IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15671

Intelligent DNA Sequence Classification Using Machine Learning Techniques

Dr.B.Nageswara Rao[1], Nalanagula Chaitanya[2]

Assistant Professor, Department of CSE, Megha Institute of Engineering and Technology for Womens, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana – 501301[1]

M.Tech Student, Department of CSE, Megha Institute of Engineering and technology for Womens, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana – 501301[2]

Abstract: The purpose of this study is to provide a new method for improving the efficiency and precision of DNA sequence classification by use of Machine Learning algorithms. To sort DNA sequences into predetermined groups, such species identification, the suggested DNA Sequencing Classifier makes use of cutting-edge Machine Learning methods. Sequencing DNA has changed the face of genetics in many fields, including medicine, evolution, and others. Still, DNA sequences may be difficult to accurately classify. Machine learning may automate this process, making it more precise and uncovering hidden patterns. Using state-of-the-art Machine Learning methods, this research seeks to develop a DNA sequence classifier that is both efficient and effective. Genetic research may be accelerated and improved with the use of automated categorisation. The first step in ensuring the accuracy of raw DNA sequences is data preparation. Machine learning models use the extracted characteristics as inputs and choose the most effective classifier according to performance. Using k-mer counting, the DNA Sequencing Classifier is compared to other approaches and reviewed thoroughly. The integration of machine learning with DNA sequencing has great potential for simplified DNA categorisation, which in turn might speed up research and enhance our knowledge of genetics.

Keywords: DNA, Natural Language Processing (NLP), k-mer counting, Naïve Bayes, Bag of words.

INTRODUCTION

One of the most cutting-edge tools that our research has developed, the "DNA Sequencing Classifier," makes use of state-of-the-art Machine Learning algorithms. We want to use this classifier to sort DNA sequences into predetermined groups, such as species identification, which will lead to exciting new opportunities in genetic analysis and study. This project's main goal is to create a classification model that can anticipate genus functions using just DNA sequencing of the coding sequence. Our goal is to simplify genus function prediction and provide useful insights by using the information already present in these sequences. There are several critical procedures that must be followed in order for this enormous undertaking to be successful. First, we do data pre-processing, which involves cleaning raw DNA sequences to get rid of noise, fix mistakes, and get them into an analysis-ready format. This first step lays the groundwork for trustworthy outcomes. Our DNA Sequencing Classifier will be fine-tuned in the following stages: feature extraction, model selection, and performance assessment. To guarantee the reliability and robustness of our classification system, we extract pertinent characteristics, choose the best Machine Learning models, and thoroughly evaluate their performance.

LITERATURE SURVEY

Riccardo Rizzo, Massimo La Rosa, Antonino Fiannaca, and Alfonso Urso took part. [1]: "A Deep Learning Approach to DNA Sequence Classification" . Using a spectral representation, this research presents a deep learning neural network for DNA sequence classification. The model was put through its paces on 16S genes and contrasted with several classifiers, such as support vector machines, naïve Bayes, random forest, and general regression neural networks. In their work, Samia M. Abd -Alhalem, El-Sayed M. El-Rabaie, Naglaa. F. Soliman, Salah Eldin S. E. Abdulrahman, Nabil A. Ismail, and Fathi E. Abd El-samie were cited as follows: "DNA Sequences Classification with Deep Learning: A Survey" . Using deep learning neural networks, this research delves into the topic of DNA sequence categorisation, paving the way for scalable and precise classification. Deep learning approaches utilised in this context are surveyed. Get Your Paper Ready Before You Style It. Ghanshyam I. Prajapati and Pooja Dixit: "Machine Learning in Bioinformatics: A Novel Approach for DNA Sequencing"In this study, we explore how bioinformatics researchers might use Machine Learning methods including artificial neural networks, genetic algorithms, and fuzzy systems to classify gene sequences. "Machine Learning for Classifying Tuberculosis Drug-Resistance from DNA Sequencing Data" [4] Li Yang, Katherine E. Niehaus, Timothy M. Walker, Zamin Iqbal, A. SarahWalker, and Daniel J. Wilson. In this work, we utilise DNA sequencing data to classify Mycobacterium tuberculosis (MTB) resistance to different tuberculosis medicines. We then

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15671

employ Machine Learning models to enhance this classification process. The authors of the paper "Classifying Cancer Patients Based on DNA Sequences Using Machine Learning" are Hussain Fahad, Saeed Umair, Muhammad Ghulam, Islam Noman Sheikh, and Ghazala Shafi. Using Machine Learning to analyse cancer patients' DNA sequences, this article [12] aims to improve cancer detection. A wide variety of classifiers are evaluated, such as SVMs, decision trees, and neural networks. *"Comparison of Monkey pox and wart DNA sequences with Deep Learning Model" [6] by Talha Burak Alakus and Muhammet Baykara. This paper demonstrates the potential of deep learning in viral sequence analysis by using an algorithm to analyse the DNA sequences of HPV (warts) and MPV (monkey pox).

