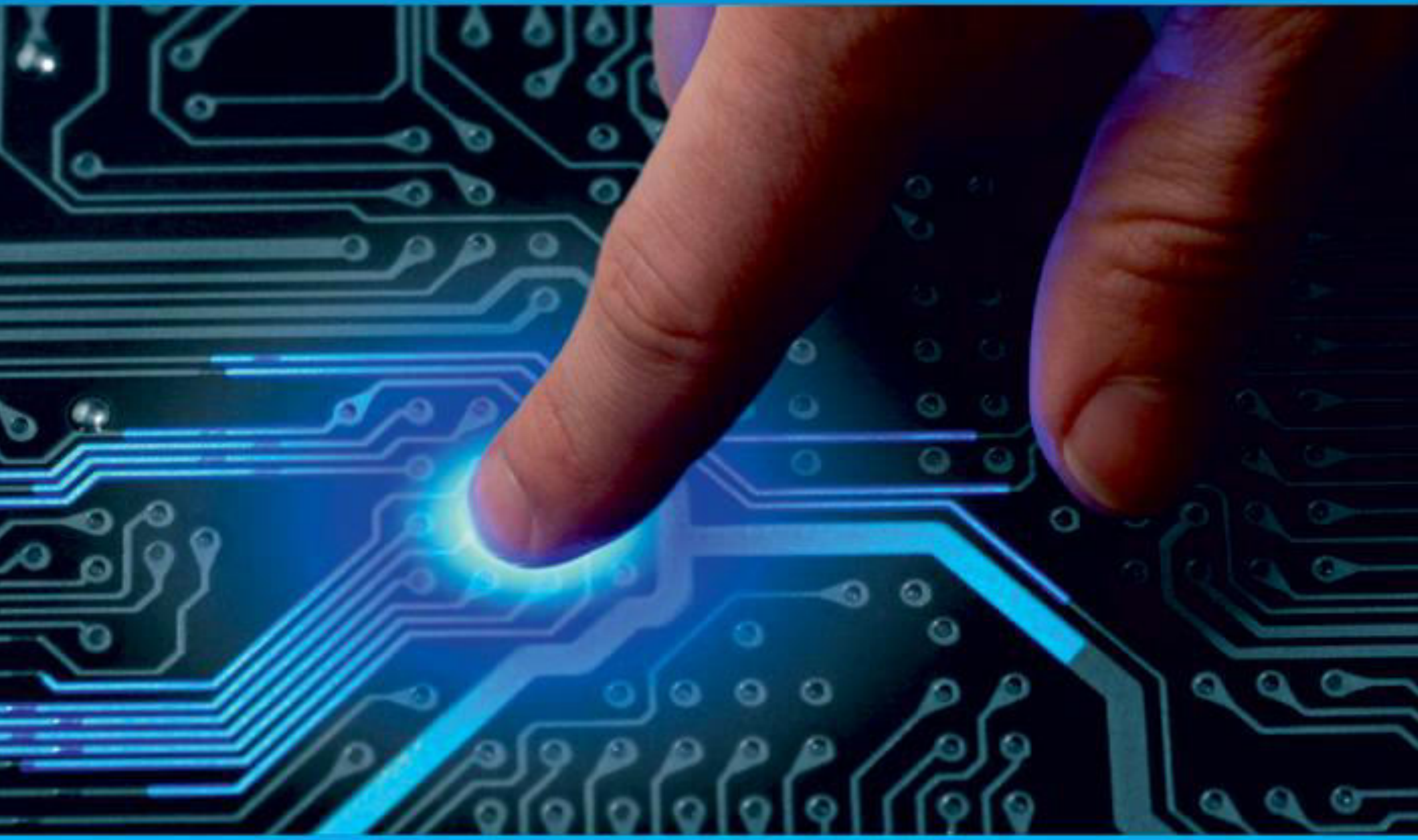
IJIRCCE

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



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 7, July 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 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207077|

DNA Sequence Classification based on Virus

Raghuveera N S, Smt. Sireesha G

Student, Department of MCA, Vishveshwara Technical University, The National Institute of Engineering, Mysuru, India

Assistant Professor, Department of MCA, Vishveshwara Technical University, The National Institute of Engineering, Mysuru, India

