IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 3, March 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203119 |

Gene Classification Based on DNA Sequence Using Deep Learning Techniques

Gotouri Bhuvaneshwari[1], Pranavi Ambati[2], Sunkoju Yogender[3],

T.Veda Reddy[4]

Students, Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India[1, 2, 3]

Assistant Professor, Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India[4]

ABSTRACT: In the field of genetics, a genome is a complete collection of DNA in an organism. The genomic data is used in the field of Bioinformatics to learn the functions of a new protein by which the researchers make the necessary drug design. Genomic data processing requires a significant amount of data storage and high-performance hardware and software for statistical analysis. This project focuses on encoding the biological DNA sequence data using Label encoding and Natural Language Processing techniques (such as k-mer counting) and based on the encoded DNA sequence, gene sequence classification is done using various Machine Learning algorithms such as Logistic Regression, Random Forest, Naive Bayes, Decision Tree, Support Vector Machine, KNN and Deep Learning algorithms which includes CNN & LSTM. Our ultimate aim is to build various classification models that is trained on the human DNA sequences data which will perform well even on unseen data. We will be testing our model's performance by passing the DNA sequences of other species (like Dogs, Chimpanzees) as input. An overall comparison of accuracy performance will be done against various models and datasets.

KEYWORDS: DNA sequence, Gene family, Machine Learning, deep learning, CNN, LSTM, NLP, K-Mer counting, One Hot Encoding, and Ordinal Encoding.

I. INTRODUCTION

The polymer known as deoxyribonucleic acid (DNA), which carries genetic instructions, is made up of two polynucleotide chains that coil around one another to form a double helix structure. At present deep learning has become the method of preference for many genomics modelling tasks including the DNA sequence classification. Understanding organisms in life science requires a thorough understanding of DNA. Most of the genetic instructions for growth, operation, and reproduction in all organisms are carried by DNA. We can now quickly read a DNA sequence thanks to advances in sequencing technology. Statistics from the NHGRI Genome Sequencing Program show that the cost of reading one million base pairs decreased dramatically, from more than $5,000 in September 2001 to only $0.014 in October 2015. The amount of knowledge we have about DNA sequences is also expanding at an exponential rate. DNA serves as the cell's architectural manual. It contains all the data and instructions necessary for the growth and operation of living organisms. However, DNA cannot do it on its own. Many DNA-binding proteins assist in modulating DNA Functions. Every individual is special. Your genetic makeup contributes to your individuality. The instructions found inside each of your cells are called genes. Your appearance and bodily functions are under their control. Everyone has a unique set of instructions since each person has slightly unique genes. You are distinct in part because of your genes! Some diseases are brought on by changes to a gene's instructions. We refer to this as a mutation. Everybody has a variety of mutations. These modifications occasionally have no impact or even marginally benefit. But on sometimes, they can spread illness. Most diseases are brought on by a confluence of environmental factors, lifestyle decisions, and genetic alterations. Even those who share the same genes may or may not get a disease depending on their lifestyle choices and environment. A unique alteration in the DNA of a single gene is the root cause of thousands of diseases. These illnesses are often uncommon. These disorders typically appear when a person is born with a mutant gene. The classifying the gene sequence is much more important in order to know the changes in the gene instruction which causes the disease. If we know to which gene family the specific DNA Sequence belongs to we can group them and discover the specific drug to cure the disease.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203119 |

II. RELATED WORK

Various studies in bioinformatics employ a range of methodologies for DNA sequence analysis. These include multinomial classification for gene class determination, Support Vector Machine (SVM) classification for identifying functional sites, and utilization of diverse machine learning algorithms like decision trees and neural networks on datasets of varying sizes. Deep learning approaches, such as convolutional neural networks (CNN) and recurrent neural networks (RNN), are also prevalent, with some studies focusing on optimizing hyperparameters for improved performance. Challenges such as dataset imbalance, computational efficiency, and resource-intensive training processes are noted across these works, highlighting the ongoing complexity and diversity of approaches in DNA sequence analysis within bioinformatics.

III. EXISTING METHOD

In the existing systems, Biopython library is used to handle the biological DNA sequence data in the datasets. In the data pre-processing step, the DNA sequences are encoded using techniques like Ordinal Encoding and Label Encoding. The existing systems uses the traditional machine learning algorithms like Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, whose accuracies were ranging from 70% to 85%.

Drawbacks of Existing Systems

The drawbacks of the current system include the following:

• Couldn't handle imbalanced classes in the dataset.

• Implemented only conventional Machine Learning Algorithms and couldn't promise high accuracy and performance.

• Didn't implement methods to handle model overfitting/underfitting issues.

