DAYANANDA SAGAR UNIVERSITY

### SCHOOL OF ENGINEERING

**KUDLU GATE, BANGALORE – 560068**



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

(DATA SCIENCE) PROJECT REPORT ON

**“FORENSIC DNA PROFILING CLASSIFIER**” 2022-2023

BACHELOR OF TECHNOLOGY IN

COMPUTER SCIENCE & ENGINEERING DATA SCIENCE

Submitted by

CHAITHRA.K -ENG21DS0014 SHUBHAM KUMAR-ENG21DS0039 NIVAS REDDY-ENG21DS0046

Under The Supervision of:

**Dr . Kakoli Bora Associate Professor**

**Department of CSE-Data science,DSU**

DAYANANDA SAGAR UNIVERSITY

#### DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING SCHOOL OF ENGINEERING

**DAYANANDA SAGAR UNIVERSITY KUDLU GATE, BANGALORE – 560068**



***CERTIFICATE***

#### It is certified that the project work entitled “ALCOHOL DETECTION AND ENGINE LOCKING SYSTEM” has been

**carried out at *Dayananda Sagar University*, Bangalore, by**

*Chaithra.K-ENG21DS0014, Shubham Kumar-ENG21DS0039, Nivas Reddy-ENG21DS0046*, **Bonafide student of fourth Semester, B.tech.in partial fulfilment for the award of degree in *Bachelor of Technology in Computer Science & Engineering, Data Science* during academic year*2022-23*. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report deposited in departmental library. The project report has been approved as it satisfies the academic requirements in respect of project work for the said degree.**

#### Signature of the Guide Signature of the Chairperson

The completion of project brings with and sense of satisfaction, but it is never completed without thanking the person who are all responsible for its successful completion. We wish to express our profound feelings of gratitude to this great institution of our DAYANANDA SAGAR UNIVERSITY for providing the excellent facilities.

I am especially thankful to our Chairperson, Dr. Shaila S G, for providing necessary departmental facilities, moral support and encouragement. The largest measure of our acknowledgment is reserved for Prof. Kakoli bora whose guidance and support made it possible to complete the project work in a timely manner.

I have received a great deal of guidance and co-operation from the staff and I wish to thank all that have directly or indirectly helped me in the successful completion of this project work.

CHAITHRA.K-ENG21DS0014 SHUBHAM KUMAR-ENG21DS0039 V NIVAS REDDY-ENG21DS0046

We hereby declare that the project entitled “FORENSIC DNA PROFILING CLASSIFIER” submitted to Dayananda Sagar University, Bengaluru, is a bona fide record of the work carried out by me under the guidance of Prof.KAKOLI BORA, associate professor- department of computer science-data science, School of Engineering, Dayananda Sagar University, and this work is submitted in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering(Data Science).

CHAITHRA.K-ENG21DS0014 SHUBHAM KUMAR-ENG21DS0039 V NIVAS REDDY-ENG21DS0046

## ABSTRACT

This project streamlines DNA analysis post-natural disasters by comparing Short Tandem Repeat (STR) regions. Our method combines DNA sequencing, machine learning classification, and gene matching on a diverse real-world dataset. To understand dataset features, we applied correlation matrices and univariate selection, utilizing confusion matrices for training and testing. Integration of Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), and other techniques enhances classification and analysis, determining the matching percentage of unknown DNA against a known dataset.

Our framework expedites DNA analysis, simplifying disaster victim identification. Automation optimizes matching, saving time and resources. This project leverages advanced technologies to address complex DNA mixtures in disaster scenarios, marking a significant step forward in disaster victim identification.

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **CHAPTER** | **PG.NO** |
| INTRODUCTION | 7 |
| OBJECTIVES AND SCOPE OF WORK | 9 |
| DESCRIPTION OF WORK | 10 |
| DESIGN | 11 |
| REQUIREMENTS | 12 |
| METHODOLOGY | 13 |
| IMPLEMENTATION-CODE | 14 |
| RESULTS | 19 |
| CONCLUSION | 23 |
| REFERENCES | 24 |

### INTRODUCTION

In the wake of natural disasters, the irreplaceable loss of life creates an immense void, leaving families and communities grappling with profound grief. Amid this tragic reality, there arises a pressing humanitarian imperative: to bridge the agonizing gap between the departed and their grieving loved ones. To address this urgent need, a pioneering initiative harnesses the formidable capabilities of DNA analysis and cutting-edge machine learning techniques.