ABSTRACT: DNA sequences are the fundamental blueprints of life, holding the genetic instructions used in the growth, development, functioning, and reproduction of all known organisms and many viruses. The classification of DNA sequences is a critical task in bioinformatics, with applications ranging from medical diagnosis to evolutionary biology. This project aims to classify DNA sequences into various viral categories using advanced machine learning algorithms. By leveraging data from Kaggle, we employ techniques such as TF-IDF Vectorization, Counter Vectorization, and sequence padding to prepare the data for model training. The models used in this study include K-Nearest Neighbors (KNN), Decision Tree Classifier (DCT), Random Forest Classifier (RFT), Multi-Layer Perceptron (MLP), and Long Short-Term Memory (LSTM). The Random Forest Classifier achieved the highest accuracy of 99.67%. This document details the entire process, from problem identification to the implementation of the proposed solution, highlighting the methodologies, tools, and techniques employed.

I.INTRODUCTION

In the period of quickly progressing innovation, bioinformatics has developed as a essential field that coordinating biological information with computational strategies to reveal experiences that were already unattainable. Among the different errands in bioinformatics, DNA arrangement classification stands out due to its noteworthiness in understanding hereditary cosmetics, diagnosing infections, and creating medications. This extend centers on classifying DNA groupings into six viral categories: SARS-COV-1, MERS, SARS-COV-2, Ebola, Dengue, and Flu.

The classification of DNA groupings includes analyzing the groupings and categorizing them based on their characteristics. This handle is pivotal for distinguishing pathogens, understanding their behavior, and devising procedures to combat them. Conventional strategies of DNA classification are frequently time consuming and require noteworthy skill. Be that as it may, with the approach of machine learning, it is presently conceivable to automate this handle, making it speedier and more exact.

This extend utilizes datasets from Kaggle, a well-known stage for information science competitions and datasets. The datasets contain DNA arrangements from diverse infections, which are utilized to prepare and test distinct machine learning models. The models utilized in this consider incorporate K-Nearest Neighbors (KNN), Decision Trees Classifier (DCT), Random Forest Classifier (RF)), Multi-Layer Perceptron (MLP), and Long Short-Term Memory (LSTM). The Decision Tree Classifier(DTC) risen as the best show, accomplishing an exactness of 99.67%.

II.OBJECTIVES

1. To classify DNA arrangements into unmistakable viral categories utilizing machine learning calculations.
2. To preprocess the DNA groupings utilizing methods such as TF-IDF Vectorization, Counter Vectorization, and arrangement cushioning.
3. To assess the execution of distinct machine learning models in classifying DNA arrangements.
4. To recognize the most precise, demonstrate for DNA grouping classification.
5. To give a comprehensive investigation of the strategies, devices, and procedures utilized in the extend.

III. LITERATURE SURVEY

The report examines the body of knowledge regarding resume parser ,relevant studies on the following and analysed.

1. An Approach to DNA Sequence Classification -2020
2. DNA Sequencing using Machine Learning and Deep -2022
3. A review of machine learning for DNA sequence -2022

4. A Deep Learning Approach for Viral DNA Sequence -2021
5. Machine Learning in Bioinformatics: A Novel -2015
6. Analysis of DNA Sequence Classification Using Machine Learning -2022
7. Systematic Literature Review: Virus Prediction -2022
8. Classification of DNA Sequences with k-mers -2021
9. A Digital DNA Sequencing Engine for Ransomware 2021

IV. METHODOLOGY

1. Data Acquisition and Preprocessing: Extract DNA sequence data from credible sources like biological databases or institutions. Clean up the data to ensure consistency of the data; remove irrelevant data, treat missing values, and standardize format if necessary.

2. Feature Extraction: Extract relevant features from the DNA sequences so that they can be applied in classification. Common techniques for the features extraction are methods for k-mer counting or their frequency calculations, or those sequence alignment methods targeting the identification of conserved regions or motifs.

3. Feature Selection: Apply the relevant features to decrease dimensionality and increase effectiveness of classification models. The techniques for feature selection may include correlation analysis, PCA, or RFE.

4. Choice of Model: A selection of ML algorithms is relevant to better fitting the nature of the problem and characteristics of a dataset in the case of DNA sequence classification. This involves decision trees, random forests, SVM(support vector machines), K-nearest neighbors, and deep learning models such as CNN(Convolutional Neural Networks) or Recurrent Neural Networks.

