International Journal of Emerging Engineering and Technology (IJEET)
ISSN (e): 2958-3764
Volume: 3, Number: 2, Pages: 6 - 12, Year: 2024

# Deoxyribose Nucleic Acid Nucleotide Virus Classification with Machine Learning

Imran Hussain[1], Muhammad Rizwan[2], Shoaib Nawaz[3], Danish Gaffar[4], Waseem Akram[4]
[1] The Islamia University of Bahawalpur, Bahawalpur, Pakistan
[2] Khwaja Fareed University of Engineering and Information Technology, RYK, Pakistan
[3] iAxon Research Centre, Rahim Yar Khan, Pakistan
[4] Virtual University of Pakistan, Pakistan
[*]Corresponding Author: Imran Hussain (Email Address: imranch9417@gmail.com)

**Received: 08-08-2024 Revised: 20-11-2024 Accepted: 20-12-2024**

*Abstract*— **Deoxyribose Nucleic Acid (DNA) viruses are a major focus in virology because they cause many different types of human illness. DNA viruses' biology, transmission, and pathogenicity can be better comprehended if they are first properly categorized. Recently, machine learning has proven to be an effective method for studying massive amounts of biological data, such as DNA viral sequences. Here, we give a high-level overview of utilizing machine learning to categorize DNA viral sequences. We address supervised, unsupervised, and deep learning strategies that have been used for DNA viral sequence classification. The data's high dimensionality, highly variable sections, and the requirement to differentiate between closely related viral strains are only a few of the difficulties we emphasize in DNA virus sequence classification. The FASTA Tool was used to retrieve the dataset from the NCBI database. The data collected included three human gene family sequences and six virus sequences. We employed six machine learning methods to classify the DNA viruses (Sars-Cov-1, Mers-Cov-2, Ebola, Dengue, Influenza, Synthase, Ion Channel, and Transcription Factor) with 98% accuracy.**

*Keywords*— **Deoxyribose Nucleic Acid (DNA)*, Machine Learning, NCBI Library, FASTA.***

## I. Introduction

The acronym DNA refers to deoxyribonucleic acid. It's a molecule that all creatures, from bacteria to plants, animals, and humans, use to grow and function. The nucleotides that makeup DNA coil around each other in a lengthy, double-stranded molecule. Sugar (deoxyribose), phosphate, and a nitrogenous base comprise each nucleotide (adenine, guanine, cytosine, or thymine)[1]. The genetic information that is transferred from one generation to the next is determined by the order in which these nitrogenous bases couple up to create the rungs of the DNA ladder.

In humans, there are two main types of DNA: nuclear DNA and mitochondrial DNA. Nuclear DNA: This is the most well-known type of DNA in humans. It is located in the cell's nucleus and contains most of an individual's genetic material. Nuclear DNA is inherited from both parents and contains approximately 3 billion base pairs that code for about 20,000 to 25,000 genes[2].

Mitochondrial DNA: This type of DNA is located in the mitochondria of the cell, which are the organelles responsible for producing energy. Unlike nuclear DNA, mitochondrial DNA is inherited only from the mother, as the sperm contributes very little, if any, mitochondria to the fertilized egg. Mitochondrial DNA contains only about 16,500 base pairs and codes for only a few dozen genes[3]. Because mitochondrial DNA is passed down exclusively from the mother, it can be used to trace maternal ancestry and relationships.

Viruses that use DNA as their own genetic material are called DNA viruses. DNA viruses are extremely versatile pathogens that can infect various living things[4]. They are categorized according to whether or not they have a double- or single-stranded DNA genome. Herpesviruses, papillomaviruses, adenoviruses, and poxviruses are all examples of DNA viruses that can infect humans. These viruses can cause various diseases, including cold sores, genital herpes, warts, respiratory infections, and smallpox[5].

Once a DNA virus infects a host cell, it uses its genetic material to hijack its machinery and produce more viral particles. Once inside a cell, the virus can replicate and infect other cells, further weakening the host. Certain DNA viruses can integrate their genetic material into the host's DNA, leading to the development of certain types of cancer.

DNA viruses are a group of viruses that have a DNA genome. They make copies of their genetic material (DNA) and new virus particles (virions) within the host cell. There are many different types of DNA viruses that can infect humans, animals, and plants.

**Some examples of DNA viruses that infect humans include:**

a) **Herpesviruses:** This family of viruses includes herpes simplex virus (HSV), varicella-zoster virus (VZV), and Epstein-Barr virus (EBV). Diseases like cold sores, chickenpox, shingles, and infectious mononucleosis can all be caused by these viruses.

b) **Papillomaviruses:** These viruses can cause warts and are associated with some types of cancer, including cervical cancer.
c) **Adenoviruses:** These viruses can cause respiratory illness, conjunctivitis, and other infections.
d) **Poxviruses:** These viruses include variola virus, which causes smallpox, and vaccinia virus, which is used as a vaccine against smallpox.
e) **The hepatitis B virus (HBV)** is a member of the hepadnaviral family and is a known cause of liver disease and cancer.

