CHAPTER-1

INTRODUCTION

The extraction of useful information from deoxyribonucleic acid (DNA) is a major component of bioinformatics research, and DNA sequence categorizations has a variety of applications, including genomic and biomedical data processing. DNA stands for deoxyribonucleic acid and is a hereditary molecule found in the cells of all humans and other living species, it contains the necessary information that determines our bodies 'basic characteristics and functions as a generic blueprint for an evolving organism. A character string consisting just of A, C, G or T can be used to represent an isolated DNA sequence. DNA is a double-standard molecule that is containing a phosphate group, a sugar group, and nitrogenous bases---thymine (T), guanine (G), cytosine (C), and adenine (A), and it is responsible for carrying and transmitting hereditary materials from parents to offspring. The base pairing is as follows: guanine pairs with cytosine (C-G) and thymine pairs with adenine (A-T). FASTA is the name of the format. DNA analysis is critical since it enables doctors to identity diseases, aids in the investigation of the spread of new infections, and can be used to solve crimes or conduct paternity testing. As a result, DNA analysis is now a hot topic in computational biology.

Primers are fundamental tools for DNA analysis in conventional biology. Primers are human's nucleotide sequences that are essential for all living organisms DNA synthesis to begin. Synthetic primers are used in molecular biology for a variety of applications, including the detection of viruses, bacteria, and parasites. These goals are served by primers, which are frequently found in human DNA sequences infected by a certain type of virus. The DNA fragment of an existing virus is amplified greatly using the polymerization chain reaction (PCR) technology, allowing researchers to detect the virus.

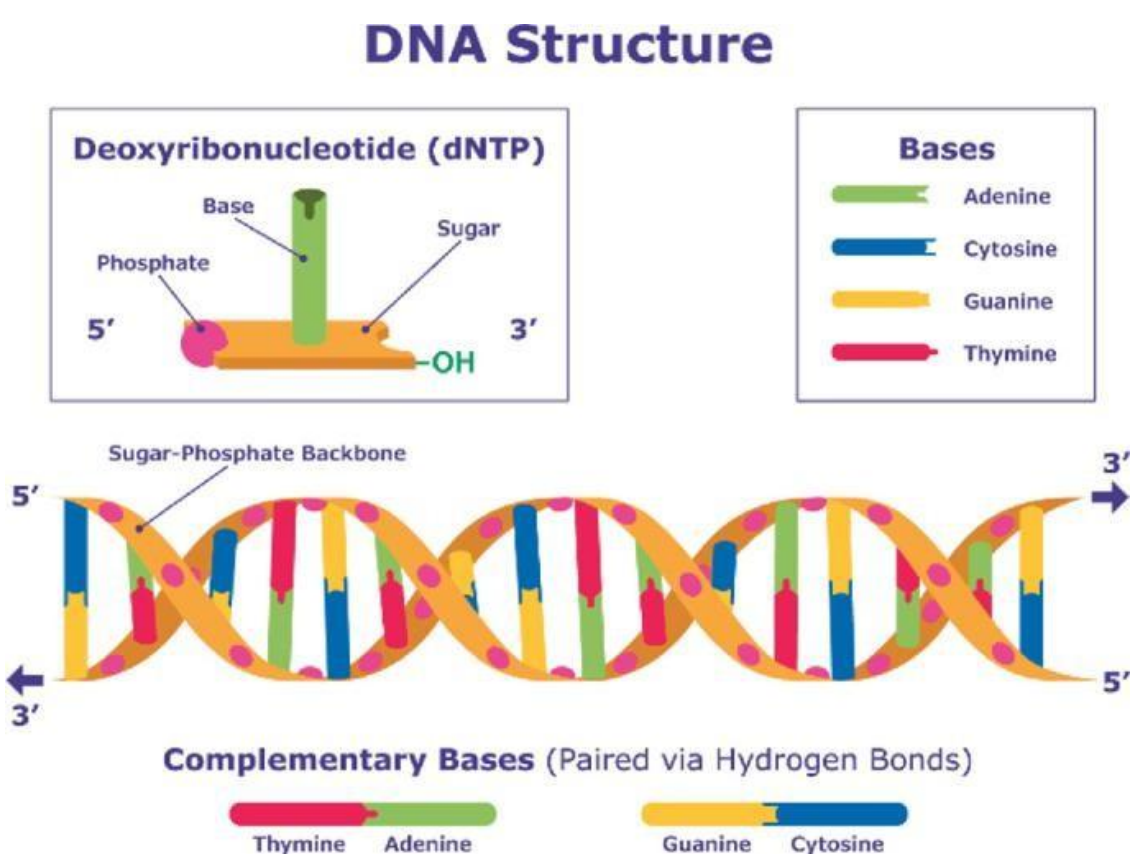

Figure 1.1 DNA sequencing

1.1 Deoxyribonucleic Acid (DNA) Sequencing

Biochemical similarities are frequently utilized to identity creatures that are closely related. The draft of one's life is frequently carried via DNA. The DNA sequence identifies the traits and types of species in simple words. At regular intervals throughout the cell, DNA include the instructions for making proteins.

