK, Tibarewala DN. Genome sequence comparison under a new form of tri-nucleotide representation based on bio-chemical properties of nucleotides. Gene. 2020;730:144257.
6. Ranwez V, Chantret NN. Strengths and limits of multiple sequence alignment and filtering methods. In: Scornavacca C; Delsuc F; Galtier N, editors. Phylogenetics in the Genomic Era. 2020;p.2.2:1–2.2:36.
7. Das S, Das A, Bhattacharya D, Tibarewala D. A new graph-theoretic approach to determine the similarity of genome sequences based on nucleotide triplets. Genomics. 2020;112(6):4701–4714.
8. Zhen H, Gong W, Wang L, Ming F, Liao Z. Two-stage data-driven evolutionary optimization for high-dimensional expensive problems. IEEE Trans Cybern. 2021;53(4):2368–2379.
9. Dey S, Das S, Bhattacharya D. Biochemical property based positional matrix: A new approach towards genome sequence comparison. Journal of Molecular Evolution. 2023; 91(1):93-131.
10. Dey S, Ghosh P, Das S. Positional difference and frequency (PdF) based alignment-free technique for genome sequence comparison. Journal of Biomolecular Structure and Dynamics. 2023;42(23):12660-12688.
11. Goshisht MK. Machine learning and deep learning in synthetic biology: key architectures, applications, and challenges. ACS Omega. 2024;9(9):9921–9945.
12. Berrar D, & Dubitzky W. Deep learning in bioinformatics and biomedicine. Briefings in Bioinformatics. 2021;22(2):1513-1514.
13. Zhao F, Li L, Chen Y, Huang Y, Keerthisinghe T P, Chow A, Dong T, Jia S, Xing S, Warth B, Huan T, Fang M. Risk-based chemical ranking and generating a prioritized human exposome database. Environmental Health Perspectives. 2021;129(4): 047014(1-12).
14. Suyama Y, Hirota SK, Matsuo A, Tsunamoto Y, Mitsuyuki C, Shimura A, Okano K. Complementary combination of multiplex high-throughput DNA sequencing for molecular phylogeny. Ecological Research. 2022;37(1):171-181.
15. Zhong HS, Dong MJ, Gao F. G4Bank: a database of experimentally identified DNA G-quadruplex sequences. Interdiscip Sci Comput Life Sci. 2023;15(3):515–23.
16. Touati R, Messaoudi I, Oueslati A, Lachiri Z, Kharrat M. New intraclass helitrons classification using DNA-image sequences and machine learning approaches. Innovation and Research in BioMedical Engineering. 2021;42(3):154–64.
17. Hamed BA, Ibrahim OAS, Abd El-Hafeez T. Optimizing classification efficiency with machine learning techniques for pattern matching. Journal of Big Data. 2023;10(1):124.
18. Ryu C, Lecroq T, Park K. Fast string matching for DNA sequences. Theoretical Computer Science. 2020;812:137–48.
19. Xu G, Li H, Ren H, Lin X, Shen X. DNA similarity search with access control over encrypted cloud data. IEEE Transactions on Cloud Computing. 2020;10(2): 1233–52.
20. Ravikumar M, Prashanth M. Analysis of DNA sequence pattern matching: a brief survey. In: Cybernetics, Cognition and Machine Learning Applications: Proceedings of ICCCMLA. 2021 Mar 31:221–229. https://doi.org/10.1007/978-981-33-6691-6_25
21. Millán Arias P, Alipour F, Hill KA, Kari L. DeLUCS: Deep learning for unsupervised clustering of DNA sequences. PLoS One. 2022;7(1):e0261531(1-25).
22. Rossi F, Paiardini A. A machine learning perspective on DNA and RNA G-quadruplexes. Curr Bioinform. 2022;17(4):305–9.
23. Zhang F, Chen F, Schwarzacher T, Heslop-Harrison J, Teng N. The nature and genomic landscape of repetitive DNA classes in Chrysanthemum nankingense shows recent genomic changes. Annals of botany. 2023;131(1):215–28.
24. Juneja S, Dhankhar A, Juneja A, Bali S. An approach to DNA sequence classification through machine learning: DNA sequencing, K Mer counting, thresholding, sequence analysis. International Journal of Reliable and Quality E-Healthcare. 2022;11(2):1–15.
25. Sarkar S, Mridha K, Ghosh A, Shaw RN. Machine learning in bioinformatics: new technique for DNA sequencing classification. In: Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2022. Springer. 2022 Aug 31:335-355. https://doi.org/10.1007/978-981-19-2980-9_27
26. Jukic S, Saracevic M, Subasi A, Kevric J. Comparison of ensemble machine learning methods for automated classification of focal and non-focal epileptic EEG signals. Mathematics. 2020;8(9): 1481(1-16).