### *A. Objectives*

Build an Artificial Intelligence Model to predict the Virus disease from the DNA/RNA Nucleotide Sequences.

### *B. Motivation*

DNA sequences, the complex and clustered data cannot openly tell us what the sequence is, what is in the sequence, and what the nucleotide associations are. From the Big Data, Humans cannot find which sequence is and what it is associated with a specific disease or specific virus.

- AI model Collected Sequences read and predict the disease in the DNA/RNA Sequences.
- Automate the Screening processes to classify the DNA diseases.

### *C. Problem Statement*

- DNA / RNA Sequences are very difficult for humans to read and understand in order to diagnose diseases.
- We cannot manually classify the multiple sequences at a time.
- Nucleotides can be read but cannot be classified at the reading stage for assessment of the Category of the disease in the Sequences.
- Required experience and knowledge of reading sequence and classification.

### *D. Statement Objectives*

The proposed study aims to develop an artificial intelligence model for DNA virus sequence classification. The main object of the proposed study is the collection and preparation of the required dataset and settlement of the data as machine learning algorithms require. After collecting and preparing, we trained and tested the algorithm and achieved good results. After the results, the model classification accuracy should be productive and efficient.

### *E. Proposed Study Methodology*

- DNA / RNA Sequences are very difficult for humans to read and understand to diagnose diseases.
- We cannot manually classify the multiple sequences at a time.
- Nucleotides can be read but cannot be classified at the reading stage for assessment of the Category of the disease in the Sequences.
- Required experience and knowledge of reading sequence and classification.

## II. Literature Review

An enhancer is a small segment of DNA (50-1500 base pairs) that stimulates the expression of genes and the subsequent synthesis of RNA and proteins. Several human diseases, including cancer, disorder, and inflammatory bowel disease, have been linked to genetic polymorphism in enhancers. Enhancers play a crucial role in genomics; hence, the classification of enhancers is an active field of study in computational biology[5]. Even if some computational tools have been used to try and solve this issue, there is certainly room for advancement in how well they perform. To categorize enhancers, we use a support vector machine technique and feed it word embedding, which includes sub-word information of its biological words[7].

We introduce iEnhancer-5Step, a server on the web that uses two-layer classifiers to determine the presence and potency of enhancers. Over two levels, our independent test accuracy is 79% and 63.5%, respectively. Our proposed technique outperforms state-of-the-art predictors when tested on the same dataset[8]. Moreover, this study lays the groundwork for future investigation that can expand our understanding of how to effectively employ natural language processing methods in the context of biological sequences. You can download iEnhancer-5Step for no cost at http://biologydeep.com/fastenc[9].

It has become common practice to analyze animal diets by applying DNA sequencing-based methods to DNA taken from environmental materials, including stomach contents and. Meta-barcoding and shotgun sequencing are utilized more frequently to answer ecological problems based on dietary interactions due to their high resolution and prey detection capacity[5]. Despite their immense potential, new studies have shown how numerous technical (relating to the methodology) and biological (relating to the study system) aspects can obscure the true signal of taxonomic diversity[10]. This article summarizes this research in light of a new method for evaluating trophic interactions that rely on high-throughput sequencing. We discuss how distortion variables can be considered in the study's design and how it is critical to recognize the limitations and biases inherent to sequencing-based diet analysis to achieve trustworthy results and draw suitable conclusions. We also provide recommendations for scaling up DNA sequencing-based dietary assessments and methods for reducing the influence of distortion factors. In doing so, we hope to assist end-users in designing robust diet studies by educating them on the complexities and limitations of DNA sequencing-based diet analyses, and we hope to inspire researchers to develop and improve the tools that will ultimately propel this field to maturity [11]. NCBI library helps us improve data collection [12].

The necessity for models that can efficiently compress DNA sequences without introducing any inaccuracies has grown as a direct result of the explosion in genomic data creation. Long-term data storage and compression analysis are two of the most important uses. A few recent studies suggest employing neural networks to achieve DNA sequence compression, which is a significant gap in the existing literature[11], [12]. However, they are inferior to more specialized DNA compression methods like GeCo2. This restriction arises because there are no models tailored to DNA sequences. Our work uses the strength of neural networks alongside targeted DNA models [5]. To achieve this goal, we developed GeCo3, a novel genomic sequence compressor that

uses neural networks to combine models of numerous contexts and substitution-tolerant contexts into a single prediction. Findings: Y-chromosome and human mito-genome, two collections of archaeal and virus genomes, four complete genomes, and two collections of FASTQ data of a human virome and ancient DNA are just some of the datasets we use to evaluate GeCo3's performance as a reference-free DNA compressor[6].

