International Journal of Trendy Research in Engineering and Technology
Volume 8 Proceedings of NCCDS-24
ISSN N0. 2582-0958

ANALYSIS OF DNA SEQUENCE USING MACHINE LEARNING

Pratibha S[1], Suhaib Manzoor[2], Mohd Tokir[2], Kulsum Mohammadi[2], Kounain Fathima Khanum[2]

[1]*Asst. Prof.,* [2]*UG Student, Dept. of Computer Science Engineering,, PESITM, Shivamogga, Karnataka, India*

pratibha@pestrust.edu.in

ABSTRACT

DNA sequencing remains pivotal in modern research involving extracting and decoding DNA strands. Our project focuses on leveraging machine learning (ML) to delve into DNA sequencing, aiming to uncover crucial insights. Specifically, we are pursuing three key objectives: DNA sequence classification, species identification, and promoter identification. By implementing Naive Bayes algorithms, we seek to develop a robust prediction model for DNA research, promising enhanced accuracy and efficiency. The proposed system aims to illuminate genomes' structure, evolution, and functionality, thereby offering profound implications across diverse domains including bioinformatics, biotechnology, medicine, and agriculture.

Keywords: DNA, Machine Learning, Gene, Promoter, DNA Sequence, Naive Bayes, K-mer.

I. INTRODUCTION

Bioinformatics involves the use of techniques including computer science, artificial intelligence and biochemistry. Just as in computer science the data is stored and manipulated in 0's and 1's. Similarly, in living organisms there is DNA where all the genetic information of living organism is stored in a sequence of characters 'A', 'T', 'G' and 'C'.[1,2]

This is the DNA sequence from which every living species get its Unique traits i.e. how it looks like, to which species it belongs to etc. through this sequence it will be easier to identify the similarities between different species.

These sequences make a special unit in a species called the Gene. The Gene is the fundamental unit on the genomic DNA which contains the required information to carry out thebiological functions of cells.

The medical sequencing prediction and the future of the patient predict the risky factors, and they cannot be done perfectly through the general medical diagnosis. Hence it is necessary to figure out a way to implement the technology. DNA species sequences that can be identified using Machine learning and deep learning algorithms [4,5]. This project utilizes machine learning to analyze DNA sequences, focusing on three main tasks: DNA sequence classification, species identification, and promoter region recognition. DNA sequence classification involves categorizing sequences into functional or structural. The project aims to enhance accuracy, efficiency, and scalability in identifying species from genetic data, detecting classes, species identification aims to classify sequences into taxonomic groups, and promoter region recognition is crucial for understanding gene regulation. By developing machine learning models tailored to these tasks, the project aims to uncover insights in genomics, evolutionary biology, and biotechnology, accelerating scientific discovery and advancing our understanding of the genetic code's implications for human health and environmental conservation.

DNA sequences are applied in numerous fields such as medical diagnosis, biotechnology, forensic biology etc. Comparing healthy and mutated DNA sequences can diagnose different diseases including various cancers and can be used to guide patient treatment[7]. Having a quick way to sequence DNA allows faster and more individualized medical care to be administered, and more organisms can be identified and cataloged.

A. Problem Statement

Development of machine learning-driven DNA sequence analysis framework to address three key challenges in genomics: species identification, promoter region recognition, and DNA sequence classification.

