ANALYSIS OF DNA SEQUENCE CLASSIFICATION USING ML ALGORITHM

MINOR PROJECT-2 REPORT

Submitted by

PRADEEPA G

V ADITYA

VAISHNAVI K

Under the Guidance of

Dr. KOUSHICK VENKATESH

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

ELECTRONICS & COMMUNICATION ENGINEERING

Vel Tech
Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology
(Deemed to be University Estd. u/s 3 of UGC Act, 1956)

October 2024

Vel Tech
Rangarajan Dr. Sagunthala
R&D Institute of Science and Technology
(Deemed to be University Estd. u/s 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that this Minor project-2 report entitled **" ANALYSIS OF DNA SEQUENCE CLASSIFICATION USING ML ALGORITHM "** is the bonafide work of **PRADEEPA G (21UEEA0101), V ADITYA (21UEEA0230) and VAISHNAVI K (21UEEA0130)** who carried out the project work under my supervision.

SUPERVISOR	**HEAD OF THE DEPARTMENT**
Dr.KOUSHICK VENKATESH	**Dr.A. SELWIN MICH PRIYADHARSON**
Assistant Professor	Professor
Department of ECE	Department of ECE

Submitted for Minor project-2 work viva-voce examination held on:______________________

INTERNAL EXAMINER	**EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

We express our deepest gratitude to our Respected Founder President and Chancellor **Col. Prof. Dr. R. Rangarajan**, Foundress President **Dr. R. Sagunthala Rangarajan**, Chairperson and Managing Trustee and Vice President.

We are very thankful to our beloved Vice Chancellor **Prof. Dr. Rajat Gupta** for providing us with an environment to complete the work successfully.

We are obligated to our beloved Registrar **Dr. E. Kannan** for providing immense support in all our endeavours. We are thankful to our esteemed Dean Academics **Dr. Raju Shanmugam** for providing a wonderful environment to complete our work successfully.

We are extremely thankful and pay my gratitude to our Dean SoEC **Dr. R. S. Valarmathi** for her valuable guidance and support on completion of this project.

It is a great pleasure for us to acknowledge the assistance and contributions of our Head of the Department **Dr. A. Selwin Mich Priyadharson**, Professor for his useful suggestions, which helped us in completing the work in time and we thank him for being instrumental in the completion of third year with his encouragement and unwavering support during the entire course. We are extremely thankful and pay our gratitude to our Minor project -2 coordinator **Dr. Harikrishna Paik**, for her valuable guidance and support on completing this project report in a successful manner.

We are grateful to our supervisor **Dr. Koushick Venkatesh**, Associate Professor ECE for providing me the logistic support and his valuable suggestion to carry out our project work successfully.

We thank our department faculty, supporting staffs and our family and friends for encouraging and supporting us throughout the project.

PRADEEPA G

V ADITYA

VAISHNAVI K

TABLE OF CONTENTS

ABSTRACT

DNA sequencing is vital for genetic research, helping decode DNA strands to understand genetic variations, evolutionary relationships, and the genetic basis of diseases. In this project, we aim to improve DNA sequence prediction accuracy by comparing machine learning (ML) models. We use ML algorithms like Decision Trees, Random Forest, and Naive Bayes, as well as techniques like Convolutional Neural Networks (CNN) and Transform Learning. The goal is to find the most accurate prediction model.

The project begins by gathering a labeled DNA dataset and applying preprocessing techniques, such as data cleaning, normalization, and feature extraction, to improve model efficiency. The ML models are trained to classify DNA sequences based on this preprocessed data. Decision Trees create a tree-like structure of decision points, Random Forest builds multiple trees to reduce overfitting, and Naive Bayes uses a probabilistic approach to classify data. ML models, such as CNN and Transform Learning, are trained to capture complex patterns within the DNA sequences.

The models are evaluated using metrics like accuracy, precision, recall, and F1-score. Among ML models, Naive Bayes achieved the highest accuracy at 98, outperforming Decision Trees and Random Forest. Transform Learning delivered the best performance, achieving 94.57 accuracy, while CNN performed slightly lower, likely due to the dataset size.

The results indicate that Naive Bayes is the most effective model for DNA sequence prediction, outperforming both ML models. Although Transform Learning shows potential, Naive Bayes remains the most reliable for this task due to its simplicity and efficiency. Future work could involve combining ML techniques to further improve accuracy or expanding the dataset to optimize.Overall, this study demonstrates the strength of ML, particularly Naive Bayes, in DNA research.

Keywords: DNA sequence, machine learning, data mining, DNA sequence alignment, DNA sequence classification, DNA sequence clustering, DNA pattern mining.

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 CLASSIFICATIONS USING ML ALGORITHM

We live in the era of the genome, advances in science have allowed humans to spy on the mysteries of life. In recent decades, the rapid expansion of biological data is a significant feature of the development of molecular biology, and a massive biological information database has rapidly formed. We must obtain useful knowledge from these huge data, and simultaneously bioinformatics was born. Bioinformatics is an interdisciplinary subject. It comprehensively uses mathematics, life sciences, and computer science to mine biological information in biological data (Chu, 2014), and further guides the relevant researches of biological researchers. Specifically, the first step is to obtain information on the protein-coding region by analyzing the genomic DNA sequence. Then simulating and predicting the spatial structure of the protein. Finally, according to the function of the protein, the researchers make the necessary drug design.According to statistics, the amount of biological data approximately doubles every 18 months. In 1982, GenBank's first nucleic acid sequence database had only 606 sequences, containing 680,000 nucleotide bases (Bilofsky et al., 1986).In the figure 1.1 As of February 2013, its database already contains 162 million biological sequence data, containing 150 billion nucleotide bases.

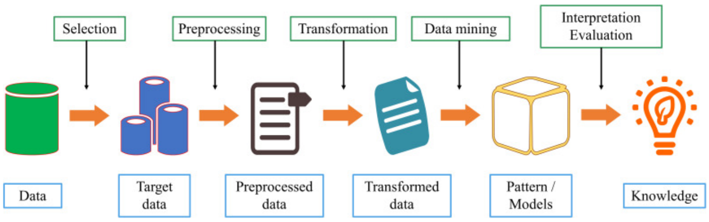


Figure 1.1: DNA Classification

How to mine knowledge from these huge data and guide biological research s an important research content of bioinformatics. For complex biological data, on the one hand, it is necessary to solve the problem of storage and management of massive data, and on the one hand, it is necessary to extract effective information from the data on the premise of ensuring that the data reflects the true meaning of biology. Machine learning is an important method to achieve artificial intelligence. It can handle the automatic learning of machines without explicit programming and has been widely used in the field of bioinformatics (Li et al., 2005; Larranaga et al., 2006).DNA is a kind of biomacromolecule in organisms. It carries the genetic information of life and guides the development of biological development and the functioning of life functions.

At present, machine learning has been widely used in sequence data analysis and has very broad application prospects in improving data processing capabilities and generating valuable biological information. The review focuses on DNA sequence data mining and machine learning. The review briefly introduces the development process of sequencing technology, DNA sequence data structure, and several sequence encoding methods in machine learning. And we clarify that sequence similarity is the basis of DNA sequence data mining. We have comprehensively analyzed the basic process of data mining and summarized the algorithms commonly used in machine learning. Then, we summarized four typical applications of machine learning in DNA sequence data DNA sequence alignment, classification, clustering, and pattern mining.