GeCo3 significantly improves compression ratios (2.4%, 7.1%, 6.1%, 5.8%, and 6.0%) compared to its predecessor, GeCo2. Tests with primate genome datasets achieve compression gains of 12.4%, 11.7%, 10.8%, and 10.1% over state-of-the-art. While GeCo3 processes data 1.7–3 times slower than GeCo2, it maintains consistent RAM usage regardless of sequence length. This improvement is achieved through a neural network mixing methodology, which can be adapted to various data compressors. GeCo3 is available under the GPLv3 license at https://github.com/cobilab/geco3.

## III. Methodology Framework

The methodology section helps us analyze the issues and track in detail what the problem is and what factors we will face.

### A. *Problem Identification*

The problem with the proposed study is that not all DNA sequences are usually directly recognized by humans. What are the reasons for this?

- The DNA Sequences are huge in terms of DNA nucleotides. The number of characters differs, like AGCT, which includes different characters due to its associated type.
- The sequences are the same for DNA, RNA, Genes, and even viruses, so the type of the sequences is also a challenge.
- Humans, animals, and other living bodies have different genes, so the sequences are difficult to understand if it's from humans or other living bodies.
- No standard defined and fixed charter is associated with the specific virus type.
- There is no standard character specification of the length of the sequences.
- No specific gap or repetition of the character helps us to find or identify the sequence type.

### B. *Problem Fragments*

There are three parts to the problem. The first one is how to find the data and recognize and verify that the sequence is perfect and related to the specific virus. So, we collected the data using the NCBI library and the DNA Sequence downloader open-source application. The verification of the dataset is also used by the DNA Sequence analyzer, which is accessible on the online services for the verification of sequence and gene family. The second is how to clean the dataset, what kind of character is important, what DNA nucleotide characters are important, and how to remove the N-based character from the sequences. The third part is how to build or extract features from the sequences because all sequences have minor differences prominently.

### C. *Machine Learning and DNA Sequences*

Based on the proposed study, we analyze the sequences and how to convert them into machine learning understandable features. So, we decided to build the K-Mer feature and K-Mer values for each sequence one K-Mar feature will consist 4 characters, previously we used 6, but it's not suitable, so we moved back to 4 characters because each DNA sequences have the core four nucleotides with the feature using analysis helps us to use K-Mer values is not more than four characters.

### D. *DNA Sequence Code*

The DNA or RNA have Nucleotide codes that are AGCT. Each A is for adenine, G for Guanine, and C for cytosine, as shown in Fig 1 T for thymine. The composition of each DNA virus sequence has many characters with specific patterns of nucleotides in the specific sequence of the virus. It may consist of N values, null characters, and some other characters, so we need to clean them and get the pure DNA sequence for the specific class. So, the sequences need to be cleaned and classified[19].

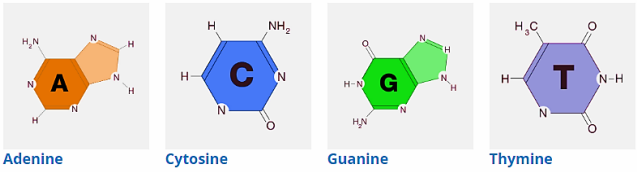



**Fig 1:** DNA Nucleotide structure

. The DNA nucleotide sequence characters have different meanings and logic, so we deal with them in a specific way to get the required data. Each code has its specific meaning and identification, and the best thing is that it changes the nature of the DNA Sequence type, and based on the type, the virus also has the same psychology.

### E. *Methodology Diagram.*

The whole process is defined in the processed model diagram. Figure 2 describes the methodology diagram, and Fig 3 describes the entire process flow of the proposed method.

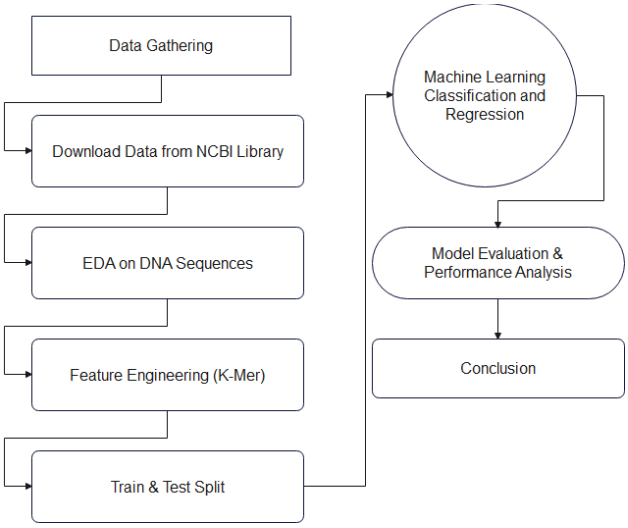



**Fig 2:** Process Model Diagram

## F. *Processed Model Diagram*

The processed model diagram delineates the proposed study's workflow from initiation to completion. It encompasses the following stages:

### *1) Data Collection*

The NCBI Library is an open public Biomedical library, and it contains many data collections that can be downloaded. The NCBI library helps us to download data directly from FTP or DNA Sequence Downloader, which is called FASTA Tool. FASTA is the file format in which the DNA Sequences are stored. This File can be converted into text using the Bio-Python Library. The FASTA File contains the Sequence ID, Type, Content, Gene, etc.