DNA sequencing is a technique for ensuring that nucleotides in a DNA segment are arranged correctly. A double stand of nucleotide can occasionally be seen in DNA. For DNA sequencing a variety of approaches have been developed. The idea of utilizing DNA sequences to identify species is being investigated in a variety of fields. Many ways are recommended for detecting species using DNA sequence, including correspondence scores, phylogeny, population heritable information, and the discovery of species- specific sequence patterns. Shotgun cloning and walking are the two processes that DNA sequencing investigation usually take. Classification is a method of identifying a single character or a group of characters. To classify a supermolecule sequence into its specific division, secondary category, or family a variety of classification techniques are used. These strategies try to eliminate a few alternatives, equalize the values of those option, and finally categorize the super molecule sequence.

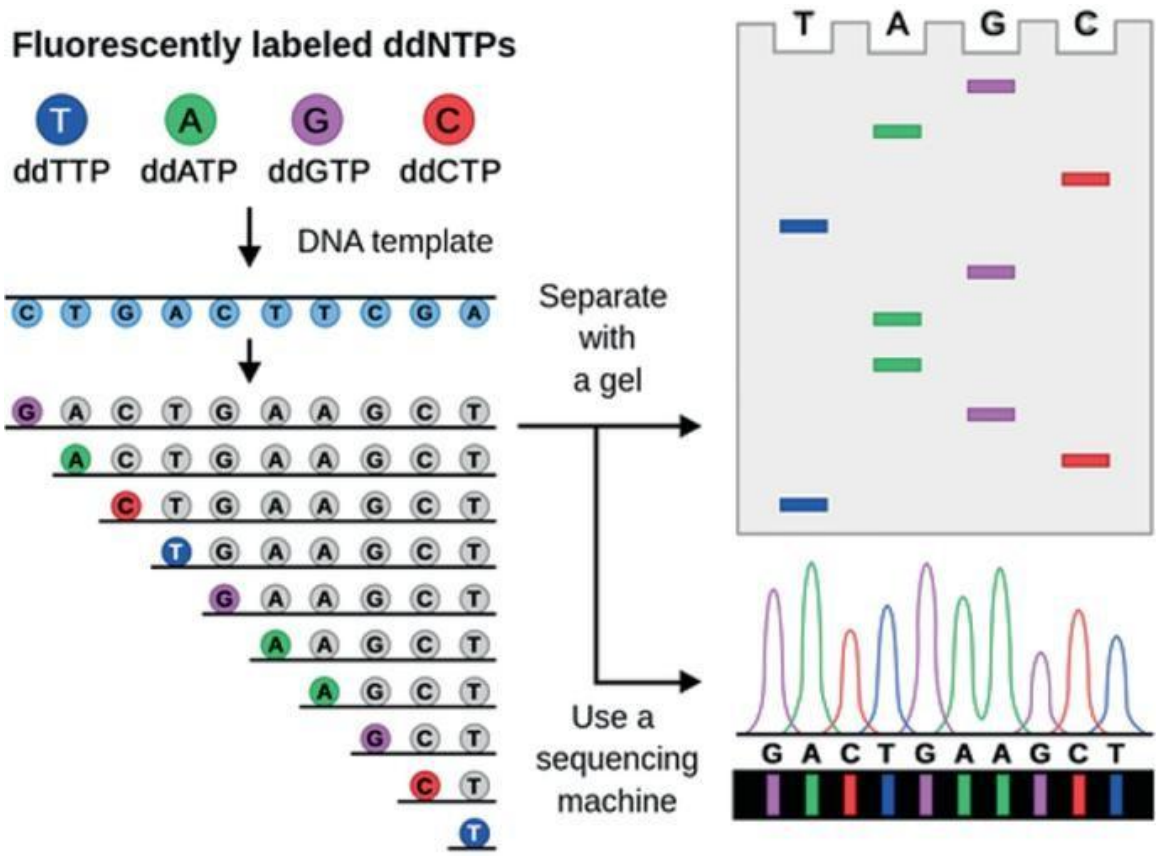

Figure1.2 Automated DNA sequencing

1.2 Deoxyribonucleic Acid (DNA) Sequencing Classification

The study of DNA sequence data is a major focus of bioinformatics. When we talk about DNA sequencing, we are talking about the process of determining the order of nucleotides in a nucleic acid sequence. The term categorization refers to the division of a nucleic acid or its combinations, referred to as a gene, into distinct regions. The selected- effect function genes are categorized into two primary groups in, which are functional DNA and rubbish DNA. Rubbish DNA and indifferent DNA are two types of functional DNA. Rubbish DNA does not have the unselected-effect function, but functional DNA does. Literal DNA and indifferent DNA, which merely the presence or lack of the sequence of indifferent DNA is selected in indifferent and garbage DNA. Junk DNA contributes to the fitness of the organisms, and as a result, it evolves under selection neutrality. Garbage DNA lowers the fitness of those who carry it. DNA in the above categories can be translated, or transcribed but not translated. The assignment of a DNA region to a specific functional category can vary throughout time. Functional DNA, for example, can be transformed into junk DNA, junk DNA into garbage DNA and so on.

1.3 Machine Learning in Bioinformatics: DNA Sequencing Classification

Many qualities of computational approaches , such as adaptability and fault tolerance, have made them appealing for bioinformatics research. For network classification, a machine learning approach is presented. Machine learnings goal is to explore, learn, and adapt to changing circumstance to improve the machines performance. The reference input is utilized for machine learning algorithms in the field of bioinformatics so that they can "learn." Soft computing techniques are appealing to use in bioinformatics because of their ability to deal with unclear and partially true data. Here, machine learning techniques can be utilized to train the network for improved performance and system accuracy. Furthermore, machine learning methods are utilized to reduce the number of false positive.

