IJIRCCE

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

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005306 |

Comparison Analysis of Various Machine Learning Classifiers in Classifying DNA Sequence

Penugonda Vedavalli[1], Mandava Monica[2], Noorbasha Sharmila[3], Namburu Prasanthi[4], Satish Kumar Parasa[5]

UG Student, Dept. of I.T., Vasireddy Venkatadri Institute of Technology, Guntur, India[1,2,3,4]

Asst Professor, Dept. of I.T., Vasireddy Venkatadri Institute of Technology, Guntur, India[5]

ABSTRACT: The biological macromolecule deoxyribonucleic acid (DNA) is a deoxyribonucleic acid (DNA). Its primary role is to store data. It contains all of life's genetic information. DNA sequencing is the process of determining the exact nucleotide order of a DNA molecule. In a strand of DNA, it is utilised to establish the order of our bases Adenine(A), Guanine(G), Cytosine(C), and Thymine(T). Researchers looking into the function of genes need to know the DNA sequence. The DNA sequence contains the necessary information for a cell to produce protein and RNA molecules. Machine learning can be used as an alternative to manual analysis for DNA sequence data analysis in order to save time and boost data processing capabilities. This study is an empirical examination of machine learning classifiers' performance in classifying DNA.

KEYWORDS: - Key word1: Big Data, Key word2: Machine Learning, Key word3: SVM, Key word4: Navies Bayes, Key word5: DNA sequence clustering

I. INTRODUCTION

The process of establishing the precise order of nucleotides in a DNA molecule is known as DNA sequencing. The true chemical depiction of DNA is the "Double-helix." It's a unique situation .four types of nitrogen bases make up a nucleotide. Adenine (A) and Guanine (G) are two different types of adenine (G), A strand of DNA contains Cytosine (C) and Thymine (T). DNA sequencing is utilised to figure out what's going on in a person's body. Individual genes, complete chromosomes, or entire genomes of an organism are sequenced. A genome is an organism's whole collection of DNA. Living creature may have the same genome, yet their sizes are vastly different. DNA sequencing has also become one of the most popular methods of research. RNA or protein sequences can now be sequenced in a more efficient manner. DNA is a type of bio macromolecule that is found in all living thing, organisms. It holds life's genetic information and directs its evolution.

II. AIM AND SCOPE

The aim of this research is to tackle the problem of DNA sequence assembly by combining a machine learning approach with commonly used assembly techniques. The research purpose is to compare the performance of these machine learning classifiers using a thorough analysis from an information theoretic perspective. The research will determine if the “divide and conquer” approach, by grouping reads, will reduce computational complexity and improve the performance of DNA assembly techniques.
Performance analysis of various machine learning classifiers in classifying the DNA Sequences. The machine learning classifiers are “Knn, Random Forest, Decision Tree, Gaussian process algorithm, Mlpc, Gaussian Naïve bayes, Svmkernels”. The performance analysis is measured by “Accuracy, Precision, Recall, F1”

III.PROPOSED SYSTEM

Instead of human classification, we use machine learning classifiers in the suggested system.
By employing these classifiers, we can improve:
Many sequencing reactions take place at the same time in a highly parallel environment.
Fast: Because reactions are carried out in parallel, findings are available significantly quickly.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005306 |

Low-cost: genome sequencing is less expensive than Sanger sequencing.
Machine learning classifiers can sequence vast amounts of DNA considerably more quickly and cheaply than manual categorization thanks to parallelization.

IV. METHODOLOGY

Importing dataset:

When running python programs, we need to use data sets for data analysis.

Python has various modules which help us in importing the external data in various file formats to a python program. It imports the necessary libraries and import the dataset from the UCI repository as a Pandas Data Frame. The csv module enables us to read each of the rows in the file using a comma as a delimiter. We first open the file in read only mode and then assign the delimiter. Finally use a for loop to read each row from the csv file. The data is not in a usable form; as a result, we will need to process it before using it to train our algorithms.