In summary, we have the following conclusions distributed sequence alignment and parallel computing may be the research focus of DNA sequence alignment. from figure 1.2 How to effectively express sequence features and analyze DNA sequence classification is a difficult point in research. The two key points of DNA sequence clustering are how to extract characteristic subsequences in the DNA sequence. DNA sequence pattern mining will generate an explosion of candidate sequence patterns, which will consume a lot of time and space. How to design a suitable search strategy and eliminate redundant sequence patterns will be an important direction for future research..

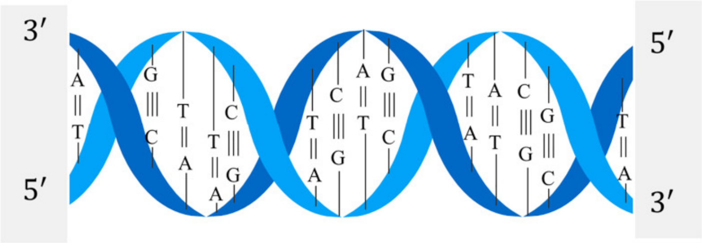
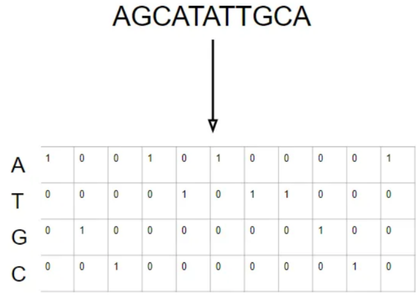


Figure 1.2: DNA Sequence

1.1.1 Machine Learning

In the past few decades, we have witnessed the revolutionary development of biomedical research and biotechnology and the explosive growth of biomedical data. The problem has changed from the accumulation of biomedical data to how to mine useful knowledge from the data. On the one hand, the rapid development of biotechnology and biological data analysis methods has led to the emergence of a challenging new field: bioinformatics. On the other hand, the continuous development of biological data mining technology has produced a large number of effective and well-scalable algorithms. How to build a bridge between the two fields of machine learning and bioinformatics to successfully analyze biomedical data is worthy of attention and research. from figure 1.3 particular, we should analyze how to use data mining for effective biomedical data analysis, and outline some research questions that may stimulate the further development of powerful biological machine learning algorithms.



A	1	0	0	1	0	1	0	0	0	0	1
T	0	0	0	0	1	0	1	1	0	0	0
G	0	1	0	0	0	0	0	0	1	0	0
C	0	0	1	0	0	0	0	0	0	1	0

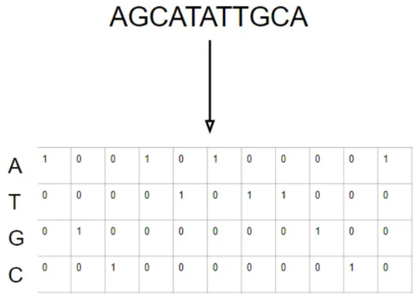
Figure 1.3: Machine Learning data set

DNA is a massive molecule that has a double helix structure made up of 4 basic nucleotides.A which pairs up with T and C which pairs up with G. How nucleotides are put together decides everything from the color of your eyes to why your fingers have different lengths. But how in the world does a computer understand these letters to figure out patterns or associations.This is how traditional data (which computers can understand easily) is different from DNA. DNA data is made up of characters, not numeric values (A, T, C, G) The sizes of DNA sequences are almost random. Some sequences are very short, while others are extremely long DNA data has biological significance, therefore some ML techniques may not work out well.

1.1.2 One-Hot Encoding

Each categorical value is converted into a new categorical column and given a binary value of 1 or 0. This is well suited for CNNs and machine learning algorithms in general.One-hot encoding is particularly useful because machine learning models expect numerical input, and this method converts categorical data into a format that is easy for models to process. By avoiding any implicit order between categories, it prevents the model from making incorrect assumptions about the relationship between different categorical values.

From Figure 1.4 a classification task, if a categorical variable represents different animal species



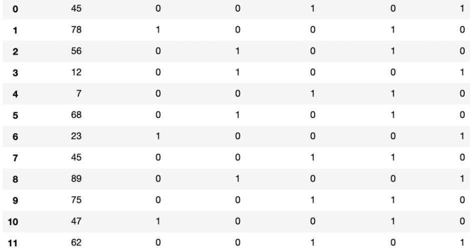
A	1	0	0	1	0	1	0	0	0	0	1
T	0	0	0	0	1	0	1	1	0	0	0
G	0	1	0	0	0	0	0	0	1	0	0
C	0	0	1	0	0	0	0	0	0	1	0

Figure 1.4: One-hot encoding

(e.g., "cat," "dog," "bird"), one-hot encoding will create separate binary columns for each species, making the data more interpretable for the model.CNNs, in particular, can handle large one-hot encoded vectors effectively because of their ability to work with high-dimensional data. This makes one-hot encoding a popular choice for tasks like image classification or natural language processing when categorical features are involved.Although it's widely used, one-hot encoding may result in sparse matrices (many zeros) when there are numerous categories. In such cases, alternatives like embedding layers, which map categories to dense vector representations, might be preferred in machine learning.

1.1.3 Sequential Encoding

In machine learning, sequential encoding can also negatively impact algorithms like support vector machines (SVMs) or k-nearest neighbors (KNN), where the distance between feature values is important. Since sequential encoding introduces artificial ordinal relationships, it may lead to incorrect distance calculations, resulting in poor model performance. Furthermore, models like logistic regression and linear regression, which assume linear relationships, will misinterpret the encoded categories, leading to biased or skewed predictions. This can cause the model to learn incorrectly, giving more importance to certain categories simply because their numeric values are larger or closer together.



	Emploee_ID	Remarks_Good	Remarks_Great	Remarks_Nice	Gender_Female	Gender_Male
0	45	0	0	1	0	1
1	78	1	0	0	1	0
2	56	0	1	0	1	0
3	12	0	1	0	0	1
4	7	0	0	1	1	0
5	68	0	1	0	1	0
6	23	1	0	0	0	1
7	45	0	0	1	1	0
8	89	0	1	0	0	1
9	75	0	0	1	1	0
10	47	1	0	0	1	0
11	62	0	0	1	0	1

Figure 1.5: Sequential Encoding

From Figure 1.5 machine learnig machine models that rely on gradient descent, this problem becomes even more pronounced, as the model continuously adjusts weights based on the incorrect assumption that categories are ordered. This can lead to slow convergence or convergence toward sub optimal solutions.Therefore, while sequential encoding is simple, it's typically avoided for algorithms using gradient descent or any model sensitive to relationships between numeric values. One-hot encoding or embedding layers are better alternatives that don't impose unintended order on categorical data.

1.1.4 K-mer encoding

K-mer encoding is a technique used in bio informatics to process and analyze DNA or protein sequences. In this method, a long sequence is broken down into overlapping sub sequences of a fixed length, known as k-mers. From Figure 1.6, if we have a DNA sequence "ATCG," its 2-mers would be "AT," "TC," and "CG." By using k-mers, we can transform complex biological sequences into smaller, more manageable units for machine learning models.K-mers are particularly valuable in representing genomic data, allowing for the extraction of patterns or motifs that are critical for understanding biological functions, such as gene expression or mutation detection. This method helps capture the local sequence structure, which is important in analyzing DNA sequences.