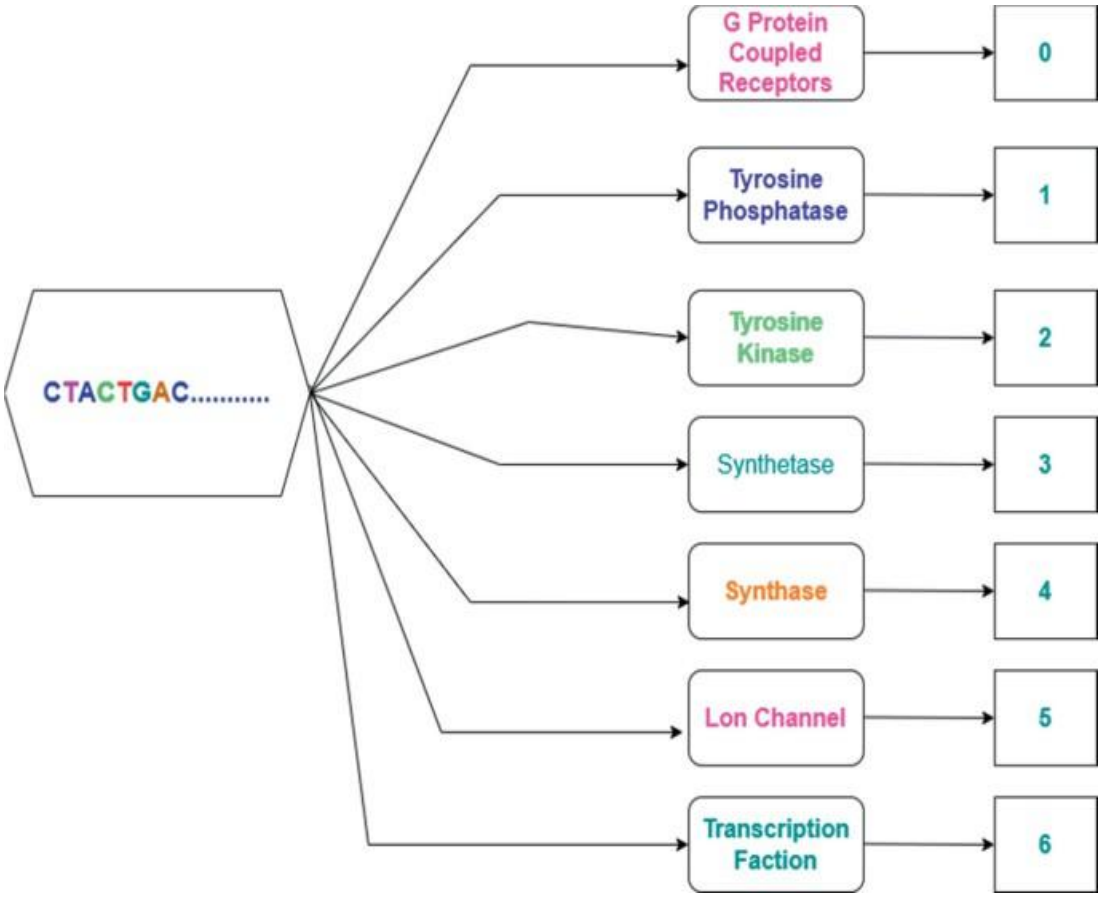

Figure 1.3 DNA sequencing type and respected class

CHAPTER-2

LITERATURE REVIEW

Individuals worked to identify some new items to apply after assessing thew importance of this area. They wanted to complete the job more quickly and accurately. Other researchers have some substantiated and fascinating contributions to this study field in the past.

Liangyou Chen and Lois Boggess used the genomic signature to research gene classification. They investigated this issue using neural network technologies and its four back-propagation algorithm methods, radial-basis functions, review panel machine, and self-organizing maps. After the experiment, the committee machine produced the best results, with an average inaccuracy of 16.88%.

Dr. P. Kiran Sree, Dr. P. S. V. Srinivasa Rao, and S. S. S. N. Usha Devi N proposed a novel concept for gene prediction and a new metagenomics classifier using a machine learning technique of transfer learning in their research [10]. To solve some bioin-formatics problems, it combined hybrid cellular automata with a deep educational environment. With this classification, they were able to reach 98.7% accuracy in just 8 ns. According to Vrinda V. Nair, Karthika Vijayan, Deepa P. Gopinath, and Achuthsankar S. Nair's research, they employed ANN and chaos game represen- tation to classify unknown genomic segments [11]. To test this proposed strategy. eight subsets from the taxonomic classification distribution of eukaryotic organisms were gathered. Qicheng Ma. Jason T L. Wang. Dennis Shasha, and Cathy H. Wu used a neural network and an expectation-maximization technique to identify DNA sequences [12] They introduced a novel method for classifying DNA sequences, to detect E. coli promoters in the bacteria's DNA. Determine whether a given DNA sequence is an E. coli promoter or not. The problem described above is known as a binary classification problem Jun Miyake, Yuhei Kaneshita, Takashi Hirano, Satoshi Asatani, Serichi Tagawa, and Hirohiko Niioka used deep learning to investigate a new approach for classifying DNA sequences [13] The DNA of human leukocyte antigen alleles was employed in this categorization, which was done as a graphical classifica-tion. Data is gathered from the Database of Immune Polymorphism and compressed into a two-dimensional format

Vrinda V. Nair examined a new approach for organism categorization based on a combination of FCGR and ANN [11], which was created by Karthika Vijayan, Deepa P Gopinath, Achuthsankar S. Nair. They used the genetic sequence to solve the difficulty of categorizing organisms into distinct groups. For this, they used a few different species. For PNN-based classification, they achieved an accuracy of 86.8%. Jonathan Auerbach, Damian Gola, Elizabeth Held, Emily R. Holzinger, Marc-Andre Legault, Rui Sun, Nathan Tintle, Hsin Chou Yang, and Inke R. König used machine learning and data mining to solve the problem of complex genomic data classification [14] They have metagenome issues as well as a combination of several data structures. Another study looked into using discriminative k-mers to classify metagenomics and genomic sequences.