### *2) BLAST TOOL*

BALAST Tool is the tool that can used to download data from the NCBI library. The tool is free and open-source for use on multiple operating system platforms.

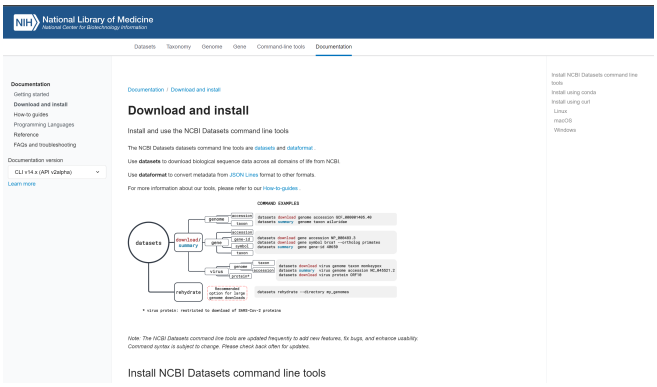



**Fig 3:** BLAST TOOL

### *3) FASTA File.*

The FASTA File contains the data and a combination of AGCT patterns representing the specific association with DNA viruses or genes. The DNA sequence file of the FASTA file looks like this, which is shown in Fig 4.

**Fig 4:** FASTA File DNA Sequence

## G. *EDA for Dataset*

After downloading the dataset, we do EDA on the data that takes us to the feature selection.

### *1) Reading Dataset:*

| | PID | CLASS | CLASSNAME | SEQ |
|---|---|---|---|---|
| 0 | AB257344 \|SARS coronavirus Frankfurt 1 genomic... | 1 | SARS-COV-1 | GATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTA... |
| 1 | AH013708 \|SARS coronavirus Sin0409\| partial se... | 1 | SARS-COV-1 | CATTCAGTACGGTCGTAGCGGTATAACACTGGGAGTACTCGTGCCA... |
| 2 | AH013709 \|SARS coronavirus Sin_WNV\| partial se... | 1 | SARS-COV-1 | CACGCGCGGGCAAGTCAATGTGCACTCTTTCCGAACAACTTGATTA... |
| 3 | AP006557 \|SARS coronavirus TWH genomic RNA\| co... | 1 | SARS-COV-1 | ATATTAGGTTTTTACCTACCCAGGAAAAGCCAACCAACCTCGATCT... |
| 4 | AP006558 \|SARS coronavirus TWJ genomic RNA\| co... | 1 | SARS-COV-1 | ATATTAGGTTTTTACCTACCCAGGAAAAGCCAACCAACCTCGATCT... |
| ... | ... | ... | ... | ... |
| 2909 | GENE1410 | 8 | 0 | ATGGAAGATTTGGAGGAAACATTATTTGAAGAATTTGAAAACTATT... |
| 2910 | GENE1411 | 8 | 0 | ATGGAAGATTTGGAGGAAACATTATTTGAAGAATTTGAAAACTATT... |
| 2911 | GENE1412 | 8 | 0 | ATGGAAGATTTGGAGGAAACATTATTTGAAGAATTTGAAAACTATT... |
| 2912 | GENE1413 | 8 | 0 | ATGGAAGATTTGGAGGAAACATTATTTGAAGAATTTGAAAACTATT... |
| 2913 | GENE1414 | 8 | 0 | ATGGAAGATTTGGAGGAAACATTATTTGAAGAATTTGAAAACTATT... |

2914 rows × 4 columns

**Fig 5:** Dataset view after reading.

Since the data consists of the DNA Nucleotide AGCT, which is called Sequence, the Outcome is the class or category of the DNA Sequence.

### *2) Classes in the Dataset:*

The type of the DNA virus sequence names are listed below, and Fig 6 describe the classes in the count of the sequence type.

1) SARS-COV-1
2) MERS
3) SARS-COV-2
4) EBOLA
5) DENGUE
6) INFLUENZA
7) SYNTHASE
8) ION CHANNEL
9) TRANSCRIPTION FACTOR

## H. *Data Encoding and its meaning in DNA Nucleotide:*

The dataset is collected from the NCBI website as DNA nucleoid sequences. These sequences have some meaning for some chargers used in the sequences. Each character belongs to some specific Mnemonics. Table 1 describes the encoding nucleotide and its meaning and Mnemonics.