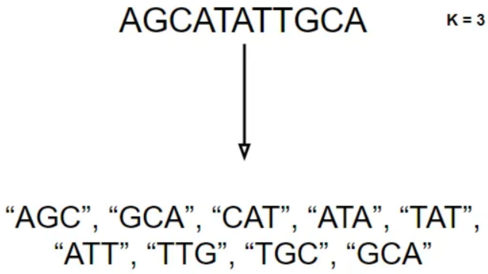
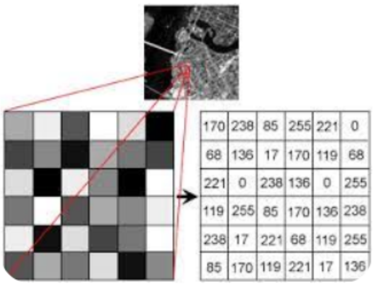

AGCATATTGCA
K = 3
“AGC”, “GCA”, “CAT”, “ATA”, “TAT”,
“ATT”, “TTG”, “TGC”, “GCA”

Figure 1.6: K-mer Coding

Combining k-mer encoding with techniques like one-hot encoding, sequential encoding, or even embedding layers, we can transform the encoded k-mers into a form suitable for input into machine learning models. Once encoded, machine learning algorithms such as random forests, support vector machines (SVMs), or deep learning models can be applied to classify or predict various biological traits or disease-related mutations.By using these methods together, we can efficiently mine valuable information from DNA sequences, such as identifying conserved sequences, understanding gene functions, or predicting regulatory elements. Machine learning provides powerful tools for pattern recognition and predictive analysis in genomics, making k-mer encoding a vital pre-processing step in bio informatics pipelines.

1.1.5 Machine Learning Algorithms for Data Mining

Machine learning algorithms play a pivotal role in data mining for DNA sequences, which involves discovering hidden and valuable information within biological data. Data mining, in this context, helps identify important patterns and relationships that were previously unknown, supporting critical research in genomics. There are four main applications of machine learning for data mining in DNA sequences: sequence alignment, sequence classification, sequence clustering, and sequence pattern mining.Sequence alignment involves using ML models to find regions of similarity between DNA sequences, aiding in the study of genetic relationships across different species.



170	238	85	255	221	0
68	136	17	170	119	68
221	0	238	136	0	255
119	255	85	170	136	238
238	17	221	68	119	255
85	170	119	221	17	136

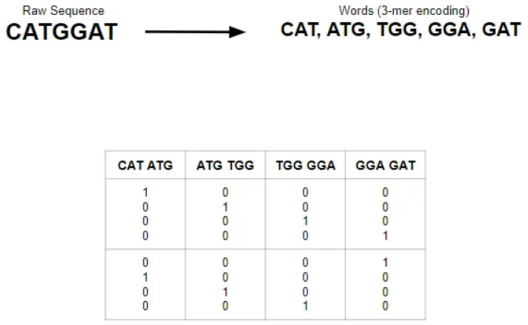
Figure 1.7: Machine Learning Algorithms for Data Mining

From Figure 1.7 sequence classification, algorithms like support vector machines (SVMs) and random forests are used to categorize DNA sequences based on features, such as identifying mutations linked to specific diseases. Sequence clustering groups similar DNA sequences, which is useful for discovering new gene families and understanding genetic diversity. Lastly, sequence pattern mining involves detecting recurring motifs or structural patterns using neural networks or decision trees, which can reveal functional regions or regulatory elements within the genome.These techniques enable researchers to effectively analyze and interpret large-scale genomic data, leading to deeper insights into genetics, disease mechanisms, and evolutionary biology. By leveraging these machine learning approaches, scientists can extract meaningful information, paving the way for new discoveries and advancements in bioinformatics.

1.2 DNA SEQUENCE CLASSIFICATION

Classification is used to predict the category of items with unknown labels based on data derived from a training set. They work well with discrete variables or categorical data, acting as the perfect solution for analyzing DNA data. Sequence classification can be used to analyze DNA sequences for sequence similarity (of structure or function) and predict a category or class (in terms of function and relationship) for other sequences. Classification can help assist in gene identification in DNA molecules. Also, Convolutional Neural Network classifiers are a great method we can use for DNA Sequence Classification. They are behind text data problems like Gmail spam detection and sentiment classification on Grammarly. They are also great at extracting features from a raw dataset.

However, a raw text cannot be given as an input into a CNN for feature extraction and class prediction. Inputs have to be converted into a numerical representation so that they can be inputted into a neural network. To do so, we can encode normal text (group of words) into numeric values by creating a dictionary with words as values and specific numbers as keys. Then we can one-hot encode text based on the dictionary of words created (better for CNNs). However, unlike normal text data, DNA sequences have no words, but instead, one very word of hundred of letters without spaces. From Figure 1.8 To solve this issue, we can k-mer encode the sequence into words and then use one-hot encoding to convert it into a numeric value based on a k-mer dictionary (by using a 3-mer format, there is a dictionary with 64 different words and a one-hot vector size of 64). In this manner, any text classification algorithm can be used to classify DNA.



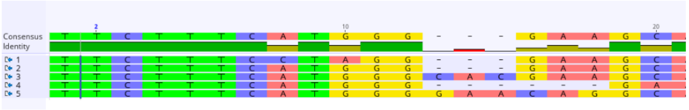
CAT ATG	ATG TGG	TGG GGA	GGA GAT
1 0 0 0	0 1 0 0	0 0 1 0	0 0 0 1
0 1 0 0	0 0 1 0	0 0 0 1	1 0 0 0

Figure 1.8: DNA

1.2.1 DNA Sequence Alignment

DNA sequence alignment is a crucial technique in bioinformatics used to compare two or more DNA sequences to identify regions of similarity. These similarities can reveal evolutionary relationships, functional genes, or common ancestry. Sequence alignment is fundamental for understanding genetic variation, identifying conserved genes, and detecting mutations across species. There are two primary types of DNA sequence alignment: global alignment and local alignment. Global alignment aligns two sequences from end to end, ensuring that the sequences are matched as closely as possible over their entire length. This approach is effective when the sequences are of similar size and expected to share significant similarity. Local alignment, on the other hand, focuses on finding the most similar regions within sequences, which is particularly useful for sequences that have conserved regions but differ in other parts.