EXISTING SYSTEM

Overfitting the model, missing data management, and the need for domain knowledge in environmental science have all been problems in previous research. The interpretability of complex models and non-stationary data have also been sources of worry. In the past, DNA Sequence classifiers have encountered several important challenges, such as: data quality and availability; feature engineering; temporal and spatial variability; overfitting and generalisation; real-time prediction; data integration; handling extreme events; updating according to environmental changes and climate variability; and regulatory considerations. Collaboration among data scientists, researchers, and healthcare providers is frequently necessary to tackle these difficulties. Enabling researchers and medical professionals with a tool that maximises the value of genetic data involves developing strong and reliable DNA Sequence classifier models, which in turn demands careful consideration of these problems throughout the research process.

PROPOSED SOLUTION

Important parts of the suggested approach to improving DNA sequence categorisation include k-mer counting, Naïve Bayes, Machine Learning (ML), and Natural Language Processing (NLP) methods like Bag of Words. To give you a quick rundown, k-mer counting is a powerful method in bioinformatics for studying DNA sequences. Inside a larger, more generic DNA sequence, you may find substrings of length k, which are called k-mers. The following three-mers are produced when we use the "ATCGATCAC" string as our DNA sequence: ATC, TCG, CGA, GAT, ATC, TCA, CAC. To count the k-mers, we need to determine how many times each kmer appears. Sequence motifs and repeating features may be more easily identified in this way. With k-mer counting, we can quantify DNA sequences for our DNA sequence classifier, making it easier for Machine Learning algorithms to process genetic information. Data Preprocessing: Ensuring the quality of input DNA sequences is the first step in the approach. To prepare the data for analysis, it is necessary to apply noise reduction, error correction, and format standardisation. Bow of Words - An NLP-Inspired Product: Using the Bag of Words (BoW) approach, which is derived from natural language processing, DNA sequences are represented as structured data. This method generates a collection of features based on counts by dividing the sequences into "words" of a predetermined length. BoW converts DNA sequences to a Machine Learning-friendly format. As a representation for DNA sequences, the Bag of Words approach is taken from the field of natural language processing. In this context, a k-mer is like a "word" in the "language" of DNA. After compiling a comprehensive list of all k-mers present in the dataset, the BoW method depicts each DNA sequence as a frequency vector of these k-mers.

For instance, for the following DNA sequences:

ATCGATCAC

GATCACATC

And with 3-mers, our dictionary would be:
[ATC, TCG, CGA, GAT, TCA, CAC, ACA]

Naïve Bayes Classifier:

A probabilistic classification approach called Naïve Bayes is used as the algorithm for classification. Genus function prediction is one application that makes use of the BoW representation to sort DNA sequences into established groups. Naïve Bayes is well-known for its simplicity and efficacy in text categorisation challenges. The k-mer features, which are used by the Naïve Bayes classifier, are assumed to be conditionally independent of the class, based on the "naïve" hypothesis and Bayes theorem. Here is one possible way to write about DNA sequence classification:

$$P(class|sequence) \propto P(class) * \prod P(k-mer|class) \quad (1)$$

The prior probability of a class is denoted as P(class), while the probability of seeing a particular k-mer in a class is P(kmer|class). Take DNA sequences as an example. To classify them into two groups, A and B, we would calculate: Afterwards, the class with the highest probability is given the sequence. A more accurate and robust model is created by

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15671

combining many decision trees in the Random Forest ensemble learning approach. This method is known as the Random Forest Classifier. Using bootstrap sampling, it generates a number of decision trees using fractions of the training data. When deciding how to divide a tree at each node, it uses a completely arbitrary subset of characteristics. The forest's trees all make their own predictions. If we're doing classification, we can use majority voting; if we're doing regression, we can average all the individual tree forecasts to get the final prediction. As a whole, Random Forest is more accurate and less prone to overfitting than individual decision trees. It is less affected by outliers and can process data with several dimensions. A decision tree classifier is a supervised learning technique that may be used for both regression and classification. Using the data's characteristics, it builds a decision-tree model. A root node represents the whole dataset and serves as the starting point. It uses metrics like information gain and Gini impurity to choose the optimal feature at each node for data splitting. It generates child nodes for every conceivable value of the selected characteristic. Afterwards, this process is repeated until a stopping requirement (such as a maximum depth or minimum samples per leaf) is met. Predictions are represented by the last tree leaves. You can use decision trees with numerical and categorical data; they're straightforward, easy to grasp, and versatile.