• Gradient boosting models didn't perform well on the dataset.

IV. PROPOSED METHOD

In the Proposed work, in order to handle large biological DNA sequence data, we are using Bio python library and in the data pre-processing step, the DNA sequences are encoded using techniques like K-Mer Counting, Ordinal Encoding and Label Encoding. To perform gene classification, we are using deep learning algorithms such as CNN & LSTM rather than conventional Machine Learning Algorithms to promise higher accuracy and performance. Proper measures are taken to handle the imbalance in the data and for testing and evaluating the performance of classification models, dog and chimpanzee dataset are used. Finally statistical analysis on various gene families will be performed.

Advantages of Proposed System

• Probability of high accuracy

• Huge amount of data can be analyzed.

• No man power is required for detection.

• Lesser possibility of false result

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203119 |

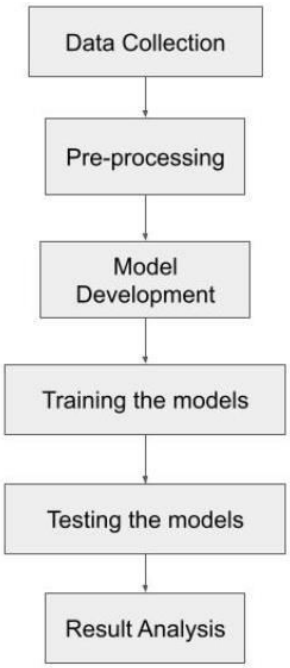

Fig 1: Flow Chart

V. SIMULATION RESULTS

The simulation results for the gene classification based on DNA sequences using machine learning and deep learning models show promising outcomes. Across various models such as KNN, decision trees, random forests, naïve bayes, logistic regression, support vector machines, CNN, and LSTM, accuracy metrics consistently demonstrate robust performance in accurately classifying gene families. The evaluation metrics including accuracy, loss, and confusion matrix reflect the effectiveness of the models in capturing the underlying patterns and relationships within the DNA sequences. Furthermore, testing the trained models on separate datasets comprising human, chimpanzee, and dog sequences showcases their generalization capability, indicating the potential applicability of the approach across diverse species. These results suggest that leveraging machine learning and deep learning techniques for gene classification using DNA sequences holds promise for advancing our understanding of genetic mechanisms underlying diseases and facilitating personalized medicine and drug discovery efforts.