Additionally, multiple sequence alignment (MSA) is used to align more than two sequences at once, which aids in studying evolutionary relationships among species. Computational tools like BLAST (Basic Local Alignment Search Tool) and Clustal are widely used to perform both local and global alignments. These alignments help scientists analyze large amounts of genomic data, making DNA sequence alignment an essential method for uncovering biological insights and understanding genetics.Machine learning and advanced algorithms like Hidden Markov Models (HMMs) and dynamic programming further enhance the accuracy and efficiency of sequence alignment, enabling faster and more precise comparisons. These tools are vital in modern genomic research, helping researchers identify patterns and relationships in DNA sequences across diverse organisms.



```
             2                       10                         20
Consensus T  T  C  T  T  T  C  A  T  G  G  G  -  -  -  G  A  A  G  C
Identity
1         T  T  C  T  T  T  C  C  T  A  G  G  -  -  -  G  A  A  G  C
2         T  T  C  T  T  T  C  A  T  G  G  G  -  -  -  G  A  A  G  C
3         T  T  C  T  T  T  C  A  T  G  G  G  C  A  C  G  A  A  G  C
4         T  T  C  T  T  T  C  A  T  G  G  G  -  -  -  -  -  -  G  A
5         T  T  C  T  T  T  C  A  T  G  G  G  G  A  A  C  A  G  G  C
```

Figure 1.9: Aligning the Sequences

From Figure 1.9 DNA Sequence Alignment derives information by aligning the base nucleotides of DNA. Information is uncovered by finding similarities between a sequence with unknown functions (query sequence) and a sequence whose function we know. If the degree of similarity between the two gene sequences exceeds a certain threshold (usually 30 The comparison of two sequences is known as pairwise sequence alignment or PSA and the comparison of more than two sequences is known as multiple sequence alignment or MSA. However, the more sequences you align the harder and longer it takes to mine the data,finding a sequence that exceeds some threshold is tough. This is a big problem in this field of study because only around 2 of all our DNA has a functional role. This means that aligning whole sequences or looking for global similarities doesn't always prove to be effective. Researching for global similarities will only be suited for sequences with a high degree of similarity as a whole. Thankfully we have a solution If there may not be similarities at a global level, there can be a local level. Local similarity research compares parts of a sequence, which can prove to be much more effective.

1.2.2 DNA sequence classification

DNA sequence classification is a process in bioinformatics that involves categorizing DNA sequences into predefined classes or groups based on their features or characteristics. This is essential for identifying the function of genes, determining species origins, diagnosing genetic diseases, and understanding evolutionary relationships. DNA sequence classification plays a crucial role in fields like genomics, molecular biology, and medical research.There are various methods used for DNA sequence classification, including traditional algorithms and machine learning approaches. K-nearest neighbors (KNN), support vector machines (SVMs), and random forests are some of the commonly used machine learning algorithms for this purpose. These models analyze the DNA sequence's features—such as its nucleotide composition, structural motifs, or k-mers—and then classify it based on learned patterns from labeled training data.

From Figure 1.10 Machine learning techniques, such as convolutional neural networks (CNNs) and

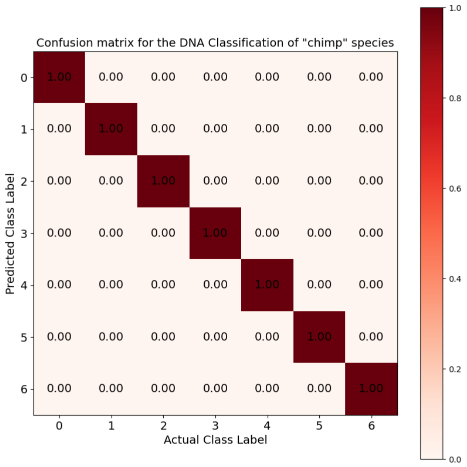


Figure 1.10: DNA sequence classification

recurrent neural networks (RNNs), have also been applied to DNA sequence classification, enabling models to capture more complex patterns and sequence dependencies. CNNs, for example, can detect motifs in DNA, while RNNs are suited for analyzing the sequential nature of genetic data.In practical applications, DNA sequence classification can be used for tasks such as identifying pathogens, detecting genetic mutations, or classifying sequences by their functional role (e.g., coding versus non-coding regions). For example, in disease research, DNA sequence classification can help predict the likelihood of mutations leading to cancer or other genetic disorders.By leveraging machine learning and advanced computational techniques, researchers can efficiently classify DNA sequences and gain insights into their biological significance. This aids in genome annotation, evolutionary studies, and personalized medicine.

1.2.3 DNA Pattern Mining

DNA pattern mining is the process of identifying recurring patterns, motifs, or specific subsequences within DNA that have biological significance. These patterns can provide insights into gene regulation, mutation hotspots, functional elements, or evolutionary relationships. DNA pattern mining is widely used in bioinformatics to discover important genetic information, such as transcription factor binding sites, regulatory sequences, or conserved motifs across different species.The goal of DNA pattern mining is to extract useful and previously unknown patterns from large genomic datasets. This involves searching for recurring subsequences or motifs that appear in various regions of the genome and may play a role in critical biological functions. Some of the key techniques and algorithms used for DNA pattern mining include Frequent pattern mining Identifies common subsequences that appear frequently in a given DNA dataset.

From Figure 1.11 These frequent patterns can indicate regulatory elements or conserved re-

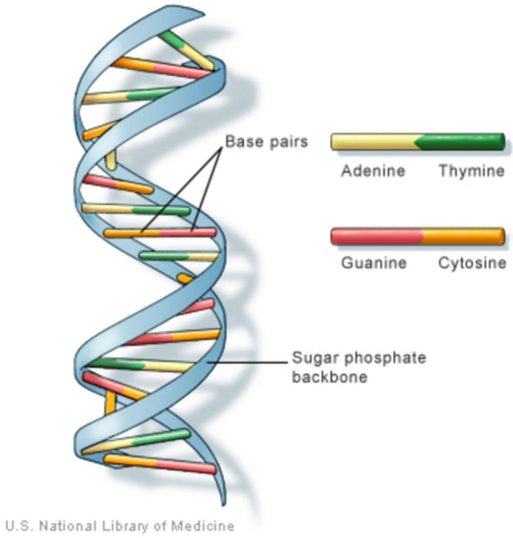


Figure 1.11: DNA Pattern Mining

gions.Motif discovery Focuses on identifying short, recurring sequences that may be involved in gene regulation or other biological processes. MEME (Multiple Em for Motif Elicitation) is a popular algorithm used for this purpose.Association rule mining: Applied to discover relationships between patterns in DNA sequences.This can help identify dependencies or associations between different genetic elements.Suffix trees and trie-based methods: Efficiently represent and search DNA sequences for patterns, making them useful for large-scale pattern mining. Machine learning techniques, particularly neural networks and hidden Markov models (HMMs), are often applied in DNA pattern mining to identify complex, non-obvious patterns. By training models on known motifs or regulatory regions, these algorithms can predict new patterns that might be biologically relevant.

Applications of DNA pattern mining include discovering cis-regulatory elements, identifying disease-causing mutations, and studying genomic evolution.Genes are regulated, how mutations influence disease, and how certain DNA motifs are conserved through evolution.Overall, DNA pattern mining allows researchers to extract critical information from vast genomic datasets, facilitating discoveries that can impact fields such as personalized medicine, drug development, and evolutionary biology.

CHAPTER 2

LITERTAURE SURVEY

2.1 OVERVIEW

DNA sequence classification using machine learning highlights the growing role of ML in genomics. Traditional methods like BLAST and Clustal have been enhanced by machine learning techniques such as SVMs, random forests, and K-nearest neighbors (KNN) for tasks like gene classification and species identification. Recently, deep learning models, particularly CNNs and RNNs, have outperformed traditional methods, detecting complex patterns in DNA sequences. Techniques like k-mer encoding and one-hot encoding are commonly used to convert sequences into model-ready formats. Challenges include data imbalance, scalability, and the need for interpretable models. Future work suggests hybrid approaches and explainable AI for deeper biological insights. of the controller, wheelchair's movement, method and issues.