TABLE I. NUCLEOTIDE CODE MEANING

| Nucleic Acid Code | Meaning | Mnemonic |
|---|---|---|
| A | A | Adenine |
| C | C | Cytosine |
| G | G | Guanine |
| T | T | Thymine |
| U | U | Uracil |
| (I) | I | Inosine (Non-Standard) |
| R | A or G (I) | puRine |
| Y | C, T or C | Base which are ketones |
| M | A or C | Base with aMino Group |
| S | C or G | Strong interaction |
| W | A, T or U | Week interaction |
| B | Not A(i.e. C,G T or U) | B comes after A |
| D | Not G (i.e. A, C , T or U) | D comes after C |
| H | Not G (A, C, T , U) | H comes after U |
| V | Neither T nor U (i.e A, C, or G) | V comes after U |
| N | A, C, G, T, U | Nucleic acid |
| | Gap of indeterminate length | |

## I. *FASTA File Structure*

Here is the representation of the FASTA File. All FASTA Files are placed in a single directory.

ID: ENST000000001234

**Sequence:**
AGCTAGTCAGTCATGCAACGTCAGTCAGGTTGCAG
GTCAGTT

The Data is collected in the form, and each file is associated with its class or category. Each file data is extracted using the Bio-Python Python library and combined with all sequences in one Comma-separated values file. The dataset's look is shown in Fig 5. The FASTA files how it's collected, managed, and extracted from the FASTA File into Comma Separated Values Files. Each DNA sequence is read through the Bio-Python, a Biomedical DNA Sequences file reading library, and the file reads as shown in Fig 4.9.


**Fig 7:** Reading and collecting FASTA File

## J. *Feature Engineering for Dataset*

As we have the dataset in the character as a string and the string we need to convert it into a feature for this purpose, we encode the character into metrics. We use the three general encoding processes.

a) Ordinal encoding DNA Sequence.
b) One-Hot encoding DNA Sequence
c) DNA Sequence is a language called K-MER counting.

With this method, each nitrogen base must be encoded as an ordinal value.

TABLE II. NUCLEOTIDE CHARACTER AND ORDINAL ENCODING

| DNA Sequence Character | DNA Sequence Encoding value |
|---|---|
| A | 0.25 |
| G | 0.75 |
| C | 1.0 |
| T | 0.5 |
| N | 0 |

"ATGC" becomes, for instance, [0.25, 0.5, 0.75, 1.0]. Any other base, like "N," can represent a 0. So, let's write a program that generates NumPy array objects from sequence strings and label encoders that use the letters "a," "c," "g," and "t" from the DNA alphabet as well as the character "n" for anything else. Figure 8 describes the encoding processes we use in the proposed study.

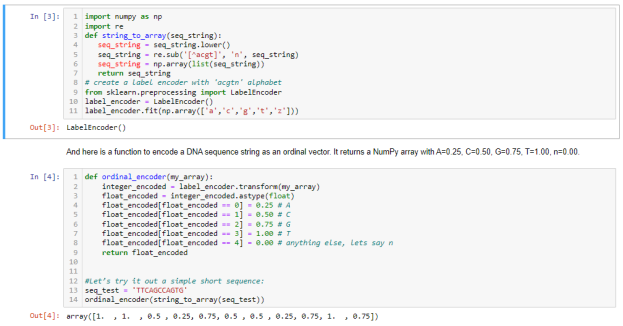



**Fig 8:** Deployment of the encoding on Dataset.

### 1) *K-MER As a Feature for Dataset*

In bioinformatics, k-mers are substrings of length k included within a biological sequence. Because k-mers are made up of nucleotides (i.e. A, T, G, and C), they are most often employed in the context of computational genomics and sequence analysis, where they are used to assemble DNA sequences,[1] enhance heterologous gene expression,[2] distinguish between species in metagenomic samples,[4] and develop attenuated vaccines. [5] For example, the sequence AGAT can be broken down into four monomers (A, G, A, and T), three 2-mers (AG, GA, AT), two 3-mers (AGA, GAT), and one 4-mer. These subsequences are collectively referred to as k-mers (AGAT). In a broader sense, given a sequence of length LL, there will be L-k+1 L-k+1 k-mers and nknk potential k-mers, where nn is the number of possible monomers (e.g. four in the case of DNA). Here is an example of two and three K-MERS contained in the DNA Sequence. The K-means are the feature of the DNA Sequence, and Fig 9 shows the features extraction processes, which k-mers used for the machine learning for model building.