27. Keith J M, Davey C M, Boyd S E. A Bayesian method for comparing and combining binary classifiers in the absence of a gold standard. BMC Bioinformatics. 2012;13:1-11.
28. Ravikumar M, Prashanth M, Guru D. Matching pattern in DNA sequences using machine learning approach based on K-mer function. In: Gunjan VK, Zurada JM, editors. Modern Approaches in Machine

Learning & Cognitive Science: A Walkthrough, Springer. 2022;p.159-171.

29. Koul N, Manvi SS, Gardiner B. Method for classification of cancers with partial least squares regression as feature selector with kernel SVM. In: 2022 International Conference for Advancement in Technology (ICONAT). IEEE. 2022 Jan 21:1-7. https://doi.org/10.1109/ICONAT53423.2022.9725968
30. Dragomir MP, Calina TG, Perez E, et al. DNA methylation-based classifier differentiates intrahepatic pancreatobiliary tumours. EBioMedicine. 2023;93(104657):1-17.
31. Chadha A, Dara R, Pearl DL, Sharif S, Poljak Z. Predictive analysis for pathogenicity classification of H5Nx avian influenza strains using machine learning techniques. Preventive Veterinary Medicine. 2023;216:105924.
32. Lu H, Karimireddy SP, Ponomareva N, Mirrokni V. Accelerating gradient boosting machines. In: International Conference on Artificial Intelligence and Statistics, PMLR. 2020;108:516-526.
33. Touzani S, Granderson J, Fernandes S. Gradient boosting machine for modeling the energy consumption of commercial buildings. Energy Build. 2018;158:1533-1543.
34. Obiora CN, Ali A, Hasan AN. Implementing extreme gradient boosting (xgboost) algorithm in predicting solar irradiance. In: PES/IAS PowerAfrica. IEEE. 2021 Aug 23:1-5. https://doi.org/10.1109/PowerAfrica52236.2021.9543159
35. Osman AIA, Ahmed AN, Chow MF, Huang YF, El-Shafie A. Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Engineering Journal. 2021;12(2):1545-1556.
36. Mandal DK, Biswas N, Manna NK, Gayen DK, Benim AC. An application of artificial neural network (ANN) for comparative performance assessment of solar chimney (SC) plant for green energy production. Scientific Reports. 2024;14(1):979.
37. Ray S, Haque M, Ahmed T, Nahin TT. Comparison of artificial neural network (ANN) and response surface methodology (RSM) in predicting the compressive and splitting tensile strength of concrete prepared with glass waste and tin (Sn) can fiber. Journal of king saud university-engineering sciences. 2023;35(3):185–99.
38. Osman A, Ibrahim A, Alsadoon A, et al. Breaking new ground in cardiovascular heart disease diagnosis K-RFC: An integrated learning approach with K-means clustering and Random Forest classifier. AIMS Mathematics. 2024;9(4):8262-8291.
39. Kouadio K, Liu J, Liu R, et al. K-Means Featurizer: A booster for intricate datasets. Earth Science Informatics. 2024;17:1203-1228.
40. Borjigin C. Data analysis with Python. In: Borjigin C, editors. Python Data Science, Springer. 2023;p.295–342.
41. Rajamani SK, Iyer RS. Machine learning-based mobile applications using Python and Scikit-Learn. In: Samanta D, editor. Designing and Developing Innovative Mobile Applications, IGI Global. 2023:282-306.
42. Lavanya A, Gaurav L, Sindhuja S, Seam H, Joydeep M, Uppalapati V, Ali W, SD VS. Assessing the performance of Python data visualization libraries: a review. Int. Journal of Computer Engineering in Research Trends. 2023;10(1):29–39.
43. Nguyen NG, Tran VA, Phan D, Lumbanraja FR, Faisal MR, Abapihi B, Kubo M, Satou K. DNA sequence classification by convolutional neural network. Journal of Biomedical Science and Engineering. 2016;9(5):280-286.
44. Gunasekaran H, Ramalakshmi K, Arokiaraj ARM, Kanmani SD, Venkatesan C, Dhas CSG. Analysis of DNA sequence classification using CNN and hybrid models. Computational and Mathematical Methods in Medicine. 2021;2021(1):1835056.
45. Nguyen NG, Tran VA, Ngo DL, Phan D, Lumbanraja FR, Faisal MR, Abapihi B, Kubo M, Satou K. DNA sequence classification by convolutional neural network. Journal of Biomedical Science and Engineering. 2016;9(05):280.
46. Soliman N, Abdelhaleem S, El-Shafai W. Hybrid approach for taxonomic classification based on deep learning. Intelligent Automation and Soft Computing. 2021;32:1881-91.