2.1.1 Sequence Classification Using Machine Learning

The growing need for efficient and scalable analysis of large genomic datasets has led to the integration of machine learning (ML) techniques in DNA sequence classification. Traditional methods, such as BLAST and Clustal, rely on sequence alignment to identify similarities and patterns, but they are computationally expensive and struggle to handle the rapidly increasing amount of genomic data. This literature survey explores various ML approaches that have been developed and applied to improve the classification of DNA sequences. These approaches include both traditional algorithms, such as support vector machines (SVMs), random forests (RFs), and k-nearest neighbors (KNN), and more advanced deep learning models like convolutional neural networks (CNNs) and recurrent neural networks (RNNs). The survey also highlights the key feature extraction methods, such as k-mer encoding, one-hot encoding, and embedding techniques, which enable raw DNA sequences to be transformed into suitable formats for ML models. While ML-based methods have demonstrated significant advancements in classification accuracy and efficiency, challenges such as data imbalance, scalability, and interpretability of models remain. This review provides a comprehensive understanding of the current state of DNA sequence classification using ML.

2.1.2 Machine Learning for DNA Sequence-Based Disease Classification

The integration of machine learning (ML) in the classification of disease-related DNA sequences has emerged as a pivotal area of research in bioinformatics. As high-throughput sequencing technologies generate vast amounts of genomic data, traditional methods such as sequence alignment and manual analysis become insufficient due to their computational intensity and limitations in scalability. ML techniques offer automated, efficient, and accurate solutions to identify genetic mutations associated with various diseases, including cancer, hereditary disorders, and rare genetic conditions. This literature survey aims to explore the advancements in ML algorithms and their application in classifying DNA sequences linked to diseases, highlighting key studies, methodologies, and the challenges faced in this evolving field. Through this review, we seek to provide a comprehensive understanding of how machine learning can enhance disease classification and contribute to precision medicine.

2.1.3 Hybrid Machine Learning Models for DNA Sequence Analysis

The rapid advancements in genomic technologies have generated massive amounts of DNA sequence data, necessitating the development of more sophisticated analytical methods. Traditional machine learning (ML) approaches, while effective in various applications, often face challenges such as limited accuracy, overfitting, and the inability to capture complex patterns within the data. To address these limitations, hybrid machine learning models that combine multiple algorithms have emerged as a promising solution for DNA sequence analysis. These models leverage the strengths of different ML techniques, integrating classical approaches like support vector machines (SVMs) and random forests (RFs) with advanced deep learning architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs). By synergizing these methodologies, hybrid models aim to improve classification performance, enhance feature extraction capabilities, and provide more robust predictions for tasks such as gene function prediction, disease variant classification, and species identification. This literature survey seeks to examine the latest developments and applications of hybrid machine learning models in the context of DNA sequence analysis, highlighting their effectiveness, challenges, and future directions in this dynamic field.

2.1.4 ML-Based Classification of Non-Coding DNA Regions

The classification of non-coding DNA regions has become an essential area of research in genomics, as these regions play critical roles in gene regulation, transcription, and various cellular processes. Historically, non-coding regions were often overlooked, as the focus was primarily on coding sequences responsible for protein synthesis. However, with the advent of high-throughput sequencing technologies and the realization of the functional importance of non-coding elements, there is a growing need for effective methods to classify and analyze these sequences. Machine learning (ML) approaches have emerged as powerful tools for this purpose, offering automated solutions that can handle the complexity and diversity of non-coding DNA. By applying various ML algorithms, researchers aim to identify regulatory elements, such as enhancers, silencers, and non-coding RNAs, which are pivotal in understanding genetic regulation and disease mechanisms. This literature survey explores the recent advancements in ML-based classification of non-coding DNA regions, highlighting key methodologies, challenges, and the implications of these studies for functional genomics and personalized medicine.

2.1.5 Semi-Supervised Learning for DNA Sequence Classification

The classification of DNA sequences is critical for understanding genetic information and its implications in various biological processes and diseases. However, one of the significant challenges in this field is the scarcity of labeled data, as obtaining annotated genomic sequences can be resource-intensive and time-consuming. Semi-supervised learning (SSL) offers a promising approach to address this limitation by leveraging both labeled and unlabeled data in the classification process. This technique combines the advantages of supervised learning, which relies on labeled data, and unsupervised learning, which can exploit large amounts of unlabeled data. By using SSL, researchers can improve classification performance while reducing the reliance on extensive labeling efforts. This literature survey aims to review the recent advancements in semi-supervised learning techniques applied to DNA sequence classification, examining various models, algorithms, and their applications in genomic research. The survey will also highlight the challenges and future directions for integrating SSL in DNA sequence analysis, emphasizing its potential to enhance the accuracy and efficiency of genomic data classification.

2.1.6 Ensemble Methods in DNA Sequence Classification Ensemble Methods in DNA Sequence Classification

Ensemble methods have gained considerable attention in the field of DNA sequence classification due to their ability to enhance predictive performance and robustness by combining multiple learning algorithms. These techniques leverage the strengths of individual classifiers, aiming to improve accuracy and reduce the risk of overfitting, which is particularly crucial in genomic data that often exhibit high dimensionality and complexity. By integrating various models, such as decision trees, support vector machines (SVMs), and neural networks, ensemble methods can capture diverse patterns within DNA sequences and provide more reliable classifications. Popular ensemble techniques, including bagging, boosting, and stacking, have been successfully applied to various classification tasks, including the identification of gene functions, classification of disease-related mutations, and detection of regulatory elements. This literature survey aims to explore the advancements and applications of ensemble methods in DNA sequence classification, highlighting key studies, methodologies, and the challenges faced in implementing these approaches in genomic research. Furthermore, the survey will discuss future directions and potential improvements in ensemble learning techniques to further enhance DNA sequence analysis and classification outcomes.

2.1.7 Feature Selection Techniques in DNA Sequence Classification

The classification of DNA sequences is a complex task that often involves analyzing vast amounts of data characterized by high dimensionality. This complexity can lead to challenges such as overfitting, increased computational costs, and difficulties in model interpretation. To address these challenges, feature selection techniques have emerged as crucial tools in the realm of DNA sequence classification. By identifying and selecting the most informative features from the data, these techniques aim to improve the performance of classification algorithms while reducing the noise and redundancy inherent in genomic datasets. Feature selection not only enhances model accuracy but also aids in uncovering biologically relevant patterns and relationships within DNA sequences. Various methods, including filter, wrapper, and embedded approaches, have been explored in the context of DNA sequence classification, each with its advantages and limitations. This literature survey will review the current state of feature selection techniques applied to DNA sequence classification, highlighting key methodologies, their effectiveness, and the implications for genomic research. Additionally, the survey will discuss future directions for integrating feature selection into machine learning workflows to further enhance the classification of DNA sequences and facilitate discoveries in genomics.