This groundbreaking project centers on the collection of DNA samples from victims of natural disasters. Through the adept utilization of advanced analysis and machine learning algorithms, its primary objective is clear: to reunite the departed individuals with their families. This initiative aims to serve as a beacon of hope, offering solace and closure during the most trying of times. By leveraging technological advancements, it endeavors to alleviate the profound suffering that ripples through our communities in the aftermath of such calamities.



Recent strides in deep learning have revolutionized the realm of automated feature extraction from complex datasets. This research marks a pivotal juncture by introducing sophisticated deep learning frameworks specifically tailored to classify DNA sequences. These frameworks adeptly navigate the challenges of feature extraction inherent in genomic data, paving the way for precise and efficient classification. Their efficacy is rigorously evaluated using a diverse array of DNA sequences, providing invaluable insights into the realm of automating DNA analysis.

In essence, this initiative amalgamates the power of DNA analysis and machine learning, offering a glimmer of hope amidst devastation. It seeks to not only expedite the process of identifying victims but also to restore a sense of closure and comfort to those left behind. Through its innovative approach, it endeavors to alleviate the anguish and turmoil that reverberate through communities grappling with the aftermath of natural disasters.

### OBJECTIVE

This project focuses on using DNA analysis and machine learning to improve the identification of natural disaster victims. The primary goal is to quickly and accurately reconnect victims with their families by applying advanced techniques. Through detailed DNA analysis and customized machine learning, the project aims to ease the uncertainty faced by families after disasters, providing closure and an opportunity for mourning.

The initiative also aims to enhance identification processes by introducing specialized deep learning frameworks. These frameworks aim to speed up identifications, improve accuracy, and reduce the distress associated with prolonged uncertainty. Additionally, the project emphasizes alleviating emotional distress within affected communities by expediting identifications and supporting collective mourning.

Furthermore, the initiative aims to pioneer automated DNA analysis methods, potentially advancing genetic research. Ultimately, this humanitarian effort seeks to promptly assist in accurate victim identification, contributing to swift responses and aid distribution crucial in post-disaster recovery efforts.

### Scope of work

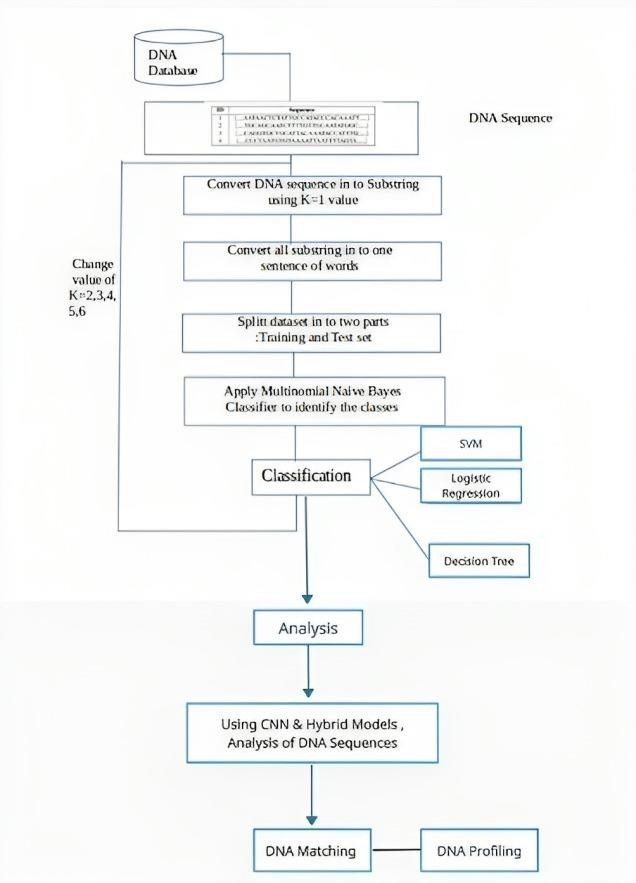
The scope involves collecting and processing DNA samples from disaster victims using advanced analysis and machine learning. This includes developing deep learning frameworks for precise DNA sequence classification and creating algorithms for automated feature extraction. The project aims to accurately identify victims, facilitate family reunification, ensure ethical compliance, and document methodologies and outcomes for broader scientific and humanitarian contributions. Collaboration with stakeholders and continuous improvement in methodologies are integral to this scope.