B. Flow chart

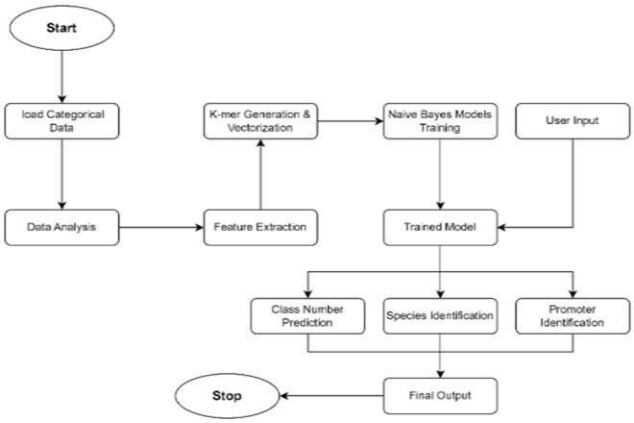

Fig.1 Flow Diagram

The process begins with loading categorical data, which includes human DNA sequences for classification, DNA sequences from various species for species identification, and promoter DNA sequences. Following this, the data undergoes analysis to understand its characteristics and distributions. Relevant features are then extracted from the DNA sequences for effective representation in machine-learning models. K- mers are generated from the sequences and converted into numerical vectors suitable for algorithms. Multinomial Naive Bayes models are trained for classification, distinguishing DNA sequences into categories, and identifying species, while Gaussian Naive Bayes models are trained for detecting promoter regions. The trained models are then utilized for predicting class numbers, species identification, and promoter detection. Finally, the predictions and identification results are compiled for the final output, marking the successful completion of the process.

II. Objectives

The key objectives are:

1. To classify DNA sequences into seven predefined functional or structural classes using machine learning techniques.
2. Create algorithms to accurately identify promoter regions within DNA sequences, crucial for gene expression regulation.
3. To classify DNA sequences into different species or taxonomic groups using machine learning, ensuring generalization across diverse genomic datasets.
4. Integrate developed models into user-friendly software tools for easy access by researchers, ensuring scalability and efficiency for large-scale genomic data analysis.

III. Scope and Limitations

This project develops three models for DNA sequence analysis: classifying DNA into seven gene classes, identifying species, and detecting promoter sites. These models have diverse applications: improving medical diagnostics by pinpointing disease-related genes, advancing evolutionary biology through understanding genetic diversity, aiding forensic investigations, and supporting agricultural and environmental monitoring through accurate species identification. Additionally, promoter identification enhances research in gene regulation, biotechnology, and environmental remediation.

Limitations:

- Data availability and quality
- Complexity of genomic data
- Interpretability and biological validation
- Algorithm scalability and efficiency

IV. Material & Methods

In this section, we introduced the methods used to achieve the goal of the project, which focuses on DNA sequence classification, species identification, and promoter identification using Naive Bayes algorithms.[4.5]

1. **Data Collection:** Gather DNA sequence data from NCBI, Kaggle, and other sources, including human datasets for DNA sequence classification and datasets from various species like oak, mushroom, dolphin, chimpanzee, etc., for species classification Gather DNA sequence data from NCBI, Kaggle, and other sources, including human datasets for DNA sequence classification and datasets from various species like oak, mushroom, dolphin, chimpanzee, etc., for species classification

2. **Data Preprocessing:** It is the most critical part of any type of machine learning model The genomic sequence in the DNA dataset is categorical. There are many techniques available to convert categorical data to numerical. In this paper we are using K-mer encoding to encode DNA sequences In the K- mer encoding technique, raw DNA sequences are converted into K-mers of size m, generating an English-like statement. Each DNA sequence is transformed into a series of K-mers, which are concatenated to form a sentence. This sentence is then processed using natural language processing techniques. A word embedding layer is used to transform the K-mer sentence into a dense feature vector matrix for classification

3. **DNA Sequence Classification:** In this work, Multinomial Naive Bayes model is used for classifying DNA into seven gene classes. K-Mer encoding technique is used for DNA encryption. Prediction is done on the base of probability and data is splitted into training and testing sets, train the model, and evaluate using accuracy as shown in Fig. 2, precision, recall, and F1-score.