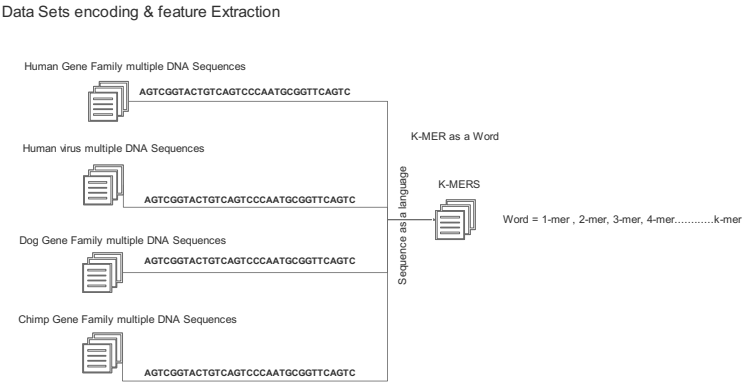



**Fig 9:** K-MER Feature engineering processes.

Figure 10 describes the sample view of the deployment of feature extraction code in the sample DNA sequence.

```
def Kmers_funct(seq, size):
    return [seq[x:x+size].lower() for x in range(len(seq) - size + 1)]


#So let's try it out with a simple sequence:
mySeq = 'GTGCCCAGGTTCAGTGAGTGACACAGGCAG'
Kmers_funct(mySeq, size=7)
```

```
['gtgccca',
 'tgcccag',
 'gcccagg',
 'cccaggt',
 'ccaggtt',
 'caggttc',
 'aggttca',
 'ggttcag',
 'gttcagt',
 'ttcagtg',
 'tcagtga',
 'cagtgag',
 'agtgagt',
 'gtgagtg',
 'tgagtga',
 'gagtgac',
 'agtgaca',
 'gtgacac',
 'tgacaca',
 'gacacag',
 'acacagg',
 'cacaggc',
 'acaggca',
 'caggcag']
```

**Fig 10:** Deployment of K-MER

*2) View of the Dataset after feature engineering*

In the Dataset, each sequence is converted into a K-MER Feature with Classes in each row. The row contains words of each K-MER, so it's called Word.

## *K. Training of the Algorithms*

Training of the Dataset is the 80 percent part of the dataset in which the sequences and classes of the sequences. The Test data is included without labels which is 20 percent of the dataset. First, we use a multinomial dataset.

```
human_data
```

| | Outcome | words |
|---|---|---|
| 0 | 1 | [gatctc, atctct, tctctt, ctcttg, tcttgt, cttgt... |
| 1 | 1 | [cattca, attcag, ttcagt, tcagta, cagtac, agtac... |
| 2 | 1 | [cacgcg, acgcgc, cgcgcg, gcgcgg, cgcggg, gcggg... |
| 3 | 1 | [atatta, tattag, attagg, ttaggt, taggtt, aggtt... |
| 4 | 1 | [atatta, tattag, attagg, ttaggt, taggtt, aggtt... |
| ... | ... | ... |
| 1495 | 6 | [agtatg, gtatgg, tatgga, atggaa, tggaaa, ggaaa... |
| 1496 | 6 | [aaagca, aagcag, agcagg, gcaggc, caggca, aggca... |
| 1497 | 6 | [agtatg, gtatgg, tatgga, atggaa, tggaaa, ggaaa... |
| 1498 | 6 | [atttga, tttgaa, ttgaat, tgaatg, gaatgg, aatgg... |
| 1499 | 6 | [agtatg, gtatgg, tatgga, atggaa, tggaaa, ggaaa... |

1500 rows × 2 columns

**Fig 11:** View of the Dataset after applying features.

*1) Multinomial Naïve Bayes*

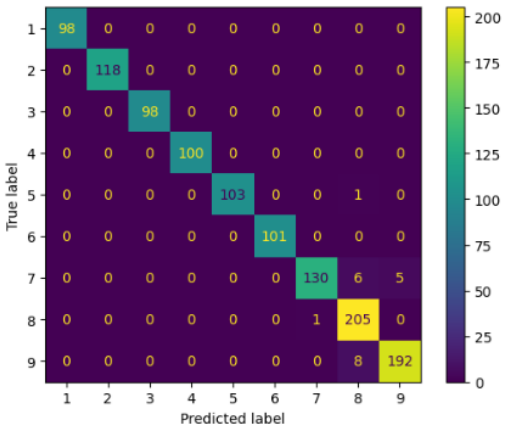
Multinomial Naive Bayes algorithm is well-suited for this type of problem because it can efficiently handle a large number of features (i.e., words or terms), and it makes the assumption that the features are conditionally independent given the class label, which simplifies the computation of the likelihood probabilities.

## IV. Results and Discussion

The results from various classification algorithms highlight the robust performance of the models used in the study for DNA virus classification. The Multinomial Naïve Bayes model achieved an impressive F1 score of 99%, indicating excellent classification accuracy across all classes. Similarly, the Support Vector Machine (SVM) exhibited comparable performance, as indicated in Fig 13, suggesting that both models are effective for this classification task. The overfitting analysis of the SVM, utilizing 10-fold cross-validation, demonstrated that the model was not over-fitted, with a training MSE of 0.0 and a test accuracy of 0.07.