### DESCRIPTION OF WORK

The work involves testing unknown DNA profiles taken often from disaster scenes, by comparing STR regions with known profiles. These incomplete profiles, common in calamities, undergo DNA sequencing and classification using machine learning classifiers. Feature understanding is achieved through correlation matrix and univariate selection techniques. Training and testing datasets are evaluated using confusion matrices. Various techniques including SVM and naïve bayes classifier are employed for classification, determining the matching percentage of unknown DNA with known profiles. This streamlined approach simplifies DNA matching, significantly saving time and effort in identification processes.

This project focuses on DNA sequencing, classification, and detailed analysis. We're also integrating gene matching to improve the accuracy and efficiency of our analysis. Gene matching utilizes deep learning to classify DNA sequences, addressing feature extraction challenges. By employing machine learning, the model interprets complex DNA profiles, identifies patterns, and enhances match accuracy, especially in cases involving mixtures or partial profiles. The model's performance is evaluated using a diverse DNA sequence dataset.

**DESIGN**



### REQUIREMENTS

DATASET :

* Dataset description: The dataset contains human DNA sequences labeled with corresponding classes, representing specific classifications . These sequences have been systematically collected and organized for the purpose of classification and analysis, facilitating research and insights into genetic pattern matching and attributes.
* Dataset used:[https://raw.githubusercontent.com/krishnaik06/DNA-](https://raw.githubusercontent.com/krishnaik06/DNA-Sequencing-Classifier/master/human_data.txt) [Sequencing-Classifier/master/human\_data.txt](https://raw.githubusercontent.com/krishnaik06/DNA-Sequencing-Classifier/master/human_data.txt)

SOFTWARE REQUIREMENTS:

* Jupyter notebook
* Git hub

### METHODOLOGY

We propose the idea in which the **testing of DNA is accomplished by comparing STR regions of known STR DNA profiles with unknown STR DNA profiles**, such as profiles that have been taken from a natural calamity or disaster scene. These profiles usually consist of incomplete STR profiles from an number of unknown contributors

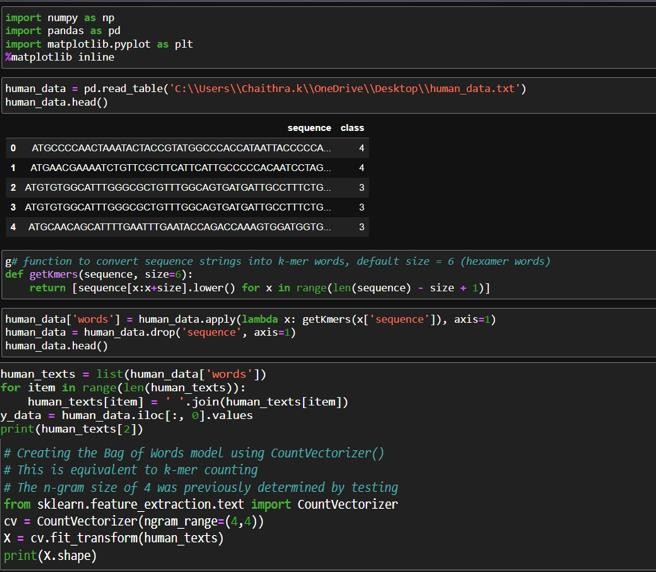
**Our dataset comprises of profiles that might be encountered in actual cases**. Numerous machine learning and deep learning interpretation tools, as well as analysis techniques, have been created to address DNA mixtures. Initially, DNA sequencing is performed. Subsequently, the dataset undergoes classification using a classifier. Following classification, gene matching is conducted, culminating in the completion of the DNA analysis process.