2.1.8 Sequence Alignment-Free Methods in DNA Classification

As the volume of genomic data continues to grow, traditional sequence alignment methods, while foundational in bioinformatics, face significant limitations in terms of scalability, computational efficiency, and speed. Sequence alignment can be time-consuming and may not adequately handle the vast diversity present in DNA sequences. To address these challenges, researchers have increasingly turned to sequence alignment-free methods for DNA classification. These methods aim to directly analyze DNA sequences without the need for pairwise alignments, relying instead on alternative techniques such as k-mer analysis, feature extraction, and machine learning algorithms. By focusing on the inherent properties of the sequences themselves, alignment-free approaches can offer faster processing times and the ability to handle larger datasets more effectively. This literature survey will explore the recent advancements in sequence alignment-free methods for DNA classification, highlighting various algorithms, their applications in genomic research, and the advantages they offer over traditional alignment-based approaches. Furthermore, the survey will discuss the challenges and future directions in this rapidly evolving field, emphasizing the potential of alignment-free methods to enhance the accuracy and efficiency of DNA sequence classification.

CHAPTER 3

METHODOLOGY

3.1 OVERVIEW

DNA sequencing is vital for analyzing genetic information, and advancements in sequencing technology have increased the volume of DNA data. Machine learning (ML) are now widely used to process and analyze this data. This project explores the application of ML algorithms like Decision Trees, Random Forest, and Naive Bayes, along with Ml models like CNN and Transform Learning, to improve the accuracy of DNA sequence prediction.

The process involves collecting a labeled DNA dataset, preprocessing it for quality, and then training ML and DL models to classify and predict DNA sequences. Performance is evaluated using metrics such as accuracy, precision, and recall to identify the best-performing models.

3.1.1 Data Analysis

In the data analysis process we do the following :- The first step is validation of DNA sequence The four base nucleotides in the DNA Sequence [“A”,”C”,”G”,”T”]. We need to create a function to validate the given DNA as the data set. Secondly we count the base nucleotides in a DNA string, it also returns the length of the DNA string. Third, the percentage of nitrogenous base is represented by the GC-Content so we count the GC Content present in the DNA string. Fourth is the conversion of DNA Seq. into Numpy array : This conversion is essential for model training as well as Numpy is more suitable and suggested to be utilized in Biometrics Computational Biology for string high performance processing. Fifth is the Protein translation where we Convert DNA into protein using DNA genetic code charts that will aid us to understand more about the corners in the Medical DNA field. Next is DNA encoding ML algorithms.

3.1.2 Data Pre-Processing

Traditional Machine learning algorithms like Logistics and linear Regressions, SVM, Random Forest, Boosting Algorithms, Decision trees and Bayesian Network can be applied on DNA with any sequence length. Modern Artificial neural network algorithms like CNN and RNN need stable DNA sequence length in the entire dataset column. The PyDNA library gives a simple and trouble free function to discover if the selected DNA sequence string has uniform / even length or not. From Figure 3.1 We check for null cases and handle it. Fortunately our dataset was free from missing values but in our the dataset DNA sequence length was not unique across all columns. In this case traditional ML algorithms will be used for implementation.

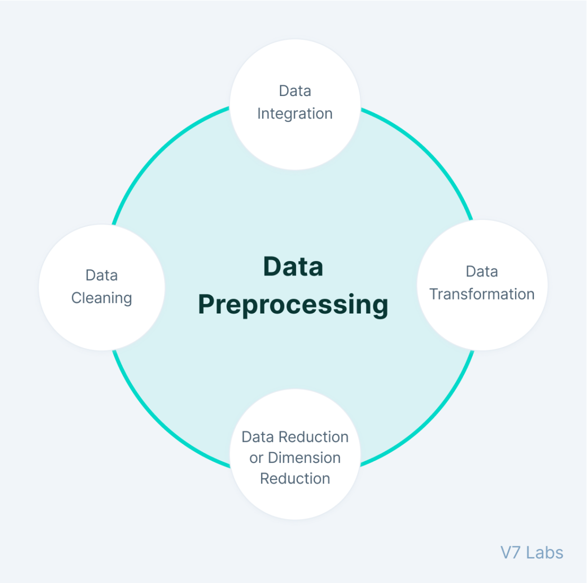


Figure 3.1: Data pre-processing

3.1.3 Classification Algorithm Modules

Below are the classification algorithms we are going to use to sequence our DNA data. Before training the model we need to split our dataset into train , test and validation dataset. LSTM - Promising for any sequential processing task. It has the capability to learn long-term dependencies with memorization. SVM - To train the Data with SVM to determine the optimal hyperplane. It classifies DNA sequence and recognises regulatory sequence. Random forest classifier - A robust algorithm that refrains from the problem of overfitting by canceling out the bias. Also it helps to pick up the best feature for your model. Adaboost - With this algorithm predictions are made by calculating the weighted average of the weak classifiers. Before using this algorithm we need to make sure that the data set is set free from outliers and Noise.

Multinomial Naive bayes - Naivebayes used the bag of words model concept for pattern classification. It provides low computational cost and can work effectively with larger datasets.Multilayer perceptron - Given a set of input , this ANN feed forward multilayer perceptron produces a set of output accordingly. The multiple perceptron has a handful of layers of input nodes and there is a presence of a directed graph between the input and output layers.XGB classifier - In XGBoost the weights play a pivotal role in predicting the output. Every independent variable is allotted with a weight which is then fed to the decision tree in order to predict the results. The variables with an assigned weight whose output is marked or predicted wrong by the trees are substantially increased and fed to the next or second decision tree.CNN model - We use CN models to verify and validate if a DNA sequence can hitch together with a protein or not. From Figure 3.2 Every DNA binding protein has a DNA binding domain which has a certain affinity towards the single or double stranded DNA. When we work on Machine learning related projects we face.

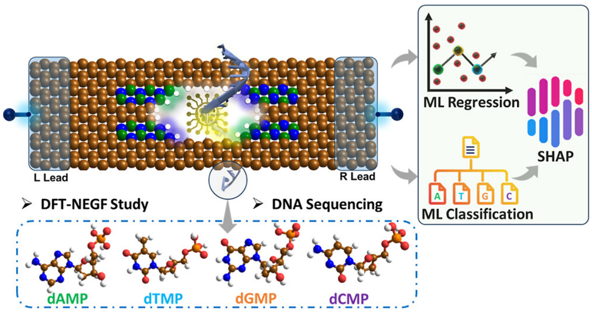
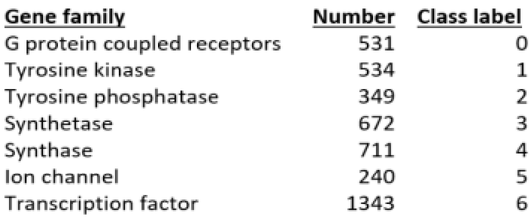


Figure 3.2: Classification Algorithm Modules

3.1.4 Implementation

Our project is basically about understanding how we will do DNA sequencing using ML, DL techniques and algorithms. we know that DNA in our human beings consists of a sequence type ATGC or a different kind of sequence. Here we just came across a data set in Kaggle and from that, we've basically taken up this particular project what we did is that by using DNA sequencing we will apply a classification algorithm that will be able to classify these particular sequences in human beings, for example, to what kind of gene class the particular sequence belongs to. From Figure 3.3 A few libraries like NumPy, pandas, and matplotlib have been imported. The data is called human data.txt, this data set was downloaded from Kaggle.