CHAPTER-3

METHODOLOGY

The research is divided into two phases: preprocessing and post-processing. The approach in the preprocessing phase focuses on data preprocessing stages, whereas the workflow in the article phase can be broken down into two subparts: model learning and framework evaluation. Figure 4 depicts a representation of a research study. The working follow of the study is related to machine learning (ML) and natural language processing (NLP). The NLP is used for processing the texting data and converting the data into a string then numerical values to fit the machine learning model. The architecture is following some steps that are discussed below.

A. Data Collection

We have collected this dataset from the Kaggle repository. This dataset is avail- able as public. The name of this dataset is "human_data.txt". The dataset contain two features one is deoxyribonucleic acid (DNA)sequencing and another one is class. The size of this dataset is (4380,2) where 4380 is the number of samples and 2 is the number of columns.

Table 1 is showing about the gene family that is class and we have six classes in this dataset. We have also the number of occurrences per class as well the numeric values of the gene family class. This means we have converted the string class to anumeric class from 0 to 6 (7 class).we have sketched one count plot to plot the occurrences per class have. The transcription factor (class 6) has the most data among all 7 classes and the number is 1343. The lowest class is the Lon channel which has 240 samples.

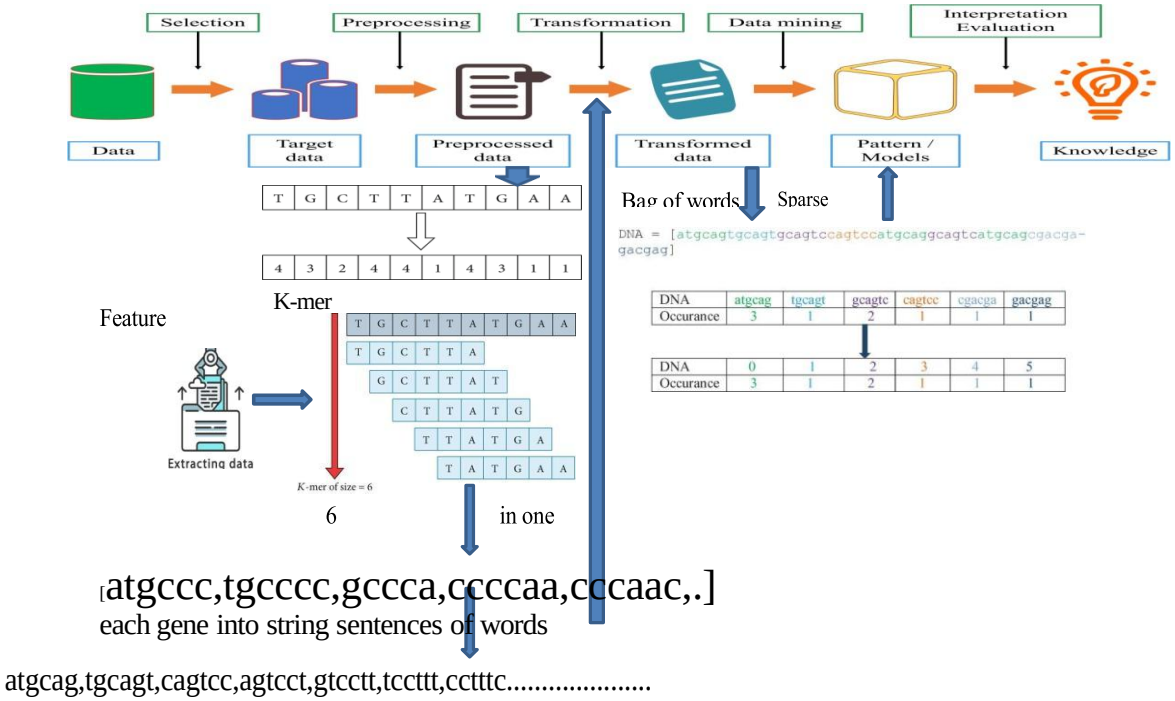

Figure 3.1 System Architecture

	class	words
0	4	[atgccc, tgcccc, gcccca, ccccaa, cccaac, ccaac...
1	4	[atgaac, tgaacg, gaacga, aacgaa, acgaaa, cgaaa...
2	3	[atgtgt, tgtgtg, gtgtgg, tgtggc, gtggca, tggca...
3	3	[atgtgt, tgtgtg, gtgtgg, tgtggc, gtggca, tggca...
4	3	[atgcaa, tgcaac, gcaaca, caacag, aacagc, acagc...