Though typically used for regression tasks, the Random Forest Regressor showed promising results with an MSE of 0.0059 and an $R^2$ value of 0.98, further validating the model's capability. The initial overfitting concerns were addressed through stratified sampling, significantly enhancing model performance to an accuracy of 94%. Finally, the Logistic Regression model achieved a classification accuracy of 96%, underscoring its effectiveness in multi-class classification scenarios. These findings illustrate that appropriate model selection and data stratification techniques can markedly improve classification performance, thereby ensuring reliable and accurate results for DNA virus classification.

TABLE IV and Fig 13 are the combined results showing that the models are outperformed and that the model accuracy for each algorithm has reached a 99% correct classification. In feature engineering, we use K-mers with 6-character values for all algorithms, which can be changed. Since the results are already out perform so there is no need to change the strategy for feature engineering. In Table 4, all models are shown in one place. The analysis shows that all models outperform the SVM in more costing algorithms than others, and the fastest learner is multinomial. Naïve Bayes gives very fast results.




```
accuracy = 0.982
precision = 0.983
recall = 0.982
f1 = 0.982
```

**Fig 12:** Naïve Bayes Classification Result.

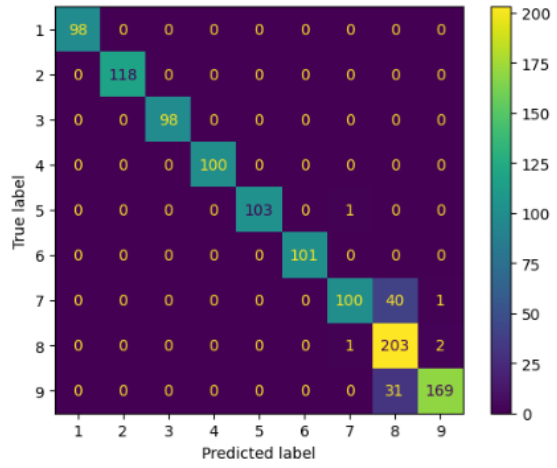



**Fig 13:** Support Vector Machine Results

TABLE III. RANDOM FOREST REGRESSOR PERFORMANCE.

| METRICS | VALUES |
|---|---|
| Mean Absolute Error (MAE): | 0.1003 |
| Mean Squared Error (MSE): | 0.1174 |
| Root Mean Squared Error (RMSE): | 0.3427 |
| Mean Absolute Percentage Error (MAPE): | 0.01478 |
| Explained Variance Score: | 0.9835 |
| Max Error: | 3.0000 |
| Mean Squared Log Error: | 0.00204 |
| Median Absolute Error: | 0.0 |
| $R^2$: | 0.9835 |
| Mean Poisson Deviance: | 0.01747 |
| Mean Gamma Deviance: | 0.0031 |

TABLE IV. RANDOM FOREST REGRESSOR PERFORMANCE.

| Algorithm | Model ACC |
|---|---|
| Multinomial Naïve Bayes | 98% |
| Support Vector Machine | 93% |
| Random Forest Regressor | 98% |
| Random Forest classifier | 94% |
| Gradient Boosting Regressor | 98% |
| Linear Regression | 96% |
| Voting Regression | 98% |
| Logistic Regression | 96% |

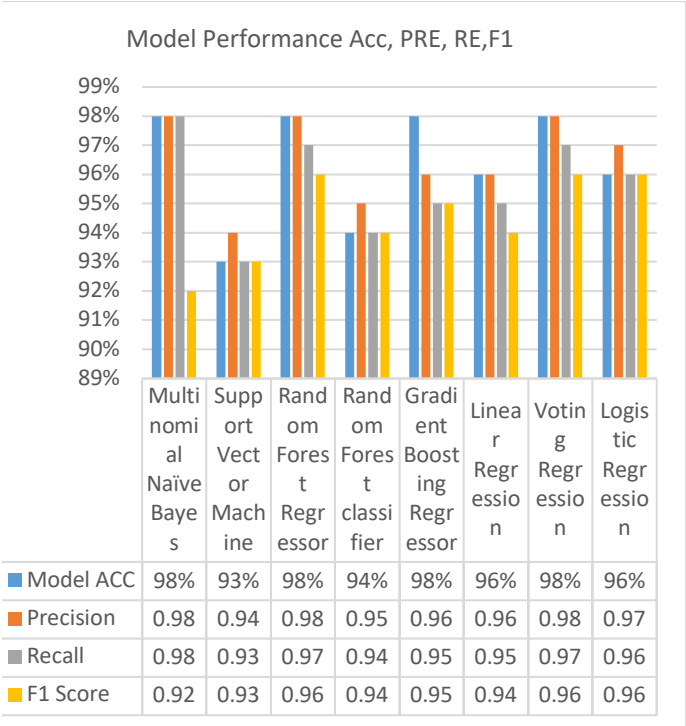



**Fig 13:** All Model Performance

The models are outperformed, and the results are perfect for classification. The dataset is small, but the qualitative features help us build the model. The Support vector machine and Multinomial Naïve Bayes outperform, but the SVM time stamping takes too much time for classification. Gradient Boosting regressor is also good, but in the classes of Class 4, which are DENGUE and INFLUENZA, some incorrect classifications cause a reduction in results.