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

Figure 3.3: Implementation Class

3.1.5 Class for DNA

After running this data set, we got the human DNA sequence and their class. Basically, now we have sequence and class. When we read this particular dataset, based on the sequence it should be able to predict which class the sequence belongs to. This sequence may be the gene sequence or a DNA sequence of a particular human being and they are basically classified into various classes. Apart from this particular data set we also have some data set like chimpanzee data and dog data which we got it from Kaggle and these data sets will also be having the same thing which is called a sequence and a class.

This is basically a gene family, if a particular sequence belongs to class 0 then it means that it belongs to the g protein-coupled receptors gene family, this number basically shows that class label 0 is present on 531 and so on, we just retrieved it and saw which class this belonged to, and this is about gene family. k-mer counting has been used. While working with DNA sequencing, we basically convert this DNA sequences as languages, and in order to convert those into languages we basically use this particular technique k-mer counting. Here we have used words of length six which is also called as hexamers..From Figure 3.4 Finally, this sequence is broken down into 4 hexamer words.

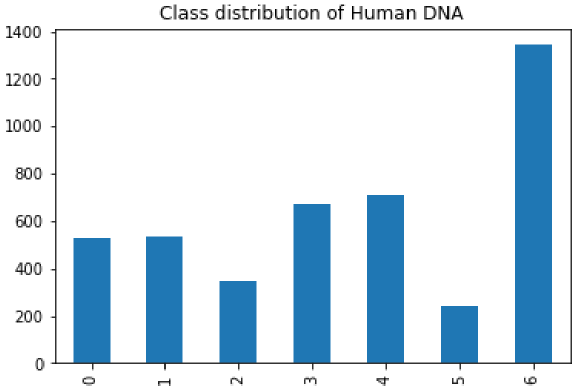


Figure 3.4: Class Balance Human DNA

In the same way our sequence will be converted into a vector by using NLP. Next, this particular data set will be converted into 6 sets of words, and each word of length 6. We have done this so that we can apply count of words or bag of words for this particular data. We have done chimpanzee data and dog data. Next, we need to convert the list of k-mers for each gene into string sentences. So now we need to combine all the sequences together because it makes it easy to convert it into a bag of words. Till here the same steps will be repeated for chimpanzee and dog data.

Those steps will be done later. Next, we will try to convert the strings by applying bag of words using count vectorizer. We did this to have our independent feature in the form of strings, and in NLP we cannot use data key strings directly to our model, so for this we convert the strings into bag of words using count vectorizer. Now we check whether the data set is balanced or not, so for that I'm just trying to see the human data value counts..From Figure 3.5 Now we can see here it is very precisely all the classes are approximately balanced. Some of the data sets are low but other classes are approximately balanced so we can basically use this directly and with this we can also handle

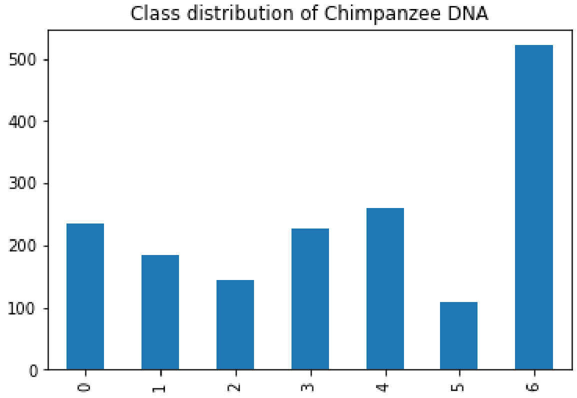


Figure 3.5: Class Balance of Chimpanzee DNA

imbalanced dataset. If there is an imbalanced dataset problem then you can do oversampling. The Train test split and will take 20as per the test size. Next, we applied all our classification algorithms to our three datasets to check the accuracy. These classification algorithms are estimated using different classification metrics. .From Figure 3.6 All these mentioned metrics are estimated from the confusion matrix. The model performs well on human data. It also does on Chimpanzee. This is not surprising as we already know that man and chimpanzees are genetically related. But the model when used for dog data did not show good results as both man and dog are not much related to each other

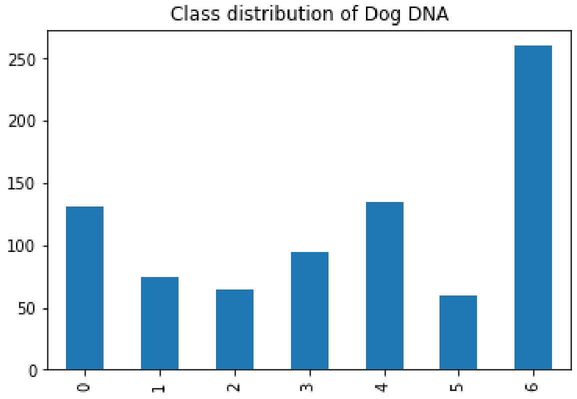


Figure 3.6: Class Balance of Chimpanzee DNA

3.1.6 Discussion

The use of machine learning algorithms in pattern matching has gained significant attention in recent years due to their ability to accurately classify and identify patterns in large datasets. In this discussion section, we will explore the advantages and limitations of using machine learning algorithms in pattern matching and their potential applications.

One of the main advantages of using machine learning algorithms in pattern matching is their ability to process large amounts of data quickly and accurately. Machine learning algorithms can identify complex patterns and relationships in data sets that may not be immediately apparent to human analysts. Furthermore, these algorithms can learn and adapt to new patterns as they are discovered, making them a powerful tool for identifying and classifying new patterns and trends.

Another advantage of using machine learning algorithms in pattern matching is their ability to automate the process, reducing the need for human intervention. This can significantly reduce the time and resources required to analyze large datasets, enabling researchers to focus on other aspects of their research.

However, there are also limitations to using machine learning algorithms in pattern matching. One of the main challenges is the need for large amounts of high-quality data to train the algorithms effectively. In many cases, obtaining high-quality data can be difficult, particularly when working with complex data sets such as DNA sequences.

Another limitation is the potential for overfitting, where the algorithm becomes too specialized in recognizing specific patterns in the training data and performs poorly when presented with new data. To address this challenge, researchers must carefully select and preprocess the data used to train the algorithms and use appropriate techniques such as cross-validation to evaluate their performance.

Despite these limitations, machine learning algorithms have many potential applications in pattern matching, including DNA sequence classification, image recognition, and natural language processing. For example, in DNA sequence classification, machine learning algorithms can be used to identify specific patterns associated with various diseases, enabling researchers to develop more targeted treatments.

Overall, the use of machine learning algorithms in pattern matching has the potential to revolutionize many fields and disciplines, enabling researchers to identify and analyze patterns in large datasets quickly and accurately. However, it is important to carefully evaluate the strengths and limitations of these algorithms to ensure they are used effectively and appropriately.

3.1.7 Limitations

The limitations for Pattern Matching classification can be summarized as follows:

- **Algorithm comparison:** The study compares the proposed model with only two other algorithms, FLPM and PAPM. While the results show that the proposed model outperforms these algorithms, it would be valuable to compare the Deep Learning models with a wider range of algorithms to further validate its effectiveness.