Figure 3.2 Sample of Dataset

Gene family	Number	Class label
G protein coupled receptors	531	0
Tyrosine kinase	534	1
Tyrosine phosphatase	349	2
Synthetase	672	3
Synthase	711	4
Ion channel	240	5
Transcription factor	1343	6

Table 1 Gene family, number of occurrences and their class label

B. Data Preprocessing:

Data preprocessing is the procedure for preparing raw data for use in a machine learning algorithm. It is the first and most important stage in building training data. In data preprocessing, we have to do the different tasks for instance getting the dataset, importing libraries, importing datasets, finding missing data, encoding categorical data, splitting the dataset into training and test sets, and feature scaling. Due to the growing volume of huge datasets, datasets frequently contain missing or ambiguity. The extraction of information will be severely hampered by poor data quality.

C Feature Extraction

After preprocessing the data, the next step is feature extraction from the processed data. Feature extraction is the most important task for the machine learning model. In feature learning extraction, we have to do some extra work on data to train the model. The model is receiving the features as input and produces the accepted output. For extracting the feature, we have so many tech- niques for the different datasets. In our case, we are using the k-mer counting algorithm for extracting the feature. From DNA sequences to amino acids, the k-mer technique simulates the process. A three-dimensional window is utilized to explore the entire DNA sequence, with a sliding unit of one at each step. The group of three nucleobases from the DNA sequence is obtained each time, and the associated amino acid is recorded. Stop codons are often overlooked All different types of amino acids are counted after the entire DNA sequence has been traversed. After that, each amino acid's proportion is determined and plotted in a histogram Consider the same DNA sequence as before.

The DNA sequence TTTGACTCGT contains eight codons TTT TTG," "TGA." are the acronyms for "TTT," "TTG," "TGA." "GAC." "ACT." "CTC." "TCG," and "CGT."

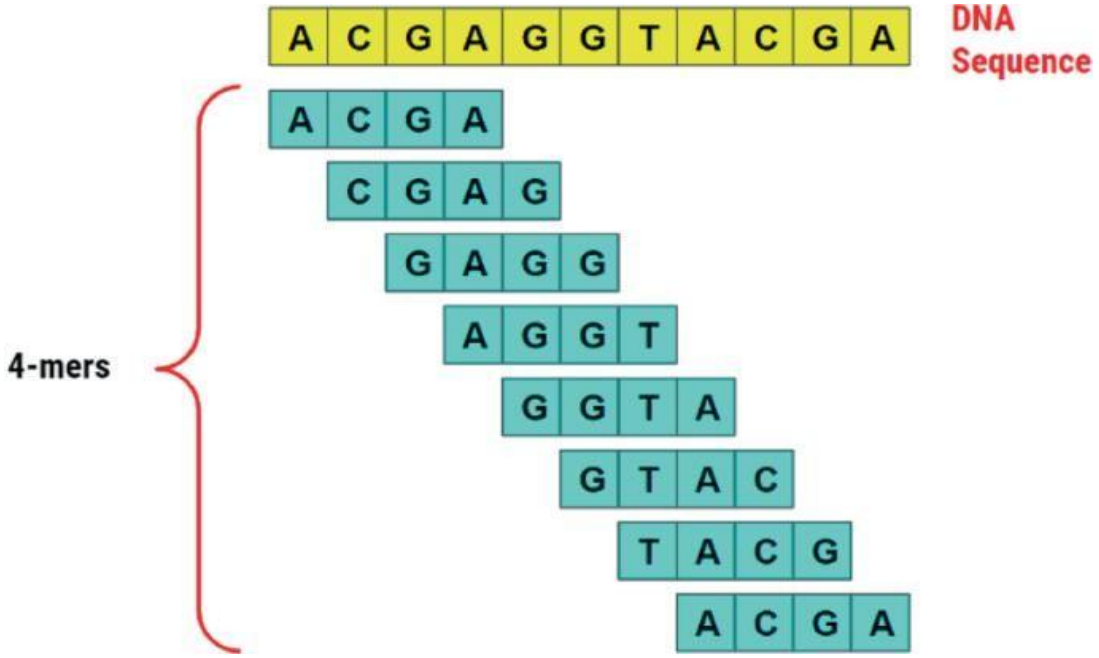

Figure 3.3 4-mers in the sequences

Algorithm 1: k-mers counting algorithm

// Start with the Empty Dictionary

1. counts Empty Dictionary for Storing the Unique k-mers values //Calculate how many kmers of length k there are
2. num kmerlen(DNA)-length_of_kmers+1
3. foriin num_kemers do
4. kmer DNA[i:i+ky/ read i to i+k //Add the kmer to the dictionary if it is not there
5. if kmer not in countsdo
6. counts[kmer] < 0 7. counts[kmer] += 1// counts value increment by 1

// Return the final Values 8. return counts

D. Transformation:

Data transformation is an important task for machine learning. For transforming the data, we used natural language processing (NLP) technique that is the Count Vectorizer method. The count is a Scikit-learn module. The vectorizer programmed converts a corpus of text into a vector of term/token frequencies. It also allows you to

preprocess your text data before producing the word vectors, making it a very versatile text feature extraction module. To show how Count Vectorizer is working, let's have an example that can be converted to a sparse matrix.

```
DNA = [atgcagtgcagtgcagtccagtccatgcaggcagtcatgcagcgacga-
gacgag]
```

DNA	atgcag	tgcagt	gcagtc	cagtcc	cgacga	gacgag
Occurance	3	1	2	1	1	1

DNA	0	1	2	3	4	5
Occurance	3	1	2	1	1	1

Figure 3.4 Example of sparse matrix.