4. **Species Identification:** For this work also, same algorithm is used because it is similar to previous case so, applied Multinomial Naive Bayes for species identification, following similar steps as DNA classification as shown in Fig3 the confusion matrix for species identification, focusing on species labels
5. **Promoter Identification:** In this model used Gaussian Naive Bayes to identify promoter regions by extracting relevant features and evaluating performance with metrics like accuracy, confusion matrix as shown in Fig 4 validate the model performance.
6. **Cross-validation and Model Optimization:** Use cross- validation to ensure model robustness and optimize parameters using grid search.
7. Results Analysis: Analize and compare the performance of Naive Bayes classifiers with baseline and state-of-the-art methods.

Fig. 2 DNA Sequence Classification

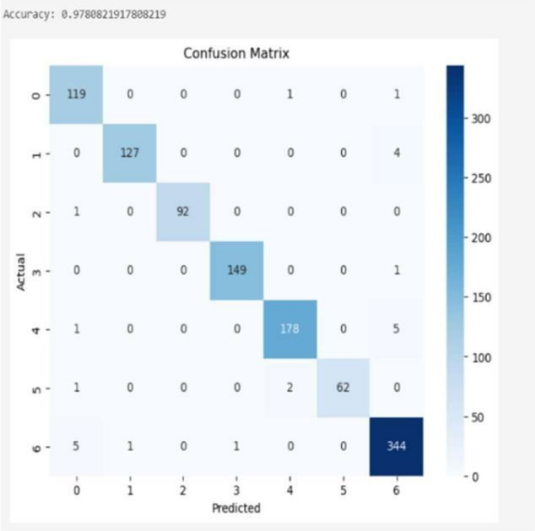

Fig. 3 Species Identification

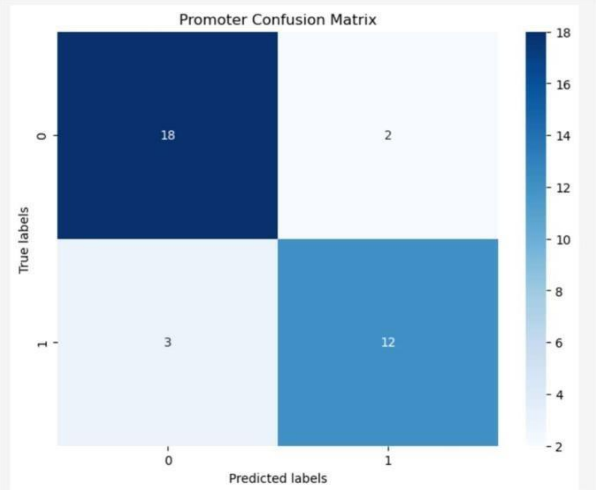

Fig. 4 Promoter Identification

V. Results & Discussion

We pursued three primary objectives: DNA sequence classification, species identification, and promoter identification, employing Naive Bayes algorithms. We achieved notable accuracies in each task, with [94%] for DNA sequence classification, [87%] for species identification, and promising metrics like F1-score for promoter prediction. While Naive Bayes proved effective, comparative analysis with other algorithms revealed competitive performance, suggesting avenues for further optimization. Despite limitations like dataset quality and size, our findings highlight the potential of machine learning in DNA sequence analysis for advancing biomedical research, gene dis