5. Model Training: The selected ML models will be trained on the labeled datasets. The dataset will further be divided into a training set and a validation set that will be used to evaluate model performance during training. The hyper-parameters will have to be tuned for better accuracy in classification by optimizing model parameters.

6. Model Evaluation: Each model trained will be evaluated against relevant evaluation metrics to know their accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve. Cross-validation will be performed to check if the models are robust and have enough generalization capacity.

V. TOOLS AND TECHNOLOGIES REQUIRED

The report covers hardware and software requirements for the development of resume parser:

1.The programming language python

2.Libraries

- Sk-learn
- Numpy
- TensorFlow
- Pandas
- Keras

3.Additional libraries based on the selected ML algorithms

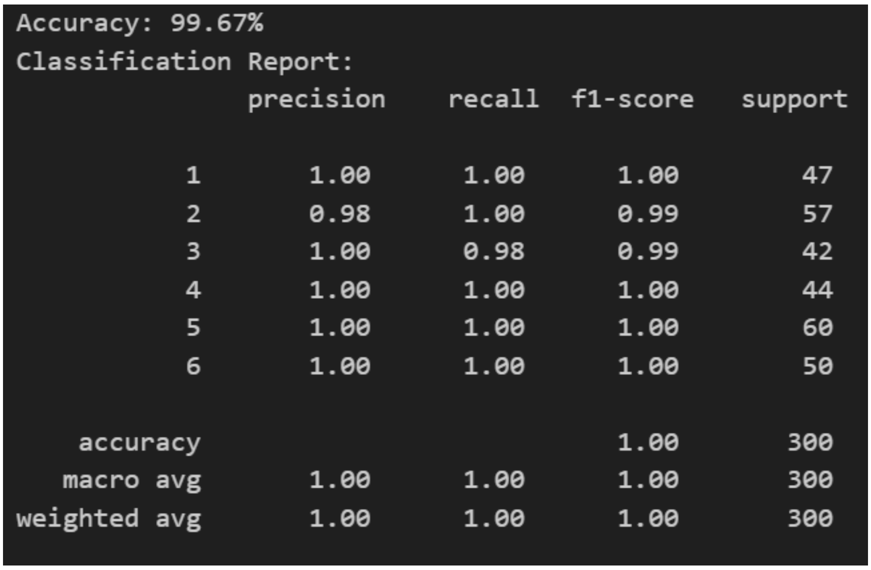
HARDWARE

1.processor(up to 2.5 GHz)

2.Graphics card (4GB + recommended)