E. Machine Learning Model:

(i) Random Forest:

It is necessary to understand how a decision tree classifier [15] operates before discussing the random forest technique. A decision tree is a tree-like construct that mimics human decision-making. Each node has a judgment, and the data is divided into separate child nodes. The final findings are displayed in the leaf nodes. The decision is evaluated using the impurity drop, and a good query should maximize the impurity drop. The decision tree is also a supervised learning model in which the model learns how to make queries and split data until specified criteria or threshold is reached using the training set. The random forest algorithm's underlying model is the decision tree.

A forest is made up of many trees, as its name suggests. The random forest model uses data to train several decision trees, with the average output of these trees being used as the final result. To construct different training datasets, the bootstrap aggregating approach is employed. This method takes some data from the original training dataset and replaces it with new data, which is then used to train a single tree. The random forest is a straightforward, easy-to-understand method that can handle difficult nonlinear classification problems. In our approach, two hyperparameters are required to be fine-tuned. The number of estimators is one of them. It determines the number of trees that should be planted during the trial. The other factor is each tree under certain. This number should be neither too high nor too

low Greater groups may result in the classifier, while smaller depths may result in parameters.

(ii) Logistic Regression:

A linear model used to do binary classification is known as logistic regres- sion. The output unit, like the MLP model, was calculated using the sigmoid function. To avoid any overfitting, L2 regularization was also applied. The regularization term is added after the loss function in this method, as shown in formula (3), L2 regularization is the word for the additional term. The L2 regularization [17] degree is controlled by the hyperparameter C in this phrase. Greater regularization is indicated by smaller values.

$$\text{Loss} \leftarrow \text{Loss} + 1/C \sum_{i=0}^{n} \omega^2_i$$

The data points in the input X are usually considered to have a nonlinear model. The majority of the time, however, this assumption is incorrect. In such cases, logistic regression is a credible alternative model.

(i) Support Vector Machine:

Another linear model for classification is the support vector machine (SVM) [18, 19]. An SVM has great generalization ability since it can handle a little amount of data and is less sensitive to noise in a dataset [20]. The goal of the SVM is to discover the hyperplane that maximizes the difference between the two classes. The Lagrange multiplier approach can be used to find the solution. If kernel functions are employed correctly, powerful nonlinear SVM models can be trained. Kernel functions generate new feature vectors, which are typically larger than the original input. In the new feature space, the SVM discovers the new hyperplane, which is linear.

CHAPTER-4

RESULTS AND DISCUSSION

The results and discussion part consist of some confusion matrix, learning curve, comparison graph, and more other information. We have sued, six machine learning algorithms to classification DNA requesting. The approach we are using is nothing but a natural language processing technique. The k-mer technique and bag-of-words are used for preprocessing and transformation the data. For evaluated the result, we are used the confusion matrix, precision, recall, and F1-score. The evaluated matrix has four major to calculated accuracy for instance true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Below the accuracy, precision, recall, and F1-score equations are written according to the confusion matrix.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

$$Precision = \frac{Tp}{tp + fp}$$

$$Recall = \frac{Tp}{tp + tn}$$

$$F1 - score = \frac{precision.\, recall}{precision + recall}$$

Table 2 displays the accuracy of a single set of training and test sets, the mean accuracy of a tenfold training and test, and the sample variance in the tenfold process of selecting test cases.

As per the results in Table 2, the multinomial Naïve Bayes algorithm not only performs the best with this data onefold but also has the least variance while completing k-fold training and test data selection and subsequent training.

Other quantitative metrics comprehensive, like accuracy, recall, and F1-scores,are shown in Table 3 for all the algorithms used in both courses. The multinomial Naïve Bayes approach outperforms the others with a precision of 0.99%.

Table 2 Comparison between 10-folder accuracy vs without cross-validation accuracy

Machine learning algorithm	K-fold Ac.(%)	Accuracy(%)
Decision tree	84.5	80.93
Support vector machine	87	80.5993
Logistic regression	92.10	94
K-nearest neighbors	82.54	76.14
Random forest	90	91.55
Multinomial Naïve Byes	97.70	98.40

Table 3 Performance measure like a precision, b recall, c F1-score

(a) Precision

Classes	Logistic Regression	KNN	Decision Tree	Multinomial Naïve Bayes	Random forest	SVM
0	0.99	0.33	0.55	0.98	0.88	0.97
1	1.00	1.00	0.85	1.00	1.00	1.00
2	1.00	1.00	0.84	1.00	1.00	1.00
3	0.97	1.00	0.83	0.99	0.79	1.00
4	1.00	1.00	0.85	0.99	0.97	1.00
5	1.00	1.00	0.84	1.00	1.00	1.00
6	0.84	0.98	0.89	0.96	0.90	0.61
Average:0.97	0.90	0.81	0.99	0.94	0.94	

(b) Recall

Classes	Logistic Regression	KNN	Decision Tree	Multinomial Naïve Bayes	Random forest	SVM
0	0.90	1.00	0.78	0.97	0.89	0.70
1	0.93	0.72	0.80	0.98	0.91	0.69
2	0.94	0.78	0.83	1.00	0.91	0.77
3	0.93	0.85	0.81	0.99	0.93	0.74
4	0.90	0.65	0.81	0.96	0.88	0.76
5	0.86	0.67	0.73	1.00	0.84	0.61
6	1.00	0.72	0.83	0.99	0.96	1.00
Average:0.92	0.77	0.80	0.99	0.90	0.75	