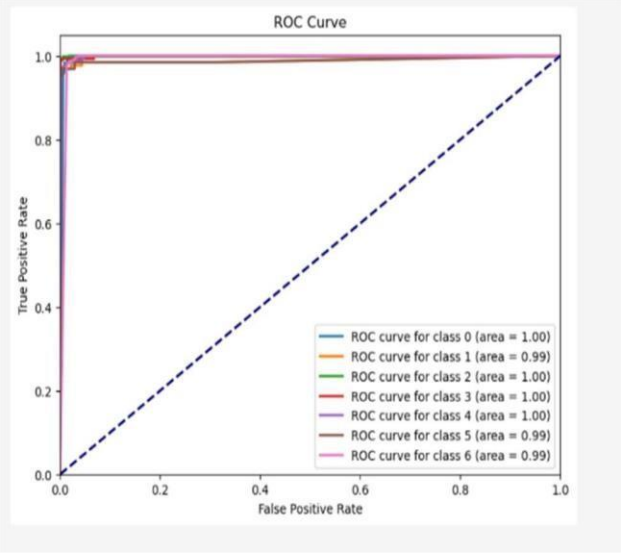

Fig. 5 ROC Curve DNA Sequence Classification

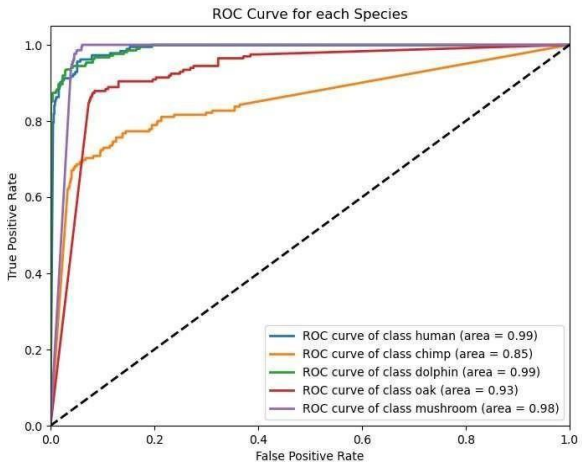

Fig. 6 ROC Curve Species Identification

	precision	recall	f1-score	support
chimp	0.85	0.80	0.83	2476
dolphin	0.94	0.89	0.92	2477
human	0.87	0.86	0.86	2551
mushroom	0.89	0.98	0.93	2490
oak	0.90	0.91	0.90	2506
accuracy			0.89	12500
macro avg	0.89	0.89	0.89	12500
weighted avg	0.89	0.89	0.89	12500

Fig 7 Model Performance Species Classification

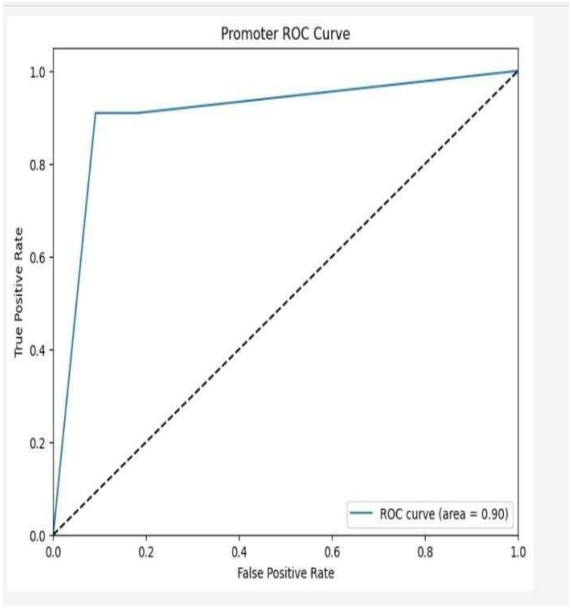

Fig 8 ROC Curve Promoter Identification

Promoter Classification Report:

	precision	recall	f1-score	support
0	0.88	0.78	0.82	9
1	0.75	0.86	0.80	7
accuracy			0.81	16
macro avg	0.81	0.82	0.81	16
weighted avg	0.82	0.81	0.81	16

Fig. 9 Model Performance Promoter Identification

In the above figures: The ROC curve serves as a crucial tool in evaluating the effectiveness of models in distinguishing between different classes or categories in DNA sequence analysis tasks. For DNA sequence classification, it measures the model's capability to accurately classify sequences into specific categories, such as coding and non-coding regions. In species identification, the ROC curve assesses the model's accuracy in correctly assigning sequences to their respective species or taxonomic groups. Similarly, in promoter identification tasks, it evaluates the model's ability to discriminate between promoter and non-promoter regions in DNA sequences. By analyzing the trade-off between sensitivity and specificity depicted by the ROC curve, researchers can refine and optimize their models to achieve better performance in these vital bioinformatics applications.