Data pre-processing:

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Data is often incomplete, inconsistent, or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a method for resolving such issues .There are four main important steps for the preprocessing of data:

- Splitting of data set in training and validation set
- Taking care of missing values
- Taking care of categorical features
- Normalization of dataset

Apply machine learning classifiers:

Machine learning classifiers are used to classify and to predict the performance analysis of DNA sequences. Classification is an important mining task in machine learning. Its purpose is to learn a classification model from the training sample set to predict the category of unknown new samples the machine learning classifier used here are listed below:

- K-Nearest Neighbour Classifier
- Decision Tree Classifier
- Random Forest Classifier
- Gaussian process Classifier
- Multi-layer Perception Classifier
- Gaussian Naïve Bayes Classifier
- Adaboost Classifier
- SVM Linear
- SVM RBF

The above all classifiers are used for DNA sequencing and analyzed and best one is suggested.

This is the system architecture used in the project**.**

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005306 |

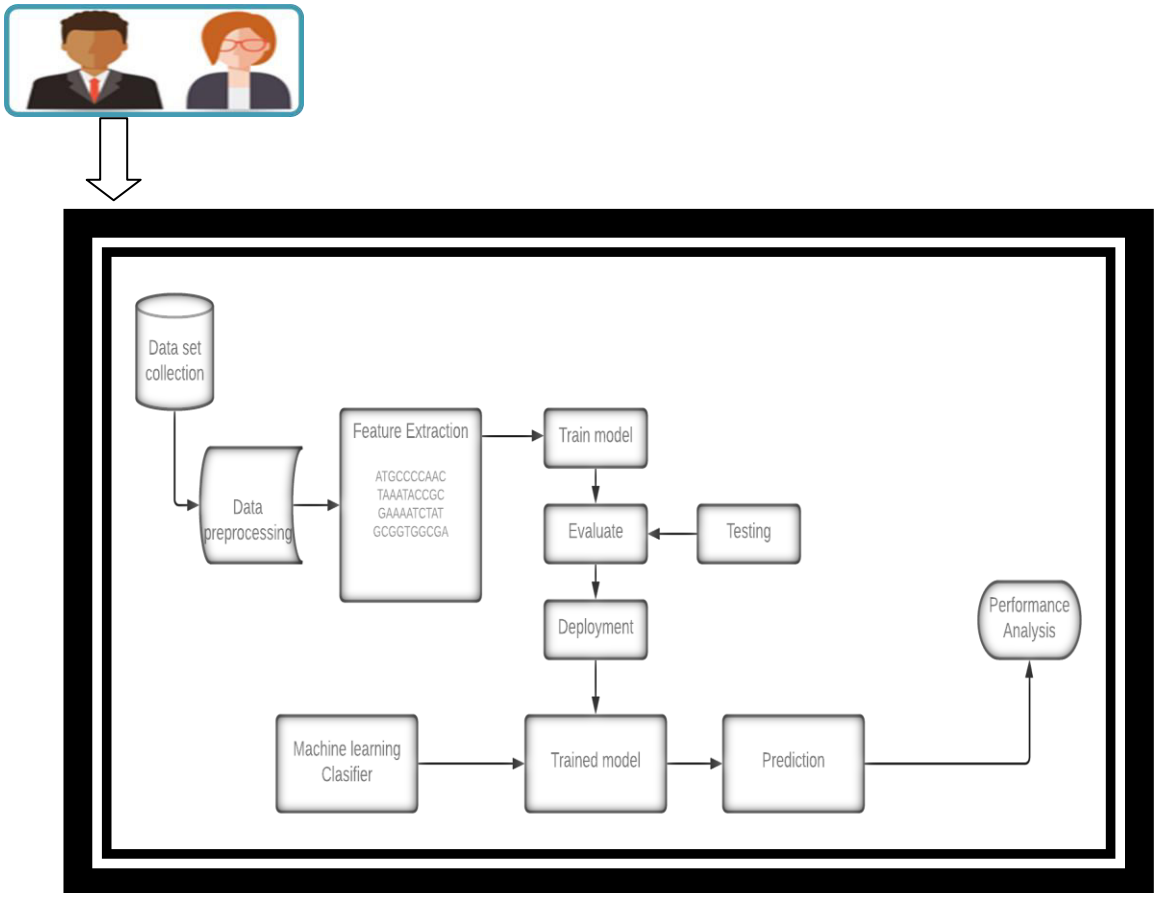

Fig. 1: System Architecture

V. RESULT AND DESCIPTION

The dataset is been pre-processed and various machine learning classifiers are applied to the obtained data after pre-processing.