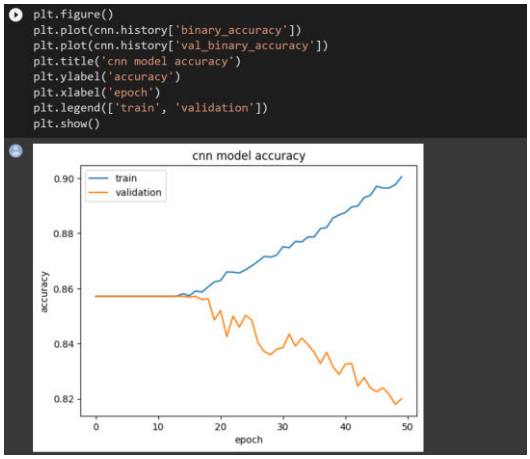

Fig 2: CNN Accuracy Graph

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203119 |

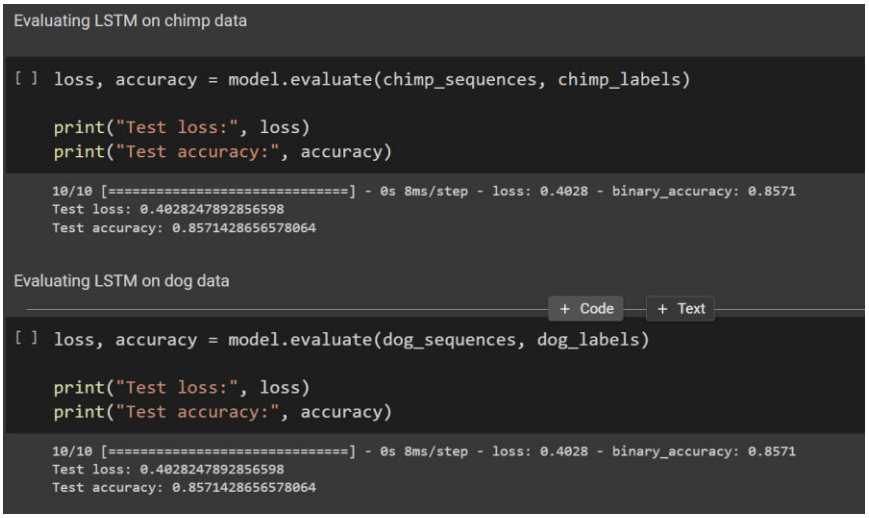

Fig 3: LSTM model evaluation on chimp & dog data

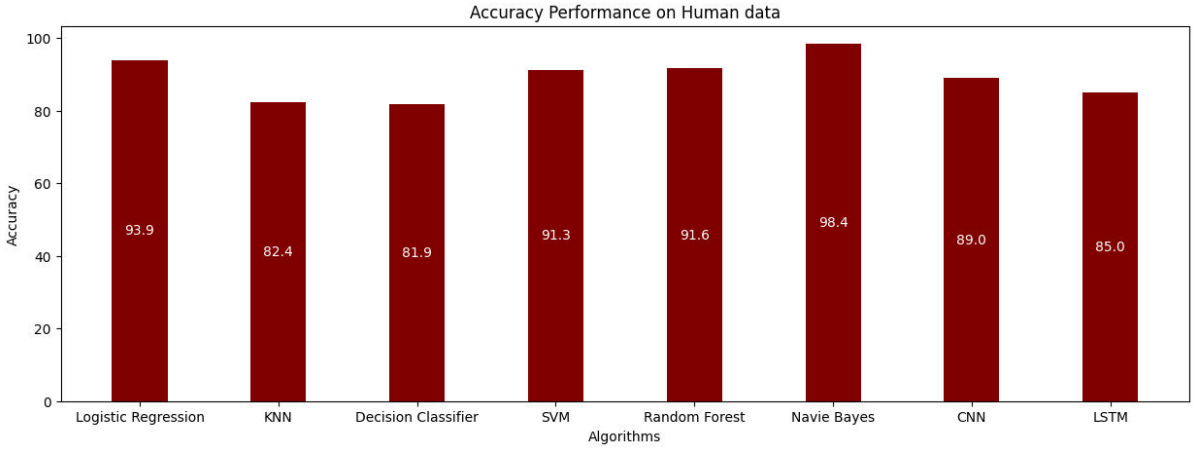

Fig 4: Performance analysis on Human data

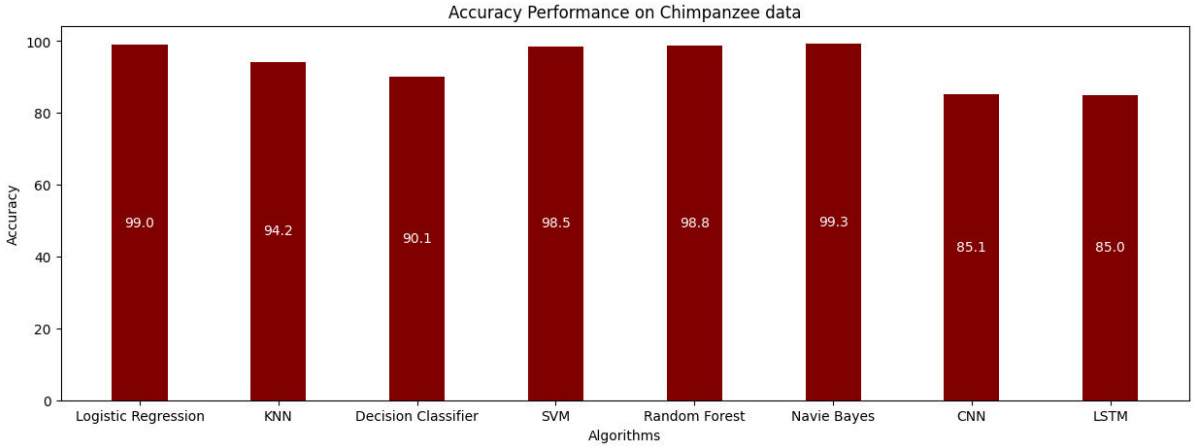

Fig 5: Performance analysis on Chimpanzee data

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203119 |

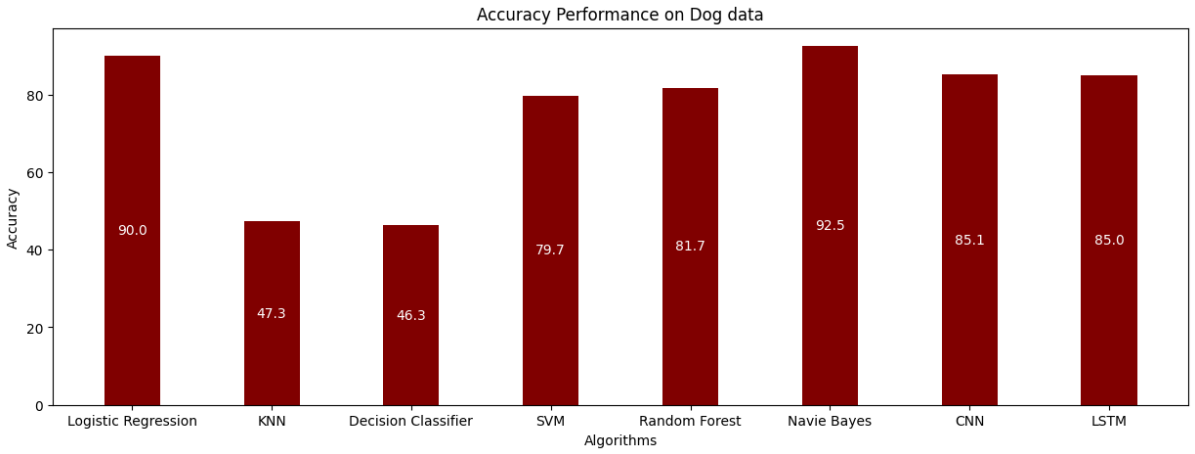

Fig 6: Performance analysis on Dog data

VI. CONCLUSION AND FUTURE WORK

In this project, DNA sequence classification using various ML & DL algorithms like Decision tree, Random Forest, KNN, Logistic Regression, Naive Bayes, SVM, CNN and LSTM are performed. All the DNA sequence strings are encoded using K-mer counting, Label and One hot encoding data preprocessing techniques and NLP bag of words algorithm by using count vectorizer was implemented for DNA sequence strings processing. With this almost every classification model has performed well, giving high accuracies ranging from 85% to 98% i.e. the Deep learning algorithms CNN & LSTM have performed well giving an accuracy around 85% which tells us that deep learning can also be preferred in the field of genomics whereas machine learning algorithm Naive bayes with k-mer encoding performs better than every other algorithm with an accuracy of 98%.

In future, we would like to enhance the application by adding following features:

• Inputs can be of type images where DNA sequences can be directly extracted.

• The presence of mutated diseases can also be predicted.

• This classification can be extended for Gene prediction.

• Can be modified and further be used for research and analysis.

• Early diagnosis of diseases by extracted patterns.

• Personalized development of medications based on genomics.

REFERENCES

[1] Juneja, Sapna, et al. "An Approach to DNA Sequence Classification Through Machine Learning: DNA Sequencing, K Mer Counting, Thresholding, Sequence Analysis." IJRQEH vol.11, no.2 2022: pp.1-15. http://doi.org/10.4018/IJRQEH.299963