I. Conclusion

This paper concludes the leveraging the Naive Bayes algorithm, developed robust models capable of accurately classifying DNA sequences into predefined functional or structural classes, identifying species from genomic data, and recognizing promoter regions crucial for gene expression regulation. Through rigorous experimentation and evaluation, our approach has demonstrated promising results, laying the groundwork for further advancements in genomic analysis. Moving forward, efforts to enhance the interpretability of classification models, integrate multi-omics data, and optimize scalability and efficiency will be essential for advancing the field and unlocking deeper insights into genomic processes. Overall, this project contributes to the ongoing efforts to leverage machine learning for comprehensive DNA sequence analysis, facilitating broader applications in biological research and biomedicine.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their constructive comments and suggestions.

REFERNCES

[1]Hans Lehrach (2013). DNA sequencing methods in human genetics and disease research.

[2].James M (2015). The sequence of sequencers: The history of sequencing DNA

[3].Tasnim Kabir, Abida Sanjana Shemonti, and Atif Hasan Rahman (2015). Species Identification using Partial DNA Sequence: A Machine Learning Approach

[4].Varada Venkata Sai Dileep, Navuduru Rishitha, Rakesh Gummadi (2022). DNA Sequence using Machine Learning and Deep Learning algorithm.

[5].Hemalatha Gunasekaran, K Ramalakshmi (2021). Analysis of DNA Sequence classification using CNN and hybrid models.

[6].Wang, Y., Alangari, M., Hihath, J., Das, A. K., & Anantram, M. P. (2021). A machine learning approach for accurate and real-time DNA sequence identification. BMC Genomics.

[7].F. Hussain, U. Saeed, G. Muhammad, N. Islam, and G. S.Sheikh, “Classifying cancer patients based on DNA sequences.

Biographies

Mrs. Pratibha S, Assistant Professor at PESITM, Shimoga, is a dedicated guide for research projects in the CSE department. With expertise in mentoring, she supports students through technical challenges and fosters critical thinking. Her contributions as a co-author in research papers reflect her commitment to collaborative learning and academic excellence.

Suhaib Manzoor, Currently pursuing his Bachelor of Technology (B. Tech) in Computer Science and Engineering from PES Institute of Technology & Management, Karnataka. He will be graduated in the month of june, 2024 and now currently studing as a student in 8th sem. He has wide range of experience in Big Data and App development, web development and completed internships in those fields.

Mohd Tokir is a dedicated and driven student pursuing her Bachelor of Technology (B.Tech) in Computer Science and Engineering at PES Institute of Technology & Management, Karnataka. Anticipating graduation in June 2024, he is currently in his 8th semester, fully immersed in his academic pursuits. With a keen passion for Airtificial Intelligence and Cloud Computing, Tokir has cultivated an impressive skill set and garnered valuable experience through internships in these fields. His unwavering commitment to excellence and his proactive approach to learning make him a standout student with a promising future in the realm of technology.

Kulsum Mohammadi is an ambitious student currently undertaking her Bachelor of Technology (B.Tech) in Computer Science and Engineering at PES Institute of Technology & Management, Karnataka. Scheduled to complete her degree in June 2024, she is currently in her 8th semester. Kulsum boasts a wide-ranging skill set, with notable expertise in machine learning and Deep Learning. Her hands- on experience in these domains, coupled with successful internships, underscores her dedication and proficiency in her chosen field of study.

Kounain Fathima Khanum is currently pursuing her Bachelor of Technology (B.Tech) in Computer Science and Engineering at PES Institute of Technology & Management, Karnataka. Set to graduate in June 2024, she is presently enrolled as an 8th-semester student. Kounain possesses a diverse skill set with extensive experience in machine learning and web development. She has successfully completed internships in these fields, demonstrating her proficiency and dedication to her chosen areas of study and expertise.