## CONCLUSION

The data collection is hectic for the NCBI library's DNA/ RNA nucleotide sequences. There are a lot of tools, but it's too slow in Pakistan. The Downloading of the Dataset automatically stops, and we try again and again to compile the complete file. The preparation of the data is so simple using Bio Python that it helps a lot in managing the data. The feature engineering also makes it easy to manage and build the data language word by word, which is what we call K-MER.

DNA / RNA sequences can be used for virus prediction, and artificial intelligence helps us predict diseases using DNA/RNA nucleotide sequences. The proposed method helps us use the K-MER feature as a language word, and based on language patterns, we build the model for other diseases using DNA/ RNA sequences. The proposed study can be deployed in the research center and hospital to diagnose the diagnosis. The results of the study are so good, and almost all the algorithms give us good performance. The Timestamp is different for all algorithms, but the SVM take much time for classification. The overall results are good.

## FUNDING STATEMENT

The author(s) received no specific funding for this study.

## CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest to report regarding the present study.

## REFERENCES


[1] H. Gunasekaran, K. Ramalakshmi, A. Rex Macedo Arokiaraj, S. Deepa Kanmani, C. Venkatesan, and C. Suresh Gnana Dhas, "Analysis of DNA Sequence Classification Using CNN and Hybrid Models," Comput. Math. Methods Med., vol. 2021, pp. 1–12, Jul. 2021, doi: 10.1155/2021/1835056.

[2] G. Q. Lee et al., "HIV-1 DNA sequence diversity and evolution during acute subtype C infection," Nat. Commun., vol. 10, no. 1, pp. 1–11, 2019.

[3] I. Ahmed and G. Jeon, "Enabling artificial intelligence for genome sequence analysis of COVID-19 and alike viruses," Interdiscip. Sci. Comput. Life Sci., vol. 14, no. 2, pp. 504–519, 2022.

[4] T. Tu, M. A. Budzinska, F. W. Vondran, N. A. Shackel, and S. Urban, "Hepatitis B virus DNA integration occurs early in the viral life cycle in an in vitro infection model via sodium taurocholate cotransporting polypeptide-dependent uptake of enveloped virus particles," J. Virol., vol. 92, no. 11, pp. e02007-17, 2018.

[5] G. G. Carvalho et al., "Virulence and DNA sequence analysis of Cronobacter spp. isolated from infant cereals," Int. J. Food Microbiol., p. 109745, 2022.

[6] V. Agarwal, N. J. K. Reddy, and A. Anand, "Unsupervised Representation Learning of DNA Sequences." arXiv, Jun. 07, 2019. Accessed: Aug. 26, 2022. [Online]. Available: http://arxiv.org/abs/1906.03087

[7] S. Choudhury, B. Kashyap, and K. Dutta, "The ITS2 DNA sequence analysis in six species of barbin fishes with phylogenetic insights," *J. Appl. Biol. Biotechnol. Vol*, vol. 10, no. 01, pp. 62–67, 2022.

[8] B. L. Elizabeth, J. Gayathri, S. Subashini, and A. J. Prakash, "HIDE: hyperchaotic image encryption using DNA computing," *J. Real-Time Image Process.*, vol. 19, no. 2, pp. 429–443, 2022.

[9] N. Q. K. Le, E. K. Y. Yapp, Q.-T. Ho, N. Nagasundaram, Y.-Y. Ou, and H.-Y. Yeh, "iEnhancer-5Step: Identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding," *Anal. Biochem.*, vol. 571, pp. 53–61, Apr. 2019, doi: 10.1016/j.ab.2019.02.017.

[10] A. Alberdi *et al.*, "Promises and pitfalls of using high-throughput sequencing for diet analysis," *Mol. Ecol. Resour.*, vol. 19, no. 2, pp. 327–348, Mar. 2019, doi: 10.1111/1755-0998.12960.

[11] M. Silva, D. Pratas, and A. J. Pinho, "Efficient DNA sequence compression with neural networks," *GigaScience*, vol. 9, no. 11, p. giaa119, 2020.

[12] F. Pina-Martins and O. S. Paulo, "NCBI Mass Sequence Downloader–Large dataset downloading made easy," *SoftwareX*, vol. 5, pp. 80–83, 2016, doi: 10.1016/j.softx.2016.04.007.

[13] R. Chikhi, J. Holub, and P. Medvedev, "Data Structures to Represent a Set of *k* -long DNA Sequences," *ACM Comput. Surv.*, vol. 54, no. 1, pp. 1–22, Jan. 2022, doi: 10.1145/3445967.