(c) F1-score

Classes	Logistic Regression	KNN	Decision Tree	Multinomial Naïve Bayes	Random forest	SVM
0	0.94	0.50	0.65	0.98	0.89	0.81
1	0.97	0.84	0.83	0.99	0.95	0.82
2	0.97	0.88	0.84	1.00	0.95	0.87
3	0.95	0.92	0.82	0.99	0.86	0.85
4	0.95	0.79	0.83	0.98	0.92	0.86
5	0.93	0.80	0.78	1.00	0.91	0.76
6	0.92	0.83	0.86	0.98	0.93	0.76
Average:0.95	0.79	0.80	0.99	0.92	0.82	

CHAPTER-5

APPLICATION AND CHALLENGES

5.1 Applications of machine learning in bioinformatics

1. Facilitating gene editing experiments

Gene editing refers to manipulations on an organism's genetic composition by deleting, inserting, and replacing a part of its DNA sequence. This process typically relies on the CRISPR technique, which is rather effective. But there is still much improvement to be desired in the area of selecting the right DNA sequence for manipulation, and this is where ML can help. Using machine learning for bioinformatics, researchers can enhance the design of gene editing experiments and predict their outcomes.

A research team employed ML algorithms to discover the most optimal combinational variants of amino-acid residues that allow genome-editing protein Cas9 to bind with the target DNA. Due to the large number of these variants, such an experiment would have been too large, but using an ML-driven engineering approach reduced the screening burden by around 95%.

2. Identifying protein structure

Proteomics is a study of proteins, their interactions, composition, and their role in the human body. This field involves heavy biological datasets and is computationally expensive. Therefore, technologies like machine learning in bioinformatics are essential here.

One of the most successful applications in this field is using convolutional neural networks to position proteins' amino acids into three classes — sheet, helix, and coil. Neural networks can achieve an accuracy of 84% with the theoretical limit being 88%– 90%.

Another usage of ML in proteomics is protein model scoring, a task essential to predict protein structure. In their machine learning approach to bioinformatics, researchers from the Fayetteville State University deployed ML to improve protein model scoring. They divided protein models under question into groups and used an ML interpreter to decide

on the feature vector to evaluate models belonging to each group. These feature vectors were used later to further improve the ML algorithms while training them on each group separately.

3. Spotting genes associated with diseases

Researchers increasingly use machine learning in bioinformatics to identify genes that are likely to be involved in particular diseases. This is achieved by analyzing gene expression microarrays and RNA sequencing.

Particularly, gene identification gains traction in cancer-related studies to identify genes that are likely to contribute to cancer, as well as classify tumors by analyzing them on a molecular level.

For instance, a group of scientists at the University of Washington used several machine learning in bioinformatics algorithms, including decision tree, support vector machine, and neural networks to test their ability to predict and classify cancer types. Researchers deployed RNA sequencing data from The Cancer Genome Atlas project, and discovered that linear support vector machine was the most precise, hitting the 95.8% accuracy in cancer classification.

In another example, researchers used ML to classify breast cancer types based on gene expression data. This team also relied on the Cancer Genome Atlas project's data. The researchers classified the samples into triple negative breast cancer — one of the most lethal breast cancers — and non-triple negative. And once again, the support vector machine classifier delivered the best results.

Speaking of non-cancerous diseases, researchers at the University of Pennsylvania relied on machine learning to identify genes that would be a suitable target for coronary artery disease (CAD) drugs. The team used the ML-powered Tree-based Pipeline Optimization Tool (TPOT) to pinpoint a combination of single nucleotide polymorphisms (SNPs) related to CAD. They analyzed the genomic data from the UK Biobank and uncovered 28 relevant SNPs. The relation between the SNPs on top of this list and CAD was previously mentioned in the literature, and this research gave a practical validation.

4. Traversing the knowledge base in search of meaningful patterns

Advanced sequencing technology doubles genomic databases each 2.5 years, and researchers are looking for a way to extract useful insights from this accumulated knowledge. Machine learning in bioinformatics can sift through biomedical publications and reports to identify different genes and proteins and search for their functionality. It can also aid in annotating protein databases and complement them with the information it retrieves from the literature.

One example comes from a group of researchers who deployed bioinformatics and machine learning in literature mining to facilitate protein model scoring. Structural modeling of protein-protein dockings typically results in several models that are further scored based on structural constraints. The team used ML algorithms to traverse PubMed papers on protein-protein interactions, searching for residues that could help generate these constraints for model scoring. And to make sure that the constraints are relevant, scientists explored the ability of different machine learning algorithms to check all discovered residues for relevancy.

This research revealed that both computationally expensive neural networks and less resource demanding support vector machine achieved very similar results.

5. Repurposing drugs

Drug repurposing, or reprofiling, is a technique scientists use to discover new applications of existing drugs that they were not intended for. Researchers adopt AI in bioinformatics to perform drug analysis on relevant databases, such as BindingDB and DrugBank. There are three major directions for drug repurposing:

- Drug-target interaction looks into the drug's ability to bind directly to the target protein
- Drug-drug interaction investigates how medications act when they are taken in combinations
- Protein-protein interaction looks into the surface of interacting intracellular proteins, and attempts to discover hotspots and allosteric sites.