[2] Perez-Rodriguez J, de Haro-Garcia A, Garcia-Pedrajas N. Floating Search Methodology for Combining Classification Models for Site Recognition in DNA Sequences. IEEE/ACM Trans Comput Biol Bioinform. 2021 Nov-Dec;18(6):2471- 2482. doi: 10.1109/TCBB.2020.2974221. Epub 2021 Dec 8. PMID: 32078558.

[3] Hussain, Fahad, et al. "Classifying cancer patients based on DNA sequences using machine learning." Journal of Medical Imaging and Health Informatics 9.3 (2019): 436-443.

[4] Rizzo, Riccardo, et al. "A deep learning approach to dna sequence classification." International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics. Springer, Cham, 2015.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 3, March 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1203119 |

[5] Mock, Florian, et al. "Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks." Proceedings of the National Academy of Sciences 119.35 (2022): e2122636119

[6] Mangkunegara, Iis Setiawan, and Purwono Purwono. "Analysis of DNA Sequence Classification Using SVM Model with Hyperparameter Tuning Grid Search CV." 2022 IEEE International.

[7] U. M. Akkaya and H. Kalkan, "Classification of DNA Sequences with k-mers Based Vector Representations," 2021 Innovations in Intelligent Systems and Applications Conference (ASYU), 2021, pp. 1-5, doi: 10.1109/ASYU52992.2021.9599084.

[8] Tampuu, Ardi, et al. "ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples." PloS one 14.9 (2019): e0222271.

[9] Bartoszewicz JM, Seidel A, Rentzsch R, Renard BY. DeePaC: predicting pathogenic potential of novel DNA with reverse-complement neural networks. Bioinformatics. 2020 Jan 1;36(1):81-89. doi: 10.1093/bioinformatics/btz541. PMID: 31298694.

[10] Gunasekaran, Hemalatha, et al. "Analysis of DNA sequence classification using CNN and hybrid models." Computational and Mathematical Methods in Medicine 2021 (2021).

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details