It is essential to know the features related to our dataset , hence we applied two techniques: **the correlation matrix and feature selection .we use confusion matrix for training and testing datasets . we incorporate [CNN]convolution neural network, naïve bayes classifier and other machine learning techniques to classify and do the analysis of the dataset and find the matching percentage of the unknown DNA among the existing known DNA dataset, for gene matching we use smith waterman algorithm.**

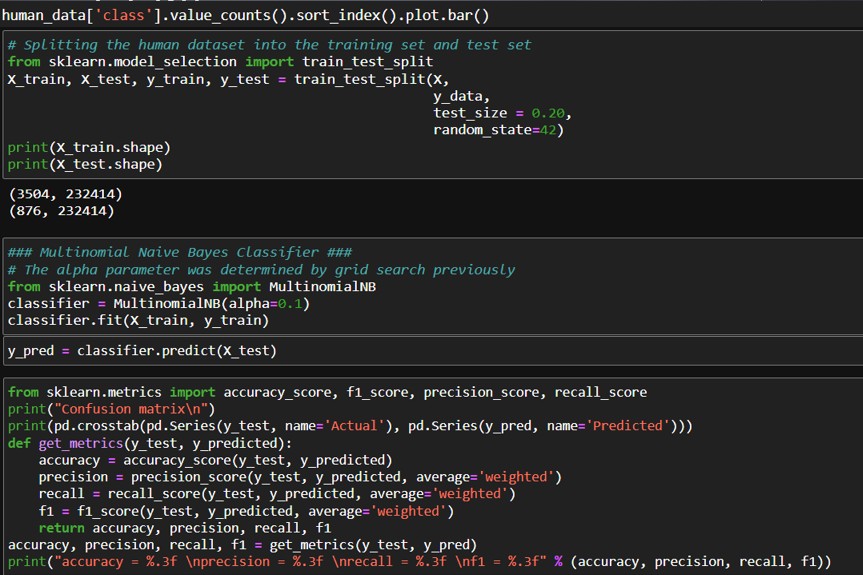
Hereby , making the task of finding a persons DNA match easier . Hence saves a lot of time and trouble.