Researchers from the China University of Petroleum and the Shandong University developed a deep neural network algorithm and used it on the DrugBank database. They wanted to study drug-target interactions between drug molecules and the mitochondrial

fusion protein 2 (MFN2), which is one of the main proteins that can possibly cause Alzheimer's disease. The study identifies 15 drug molecules with binding potential. Upon further investigation, it appeared that 11 of them can successfully dock with MFN2. And five of them have medium to strong binding force.

5.2 Challenges presented by machine learning in bioinformatics

1. **Bioinformatics AI is expensive**.

For the algorithm to perform properly, you need to acquire a large training dataset. However, it's rather costly to obtain 10,000 chest scans, or any other type of medical data for that matter.

2. **Difficulties associated with the training datasets**.

In other fields, if you don't have enough training data, you can generate synthetic data to expand your dataset. However, this trick might not be appropriate when it comes to human organs. The problem is that your scan generation software might produce a scan of a real human. And if you start using that without the person's permission, you will be in gross violation of their privacy.

Another challenge associated with the training data is that if you want to build an algorithm that works with rare diseases, there will not be much data to work with in the first place.

3. **The confidence level must be very high**

When human life depends on the algorithm's performance, there is just too much at stake, which does not leave room for error.

4. **Explainability issue**.

Doctors will not be open to using the ML model if they don't understand how it produced its recommendations. You can use explainable AI instead, but these algorithms are not as powerful as some black-box unsupervised learning models.

CONCLUSION AND FUTURE SCOPE

The DNA sequence database is still growing. This is because many new viruses, bacteria, or genomics sequences have been discovered as a result of research, and this has contributed to many unsolved difficulties affecting the DNA sequence of organisms. Deep learning is also evolving. As a result, deep learning appears to be much more promising in the future. It is reasonable to expect the discovery of a new invention or discovery, such as a new hybrid approach combining two procedures. A more elegant future will emerge when deep learning is discovered and can be used to design and integrate a system that can evolve and adapt to any environmental or contextual variation.

The purpose of this study was to assess the suggested model's ability to classify DNA sequences. The results were also compared to the prior best performance to see if the proposed model had improved. Furthermore, this study looked at different lengths of sequence to see how well the proposed model performed. The overall results of this study are within the expected range.

REFERENCE

[1]. Gelfand, M.S.: Prediction of function in DNA sequence analysis. J. Compute. Biol. 2(1), 87-115 (1995)

[2]. Bukh, J., Purcell, R.H., Miller, R.H.: Importance of primer selection for the detection of hepatitis C virus RNA with the polymerase chain reaction assay. Proc. Natl. Acad. Sci. 89(1), 187-191 (1992)

[3]. Dorn-In, S., Bassitta, R., Schwaiger, K., Bauer, J., Hölzel, C.S.: Specific amplification of bacterial DNA by optimized so-called universal bacterial primers in samples rich in plant DNA. J. Microbiol. Methods 113, 50-56 (2015) ,

[4]. Pacheco, M.A., Cepeda, A.S., Bernotienė, R., Lotta, LA., Matta, N.E., Valkiūnas, G., Escalante AA Primers targeting mitochondrial genes of avian hacmosporidian: PCR detection and differential DNA amplification of parasites belonging to different genera. Int. J. Parasitol 48(8), 657-670 (2018)

[5]. Mridha, K. et al. Deep learning algorithms are used to automatically detection invasive ducal carcinoma in whole slide images. In: 2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA), pp. 123-129 (2021). https://doi.org/10.1109/ICC CAS2192.2021 9666302

[6]. Mridha, K., et al: Web based brain tumor detection using neural network In: 2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA), pp. 137-143 (2021). https://doi.org/10.1109/1CCCA52192 2021 9666248

[7]. Zheng, Y., Azevedo, R.B.R., Graur, D: An Evolutionary Classification of Genomic Function, vol 7, no. 3. p. 4 (2015)

[8]. Mridha, K. Pandey, A.P. Ranpariya, A. Ghosh, A., Shaw, R.N: Web-based brain tumor detection using neural network In: 2021 IEEE 6th International Conference on Computing. Communication and Automation (ICCCA), pp. 137-143 (2021)

[9]. Boggess, L. Chen, L: Neural networks for genome signature analysis. In: 9th International Conference on Neural Information Processing (ICONIP OZ)

[10]. Srinivasa Rao, P.S.V. Usha Devi NSS.SN. Kiran Sree, P. CDLGP a novel unsupervised classifier using deep learning for gene prediction In. IEEE International Conference on Power, Control, Signals, and Instrumentation Engineering (2017)

[11]. Vijayan, K., Gopinath, D.P., Nair, A.S., Nair, V.V. ANN-based classification of unknown genome fragments using chaos game representation. In: Second International Conference on Machine Learning and Computing (2010)

[12]. Wang, J.T.L., Shasha, D., Wu, C.H., Ma, Q: DNA sequence classification via an expectation-maximization algorithm and neural networks: a case study. In: IEEE Transactions on Systems,Man, and Cybernetics-Part C: Applications and Reviews (2001)

[13]. Sinha, T., et al: Analysis and prediction of COVID-19 confirmed cases using deep learning models: a comparative study In: Bianchini, M., Piuri, V., Das, S., Shaw, R.N. (eds) Advanced Computing and Intelligent Technologies. Lecture Notes in Networks and Systems, vol. 218. Springer, Singapore (2022). https://doi. org/10.1007/978-981-16-2164- 2 18