The performance of various machine learning classifiers is analyzed using performance metrics (accuracy, recall, precision, f1) and is plotted below**.**

• Accuracy Score: proportion of correct predictions out of the whole dataset. It is the ratio of number of correct predictions to the total number of input samples.

• Precision Score: proportion of correct predictions out of all predicted diabetic cases. Precision is one indicator of a machine learning model's performance – the quality of a positive prediction made by the model.

• Recall Score: proportion of correct predictions out of all actual diabetic cases. Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made.

In the F1 score, we compute the **average of precision and recall**. They are both rates, which makes it a logical choice to use the harmonic mean.

Accuracy= TP+TN/TP+TN+FP+FN

Precision = TP/TP+FP

Recall = TP/TP+FN

F-Score = 2 * (precision * recall/ precision + recall)

TP = True Positives, *TN* = True Negatives, *FP* = False Positives, and *FN* = False Negatives.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 5, May 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1005306 |

S.no	Model	Accuracy	Precision	Recall	F1score
1	Gaussian Naive Bayes	0.914	0.916	0.914	0.914
2	SVM Rbf	0.886	0.906	0.886	0.884
3	SVM Sigmoid	0.886	0.906	0.886	0.884
4	MLPC	0.857	0.867	0.857	0.856
5	Gaussian Process	0.829	0.871	0.829	0.822
6	AdaBoost	0.829	0.846	0.829	0.826
7	SVM Linear	0.829	0.846	0.829	0.826
8	Decision Tree	0.800	0.856	0.800	0.790
9	Nearest Neighbour	0.771	0.842	0.771	0.757
10	Random Forest	0.514	0.511	0.514	0.502

VI. CONCLUSION

In general, there is not a very sharp difference in precision, recall and F1 scores among the classifiers. Gaussian Naive Bayes is still a clear winner here by having a F1-score of 0.914 for class, respectively. Besides, its high recall rate, 0.914, has proven itself as a reliable classifier to make DNA sequencing on the Dataset.

VII. FUTURE ENHANCEMENTS

Unlike next-generation sequencing technologies like Illumines, nanopore sequencing can analyze long sections of DNA, with no need to fragment the DNA before sequencing, and almost no sample preparation is required. Because no DNA is synthesized in the sequencing process, there's no need to add nucleotides, enzymes or other chemical reagents. As a result, nanopore sequencing dramatically reduces the cost of sequencing and removes the need for a highly equipped laboratory

REFERNCES

Bilofsky, H. S., Burks, C., Fickett, J. W., Goad, W. B., Lewitter, F. I., Rindone, W. P., et al. (1986). The GenBank genetic sequence databank. Nucleic Acids Res. 14, 1–4. doi: 10.1093/nar/14.1.1
PubMed Abstract | CrossRef Full Text | Google Scholar
Bosco, G. L., and Di Gangi, M. A. (2016). "Deep learning architectures for DNA sequence classification," in Proceedings of the International Workshop on Fuzzy Logic and Applications (Cham: Springer), 162–171. doi: 10.1007/978-3-319-52962-2_14
CrossRef Full Text | Google Scholar
Chen, L., and Liu, W. (2011). "An algorithm for mining frequent patterns in biological sequence," in Proceedings of the 2011 IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCABS) (Piscataway, NJ: IEEE), 63–68. doi: 10.1109/ICCABS.2011.5729943
CrossRef Full Text | Google Scholar
Choong, A. C. H., and Lee, N. K. (2017). "Evaluation of convolutionary neural networks modeling of DNA sequences using ordinal versus one-hot encoding method," in Proceedings of the 2017 International Conference on Computer and Drone Applications (IConDA) (Piscataway, NJ: IEEE), 60–65. doi: 10.1109/ICONDA.2017.8270400

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

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