**IMPLEMENTATION**

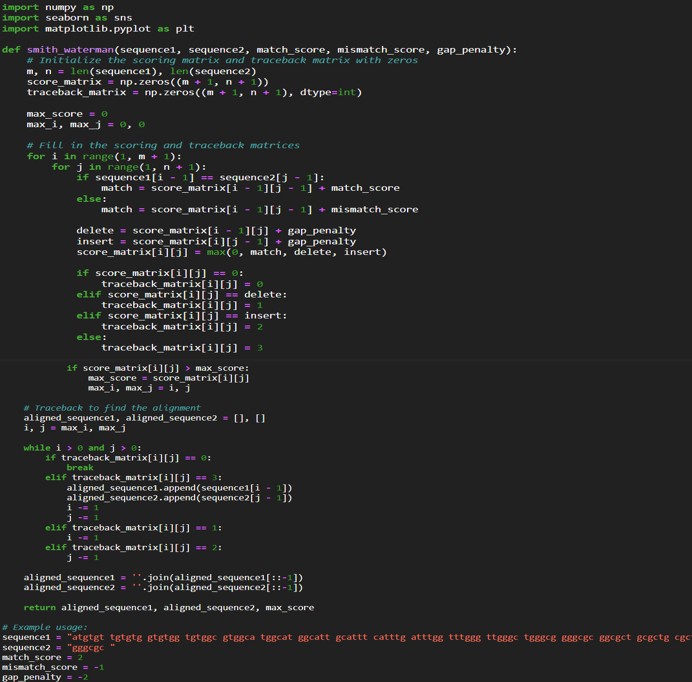
# SOURCE CODE

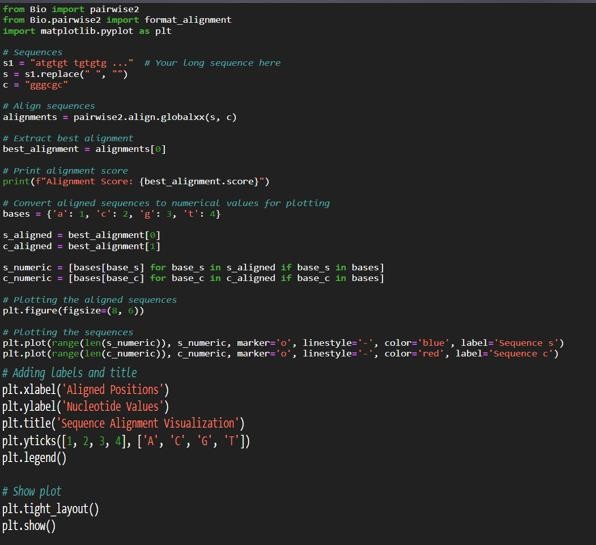


CLASSIFICATION – FEATURE SELECTION



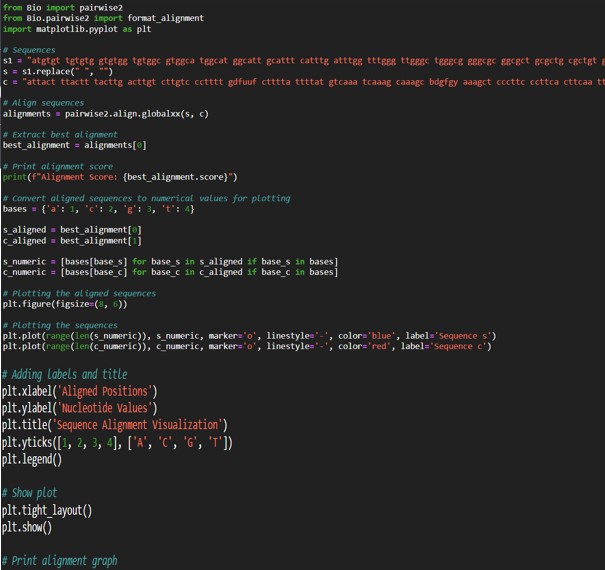
GENE MATCHING-SMITH WATTERMAN ALGORITHM





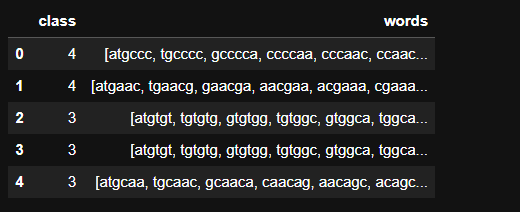
VISUALIZATION



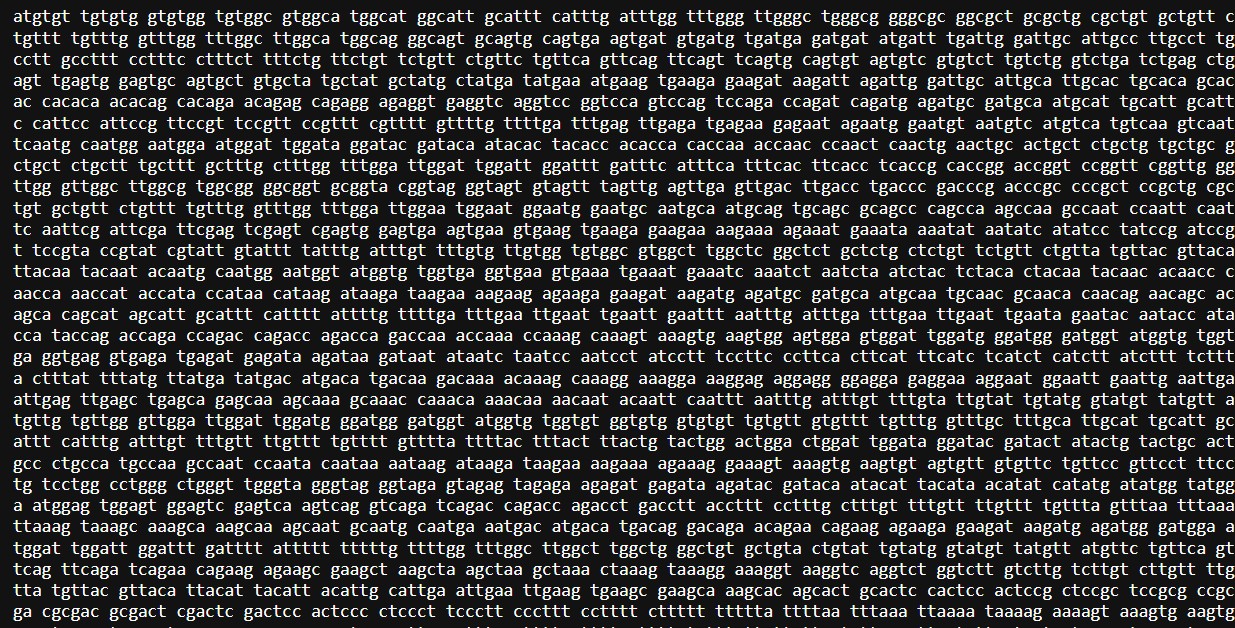


**RESULTS**

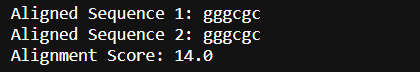
DIVIDING THE DATASET INTO CLASSES:



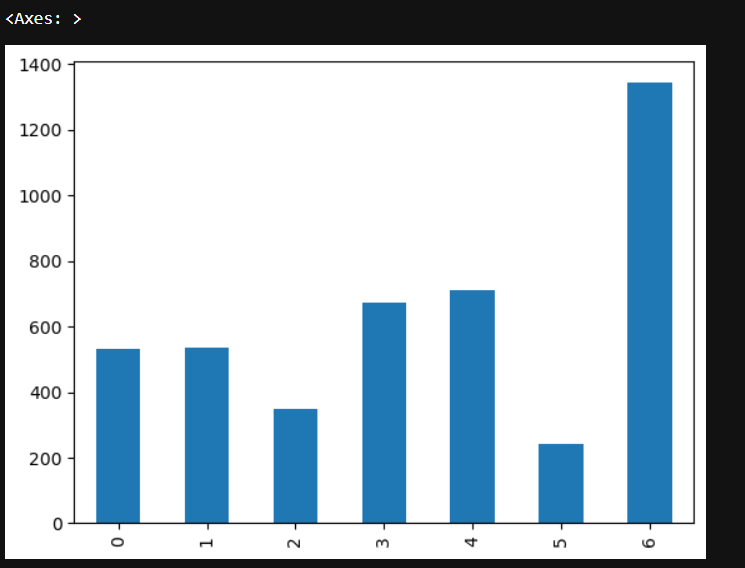
SEQUENCING THE DNA



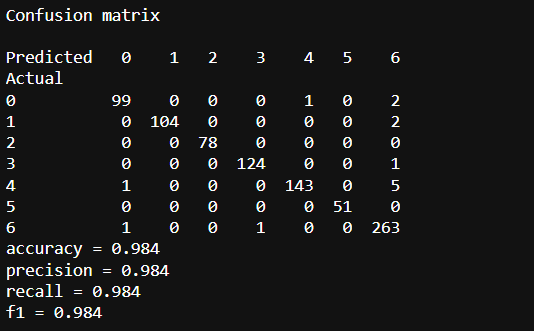
GENE MATCHING



FEATURE EXTRACTION FROM CLASSES:

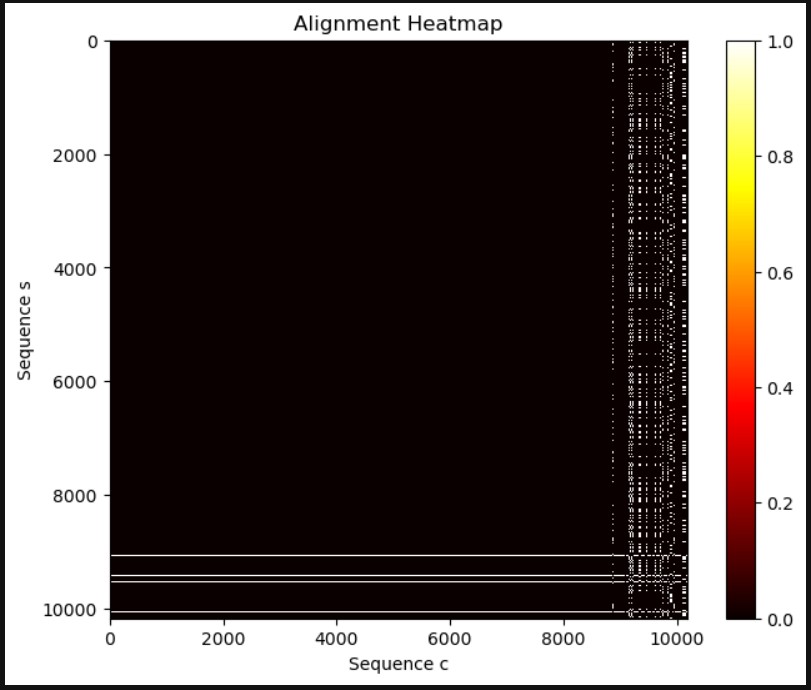


CONFUSION MATRIX

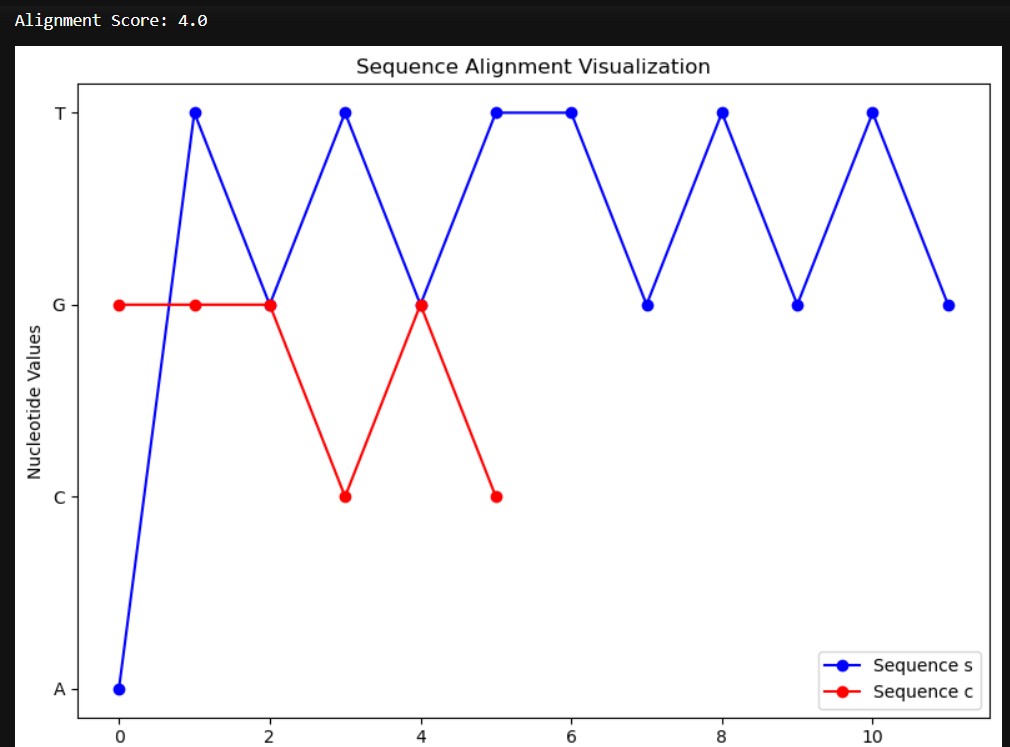


VISUALIZATION:

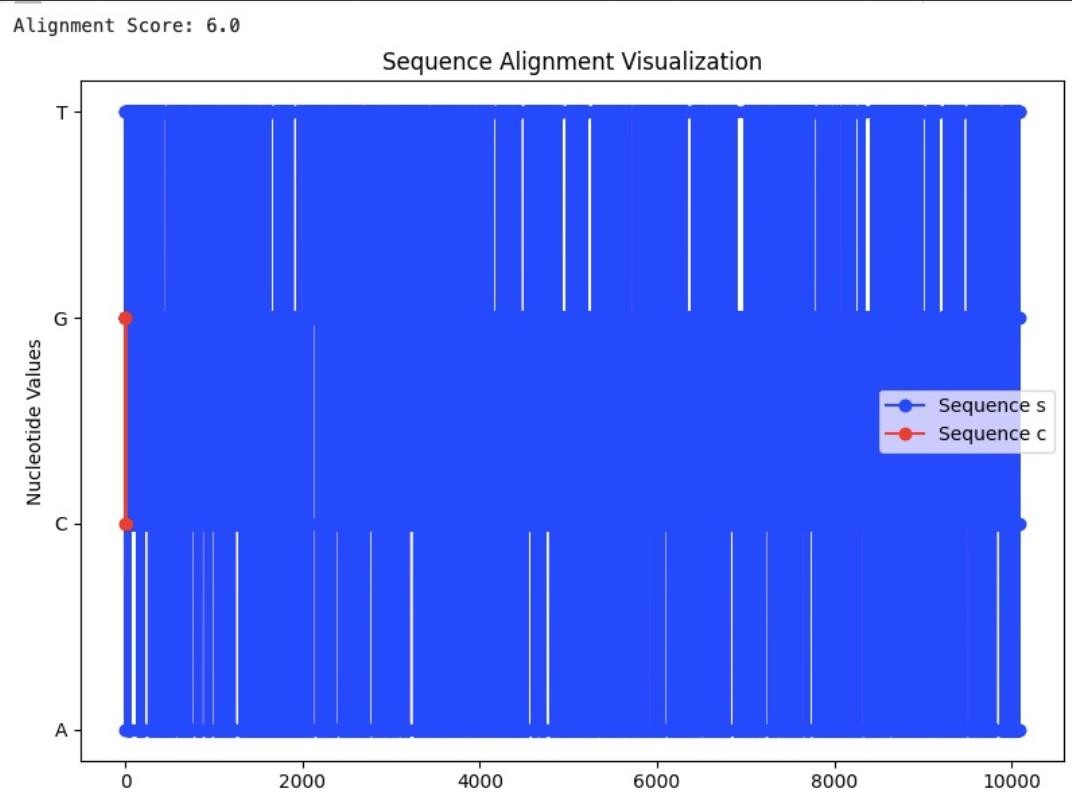
HEAT MAP



LINE GRAPH:



SEQUENCE ALIGNMENT VISUALIZATION



# CONCLUSION

The project uses DNA profiling, machine learning, and deep learning to speed up DNA analysis during disasters. It starts with DNA sequencing and compares known and unknown STR DNA profiles. Classification involves tools like CNN, SVM, and other machine learning methods. Correlation matrices and feature selection help identify essential features for accurate classification.

Introducing gene matching brings a new approach, using deep learning to interpret complex DNA sequences and improve pattern identification. This innovation significantly enhances accuracy, especially in cases with partial or mixed profiles, providing a strong solution for matching unknown DNA with existing records.

Beyond scientific progress, the project aims to bring comfort to those affected by natural disasters. Reuniting the DNA of the deceased with their families offers hope amid grief, easing suffering in communities. This holistic approach streamlines DNA analysis and serves as a compassionate support during tragic events, underlining the project's significant societal impact.

**REFERENCES**

1. [https://thesai.org/Publications/ViewPaper?Volume=12](https://thesai.org/Publications/ViewPaper?Volume=12&Issue=11&Code=IJACSA&SerialNo=15) [&Issue=11&Code=IJACSA&SerialNo=15](https://thesai.org/Publications/ViewPaper?Volume=12&Issue=11&Code=IJACSA&SerialNo=15)
2. <https://www.hindawi.com/journals/cmmm/2021/1835056/>
3. [https://medium.com/mlearning-ai/apply-machine-learning-](https://medium.com/mlearning-ai/apply-machine-learning-algorithms-for-genomics-data-classification-132972933723) [algorithms-for-genomics-data-classification-132972933723](https://medium.com/mlearning-ai/apply-machine-learning-algorithms-for-genomics-data-classification-132972933723)
4. [https://www.researchgate.net/figure/Steps-for-](https://www.researchgate.net/figure/Steps-for-classification-of-DNA-sequence-feature-extraction-feature-encoding-and_fig1_349147096)

[classification-of-DNA-sequence-feature-extraction-feature-](https://www.researchgate.net/figure/Steps-for-classification-of-DNA-sequence-feature-extraction-feature-encoding-and_fig1_349147096) [encoding-and\_fig1\_349147096](https://www.researchgate.net/figure/Steps-for-classification-of-DNA-sequence-feature-extraction-feature-encoding-and_fig1_349147096)

1. [https://github.com/krishnaik06/DNA-Sequencing-](https://github.com/krishnaik06/DNA-Sequencing-Classifier/blob/master/DNA%20Sequencing%20and%20applying%20Classifier.ipynb)

[Classifier/blob/master/DNA%20Sequencing%20and%20appl](https://github.com/krishnaik06/DNA-Sequencing-Classifier/blob/master/DNA%20Sequencing%20and%20applying%20Classifier.ipynb) [ying%20Classifier.ipynb](https://github.com/krishnaik06/DNA-Sequencing-Classifier/blob/master/DNA%20Sequencing%20and%20applying%20Classifier.ipynb)

1. [https://medium.com/mlearning-ai/apply-machine-learning-](https://medium.com/mlearning-ai/apply-machine-learning-algorithms-for-genomics-data-classification-132972933723) [algorithms-for-genomics-data-classification-132972933723](https://medium.com/mlearning-ai/apply-machine-learning-algorithms-for-genomics-data-classification-132972933723)