ARCHITECTURE DIAGRAM

Fig.1: The DNA Sequence classifier system architecture using MO. Usually requires a number of interconnected parts and data sources to provide up-to-the-minute results.

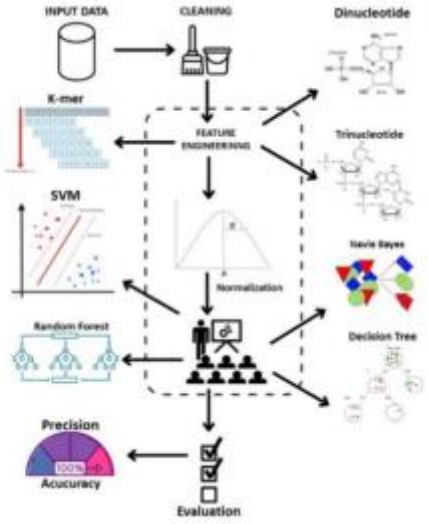

Fig.1. Architecture Diagram

Parts of the design comprise: Step 1: Collect Dataset from Kaggle. In order to prepare the data, we must first examine it for any variables that may not be connected to our dependent variable. Get rid of any blanks in the dataset. A lowercase version of each DNA sequence was generated. Data Segmentation: Create two sets of data, one for training and one for testing, to help in building and improving the model. Examples of models used for model selection include Naïve Bayes and several approaches from Natural Language Processing such as kmer counting and count vectorizer. • Support Vector Machine • Monte Carlo Simulation • Bayesian To assess the efficacy of a decision tree model, one must compute its f1_score, Accuracy, Precision, and Recall.

RESULTS AND DISCUSSION

The "DNA Sequencing Classifier," a state-of-the-art tool that takes use of the most recent developments in Machine Learning techniques, is introduced in this study. We want to use this classifier to sort DNA sequences into predetermined groups, such as species identification, which will lead to exciting new opportunities in genetic analysis and study. Using Machine Learning to classify different species based on their DNA sequences is the objective of this project. Numerous species' DNA is represented in the collection by lengthy sequences of nucleotide bases (A, T, C, G). By converting the DNA sequence data to a machine learning-friendly format and using classification algorithms, we aim to determine the species type. As seen in Figure 2. Here, we show how several classes (species) of kmer words are distributed graphically. The x-axis shows the different taxonomic groups, while the y-axis shows the number of occurrences of distinct words (k-mers) inside each group. This graph is useful for comprehending the variety and complexity of k-mers in different species, which might provide light on patterns in DNA sequences that are peculiar to each species.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15671

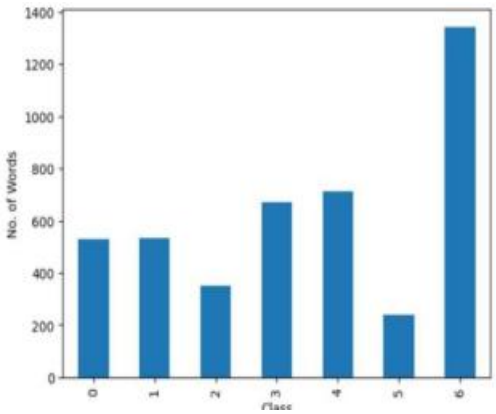

Fig.2. Class Distribution of Human DNA Naïve Bayes Classifier:

The confusion matrix in Figure 3 represents the performance of the Naïve Bayes classifier on the species categorisation problem. The rows of the matrix show the expected classes, whereas the columns show the actual classes. Each item in the matrix shows the ratio of the classifier's predictions to the actual labels for that class. Analysis of Records: • TPs, or true positives, are cases when the classifier correctly identifies a species. When a classifier accurately determines that a given species does not belong to a given class, this is known as a true negative (TN). An FP occurs when a classifier incorrectly assigns a species to the wrong category. When a classifier incorrectly assigns a species to the wrong category, this is known as a false negative (FN).