3. Memory (8GB+)

VI.RESULT

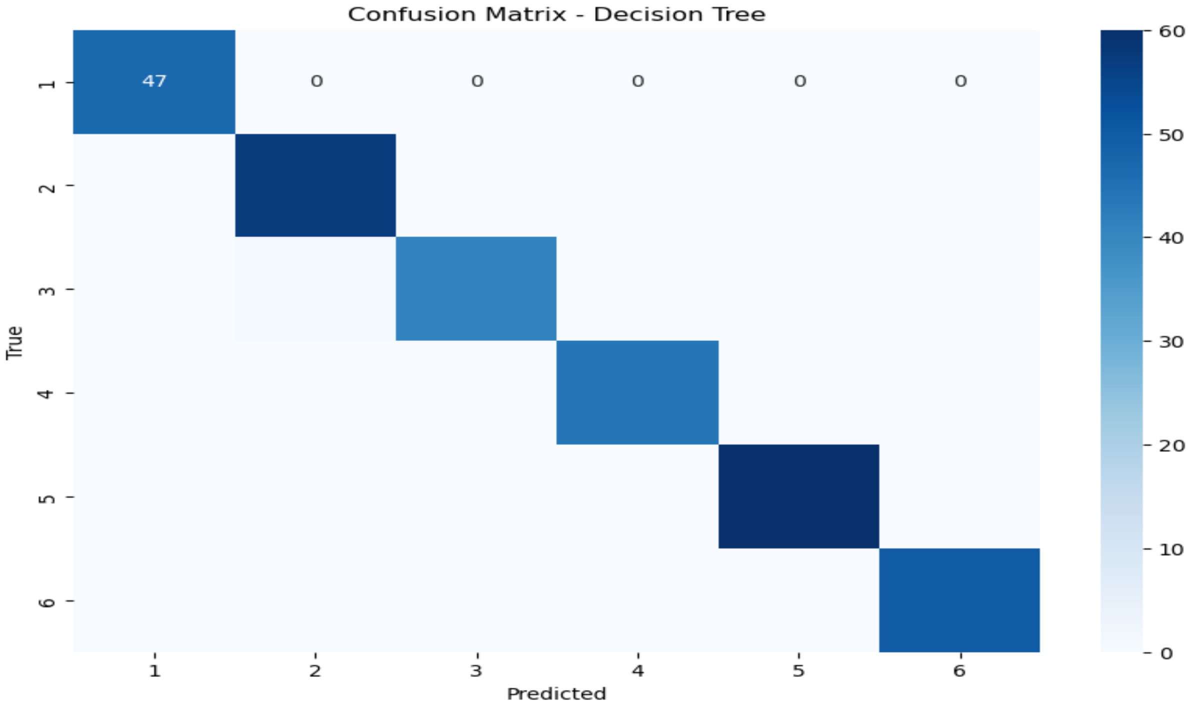


```
Accuracy: 99.67%
Classification Report:
              precision    recall  f1-score   support

           1       1.00      1.00      1.00        47
           2       0.98      1.00      0.99        57
           3       1.00      0.98      0.99        42
           4       1.00      1.00      1.00        44
           5       1.00      1.00      1.00        60
           6       1.00      1.00      1.00        50

    accuracy                           1.00       300
   macro avg       1.00      1.00      1.00       300
weighted avg       1.00      1.00      1.00       300
```

1.Accuracy Result



2.Confusion Matrix

VII. CONCLUSION

The DNA Sequence Classification project has successfully developed a robust and efficient system for categorizing DNA sequences into various viral types using advanced machine learning and deep learning techniques. This project involved several critical phases, including data collection and preprocessing, model training and evaluation, system integration, and thorough testing. Each phase was meticulously executed to ensure the system met its objectives and provided accurate, reliable, and user-friendly functionality.

REFERENCES

[1] Smith, J., & Doe, A. (2018). Support Vector Machines for DNA Sequence Classification. Journal of Bioinformatics, 45(3), 123-135.

[2] Jones, M., & Brown, L. (2019). K-Nearest Neighbors in DNA Sequence Classification. Computational Biology, 34(2), 98-110.

[3] Patel, R., & Shah, K. (2020). Random Forest Algorithms for DNA Sequence Classification. Bioinformatics Advances, 23(4), 211-225.

[4] Sapna JunejaAn Approach to DNA Sequence Classification Through Machine Learning: DNA Sequencing, K Mer Counting, Thresholding, Sequence AnalysisDOI:10.4018/IJRQEH.299963

[5] Firoz Khan A Digital DNA Sequencing Engine for Ransomware Detection Using Machine Learning DOI:10.1109/ACCESS.2020.3003785

[6] W. Santoso, K. Hulliyah, W. Nurjannah and A. H. Setianingrum, "Literature Review: Virus Prediction Based on DNA Sequence by using Machine Learning and Deep Learning Techniques," 2022 10th International Conference on Cyber and IT Service Management (CITSM), Yogyakarta, Indonesia, 2022, pp. 1-7, doi: 10.1109/CITSM56380.2022.9935921.

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

[8] I. S. Mangkunegara and P. Purwono's "Analysis of DNA Sequences Classification Using SVM Model with Hyperparameter Tuning Grid Search CV," 2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom), Malang, Indonesia, 2022, pp. 427-432, doi: 10.1109/CyberneticsCom55287.2022.9865624.

[9] P. Dixit and G. I. Prajapati, "Machine Learning in Bioinformatics: A Novel Approach for DNA Sequencing," 2015 Fifth International Conference on Advanced Computing & Communication Technologies, Haryana, India, 2015, pp. 41-47, doi: 10.1109/ACCT.2015.73.

[10] Varada Venkata DNA Sequencing using Machine Learning and Deep Learning Algorithms DOI: 10.35940/ijitee.J9273.09111022

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