Confusion matrix

Predicted	0	1	2	3	4	5	6
Actual							
0	119	0	0	0	1	0	1
1	0	127	0	0	0	0	4
2	1	0	92	0	0	0	0
3	0	0	0	149	0	0	1
4	1	0	0	0	178	0	5
5	1	0	0	0	2	62	0
6	5	1	0	1	0	0	344

accuracy = 0.978
precision = 0.978
recall = 0.978
f1 = 0.978

Fig.3. Confusion matrix for Naïve Bayesian classifier Random Forest Classifier:

One ensemble approach that may detect non-linear patterns in data is Random Forest. Despite its potential for great accuracy, it often requires greater processing resources and, in smaller datasets, may overfit on k-mer features. In comparison, the consistent accuracy is better provided by Naïve Bayes due to its simplicity and ability to generalise effectively across the dataset. Figure 4 shows that the Random Forest Classifier achieved an outstanding accuracy of 0.458 and an F1 score of 0.392 when applied to human DNA sequences. There are a lot of uses for this, and mistakes in either direction might have serious repercussions, so it's crucial. Random Forest is well-suited to high-dimensional biological data because of its ensemble method, which allows it to withstand overfitting significantly better than individual decision trees.

Confusion matrix

Predicted	0	1	2	3	4	5	6
Actual							
0	28	0	0	0	0	0	93
1	0	45	0	0	0	0	86
2	0	0	38	0	0	0	55
3	0	0	0	18	0	0	132
4	0	0	0	0	13	0	171
5	0	0	0	0	0	9	56
6	0	0	0	0	0	0	351

accuracy = 0.458
precision = 0.799
recall = 0.458
f1 = 0.392

Fig.4. Confusion matrix for Random Forest Classifier

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15671

Though the Decision Tree classifier's accuracy metrics are shown in Figure 5. was not identified in the search results; it is well-known that Decision Trees, while generally effective, are susceptible to overfitting, particularly when dealing with complex datasets like DNA sequences. Decision Trees have shown good results in numerous studies, however ensemble approaches, such as Random Forest, tend to outperform them because to their susceptibility to data noise. To help visualise the judgements made for object classification, decision trees are helpful for analysing the decision routes.

Confusion matrix

Predicted	0	1	2	3	4	5	6
Actual							
0	39	0	0	0	0	0	82
1	1	48	0	3	2	0	77
2	0	0	34	1	0	0	58
3	1	1	0	16	0	0	132
4	1	0	0	2	12	0	169
5	1	0	0	0	0	11	53
6	2	1	0	0	0	0	348

accuracy = 0.464
precision = 0.720
recall = 0.464
f1 = 0.400

Fig.5. Confusion Matrix for Decision Tree Classifier

To maximise the potential of this DNA Sequence Classification, the next procedures are to extract features, pick a model, and assess its performance. To guarantee the reliability and robustness of our classification system, we extract pertinent characteristics, choose the best Machine Learning models, and thoroughly evaluate their performance. Fig.6 shows a bar graph that compares the precision, accuracy, and F1 score of three classifiers: Decision Tree, Random Forest, and Naïve Bayes. In order to get a full picture of how well each model does on the species categorisation job, we use metrics that assess many aspects of the models.

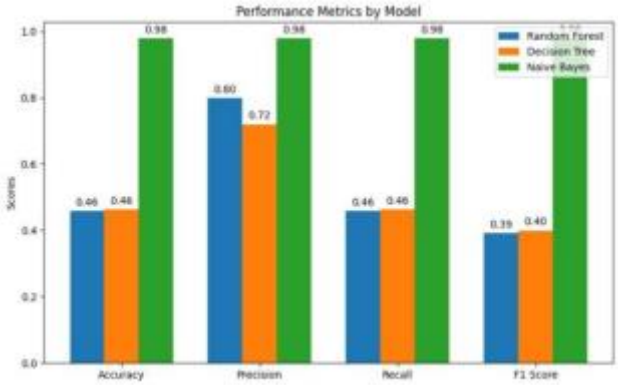

Fig.6. Performance Metrics by Model

CONCLUSION

As a first step towards a paradigm shift in the analysis and interpretation of genetic data, DNA sequence classification has great potential. In fact, we enhanced classification accuracy using Machine Learning breakthroughs while simultaneously laying the groundwork for future genomics discoveries. This new study highlights the exciting potential of combining computational approaches with biological research to uncover molecular pathways at the foundation of life.

FUTURE DEVELOPMENT

Several promising avenues for further study and development are now available thanks to the project: 1. Improved Feature Engineering: Going forward, studies may look at more sophisticated feature engineering methods to better depict DNA sequences. More intricate data patterns and characteristics may need to be considered for this. 2. Integrating Deep

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15671

Learning: Recurrent neural networks (RNNs) and convolutional neural networks (CNNs) are examples of deep learning technologies that have the potential to increase the accuracy of DNA sequence categorisation. 3. Enterprise-Level Genomic Data: Possible real-world uses in genetics, personalised medicine, and illness research could result from expanding the project to manage massive genomic data sets and real-time sequencing data. Fourth, work together across disciplines: By bringing together medical researchers, biologists, and geneticists, we can create solutions and tools that are tailored to the demands of genetic research. 5. Practical Use: Expanding the project's scope to include practical uses like clinical diagnostics, drug development, and evolutionary research would constitute a substantial stride in enhancing the accessibility and influence of genetic findings. 6. A User-Friendly Interface: Building a DNA Sequencing Classifier with an intuitive interface would make the automated classification system accessible to medical professionals and other non-technical users.

REFERENCES

[1]. Wang D, Chen X, Wu Y, Tang H and Deng P (2022) Artificial intelligence for assessing the severity of microtia via deep convolutional neural networks. Front. Surg. 9:929110. doi: 10.3389/fsurg.2022.929110

[2]. Sarkar, S., Mridha, K., Ghosh, A., Shaw, R.N. (2022). Machine Learning in Bioinformatics: New Technique for DNA Sequencing Classification. In: Shaw, R.N., Das, S., Piuri, V., Bianchini, M. (eds) Advanced Computing and Intelligent Technologies. Lecture Notes in Electrical Engineering, vol 914. Springer, Singapore. https://doi.org/10.1007/978-981- 19-2980-9_27

[3]. Annual Review of Genomics and Human Genetic Vol. 9:387-402 (Volume publication date 22 September 2008) First published online as a Review in Advance on June 24, 2008 https://doi.org/10.1146/annurev.genom.9.081307.164 359

[4]. Mohamed, O. (2021, April 5). DNA sequencing with Machine Learning. DataValley. Retrieved January 24, 2023, from https://datavalley.technology/dna-sequencing-with- machine-learning/

[5]. Doppala, Thrisha, "Differentiating Human Populations Based on kmer Classification of Hand Bacteria" (2018). Graduate Theses, Dissertations, and Problem Reports. 3720. https://researchrepository.wvu.edu/etd/3720

[6]. Bianchini, M., Shaw, R. N., Das, S., & Piuri, V. (2021). Advanced computing and intelligent technologies. Lecture Notes in Networks and Systems. https://doi.org/10.1007/978-981-16-2164-2

[7]. Mardis, E. R. (2008). Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics, 9(1), 387–402. https://doi.org/10.1146/annurev.genom.9.081307.164 359

[8]. Chauhan, N. S. (2021, January 15). DNA sequence dataset. Kaggle. Retrieved December 20, 2022,from https://www.kaggle.com/datasets/nageshsingh/dna-sequence-dataset

[9]. Gu, M. (2021, September 6). DNA sequence classification based on Milvus. Milvus. Retrieved December 21, 2022, from https://milvus.io/blog/dna-sequence-classification-based-onmilvus.md

[10]. Dixit, P., & Prajapati, G. I. (2015). Machine Learning in Bioinformatics: A novel approach for DNA sequencing. 2015 Fifth International Conference on Advanced Computing & Communication Technologies. https://doi.org/10.1109/acct.2015.73

[11]. K.A. Mohamed Junaid, T. Sethukarasi, M. Vigilson Prem, Adi Alhudhaif, Norah Alnaim, "A novel efficient Rank-Revealing QR matrix and Schur decomposition method for big data mining and clustering (RRQR-SDM)", Information Sciences, Volume 657, 2024, 119957, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2023.119957.

[12]. Visuwasam, L.M.M., A distributed intelligent mobile application for analyzing travel big data analytics. Peer-to-Peer Netw. Appl. 13, 2036–2052 (2020). https://doi.org/10.1007/s12083-019-00799-z

[13]. Sethukarasi T, Prakash M, Baburaj E (2023) An Efficient Hybrid Job Scheduling Optimization (EHJSO) approach to enhance resource search using Cuckoo and Grey Wolf Job Optimization for cloud environment. PLOS ONE 18(3): e0282600. https://doi.org/10.1371/journal.pone.0282600

[14]. A. M. Sermakani, "Effective Data Storage and Dynamic Data Auditing Scheme for Providing Distributed Services in Federated Cloud", Journal of Circuits, Systems and ComputersVol. 29, No. 16, 2050259 (2020), https://doi.org/10.1142/S021812662050259X