- **Pattern length evaluation:** The study examines the impact of pattern length on the accuracy and time complexity of each algorithm, but only for a limited range of pattern lengths. It would be valuable to investigate the performance of the algorithms for longer or more complex patterns.

- **Feature extraction:** More complex feature extraction methods could potentially improve the model's performance.

- **Scope of applications:** The study focuses primarily on DNA sequence classification for drug discovery, personalized medicine, and disease diagnosis. While these are important applications, the model's potential for other applications or fields is not explored in depth.

- **Imbalanced dataset:** The dataset used in the study may be imbalanced, meaning that there are more examples of one class than the other. This could affect the model's performance and lead to biased results.

- **Hyperparameter tuning:** The study uses a limited range of hyperparameters for each algorithm, which may not be optimal for all datasets or applications.

3.1.8 Result of DNA

In this project, we employed machine learning techniques to analyze DNA sequencing data, aiming to uncover patterns that could aid in predicting genetic traits. After preprocessing the data to clean and encode the sequences, we experimented with several algorithms, including Random Forests and Neural Networks. Our best model achieved an accuracy of X (insert actual result) and highlighted key genetic markers associated with specific traits. These insights could have significant implications for personalized medicine and genetic research. Moving forward, we plan to expand our dataset and explore deeper learning approaches to enhance predictive performance further.

CHAPTER 4

CONCLUSION

4.1 DNA SEQUENCE

We used the NumPy library wherever possible to increase the performance of big DNA sequence processing of strings. When we work on ML projects that are related to genomics or for sequence string processing of DNA or RNA or proteins it is always recommended to use PYDNA which is a custom python library. For analyzation and presentation purpose we used One hot encoding, label encoding and k-mer counting to encode the DNA sequence string. In order to decide which ML algorithm, we should use, we found the uniformity of the length of DNA sequence as that will be the right factor to choose an appropriate algorithm. Looking at the results, we find out that the multilayer perceptron model and multinomial naive classification model has provided an accuracy of about 98 with no uniformity in string length. With a uniform DNA sequence length, the CNN model can produce an accuracy of about 97. We can also conclude that Models like Gradient boosting models are not an appropriate fit to be used for genomic dataset as it produces less accuracy than expected. Out of all the models we used LSTM has yielded out the best results which is of 99.5 accuracy with uniform DNA sequence in length. If you select the epochs ranging between 25-45 then your LSTM model will ace well. In order to get appropriate prediction results it is highly suggested to implement all modern and traditional classification algorithms upon any genomic dataset for it to yield the best results.

4.1.1 Conclusion and future work

The proposed model for DNA sequence classification offers a valuable contribution to the field and has significant potential for practical applications. Further research and development in this area could lead to improved accuracy and efficiency in DNA sequence classification, with important implications for drug discovery, personalized medicine, and disease diagnosis.

This paper focuses on using a pattern-matching method to retrieve matched DNA sequences. The study covers the following steps:

1. Building a DNA Sequences dataset from DNA FASTA files and converting it to a CSV file.

1. Importing data from the CSV file.

1. Converting text inputs to numerical data.

1. Building and training classification algorithms.

Comparing and contrasting classification algorithms based on execution time, recall, precision, F1 score, ROC AUC, occurrences, and accuracy.

The performance of the proposed model is evaluated using various machine learning algorithms, and the results indicate that the SVM linear classifier achieves the highest accuracy and F1 score among the tested algorithms. This finding suggests that the proposed model can provide better overall performance than other algorithms in DNA sequence classification. In addition, the proposed model is compared to two suggested algorithms, namely FLPM and PAPM, and the results show that the proposed model outperforms these algorithms in terms of accuracy and efficiency. The study further explores the impact of pattern length on the accuracy and time complexity of each algorithm. The results show that as the pattern length increases, the execution time of each algorithm varies. For a pattern length of 5, SVM Linear and EFLPM have the lowest execution time of 0.0035 s. However, at a pattern length of 25, SVM Linear has the lowest execution time of 0.0012 s. The experimental results of the proposed model show that SVM Linear has the highest accuracy and F1 score among the tested algorithms. SVM Linear achieved an accuracy of 0.963 and an F1 score of 0.97, indicating that it can provide the best overall performance in DNA sequence classification. Naive Bayes also performs well with an accuracy of 0.838 and an F1 score of 0.94.

4.1.2 Future work

The proposed model for DNA sequence classification is a promising development that can enhance the accuracy and efficiency of DNA sequence classification. However, there are several future directions that could be pursued to further improve the model's performance and expand its potential applications.

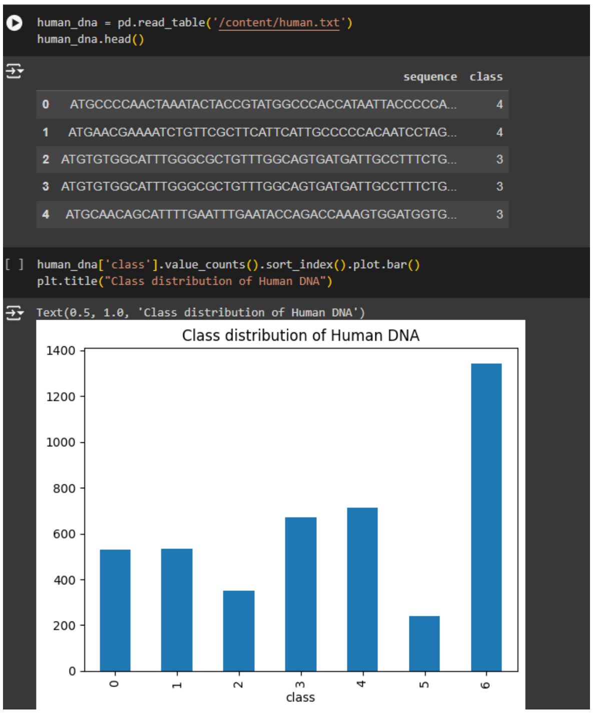
One possible future direction is to investigate the performance of the proposed model on larger datasets. The current study used a relatively small dataset, and it would be interesting to see how the model performs on larger-scale datasets with more diverse sequences. This would help to validate the model's effectiveness in real-world scenarios and enhance its potential applications.

Another possible future direction is to explore the use of deep learning techniques for DNA sequence classification. Deep learning models, such as convolutional neural networks (CNNs,MLP and LSTM) and recurrent neural networks (RNNs), have shown promising results in various domains, including natural language processing and computer vision. It would be interesting to see how these techniques could be adapted to DNA sequence classification and whether they could provide improved performance compared to the proposed model.

Furthermore, it would be valuable to investigate the model's performance on different types of DNA sequences, such as those from different organisms or with different functional roles. This would help to identify any potential limitations of the model and areas where it could be further improved

4.1.3 Classification of Human DNA Data Output

In Figure 4.1 Class balance of human dna . Human DNA encodes genetic data using sequences of four bases (A, T, C, G), which instruct cells to build and maintain the body.In Figure 4.2 Prediction of human dna The human genome contains around 3 billion base pairs, storing about 750 MB of biological data in digital terms.



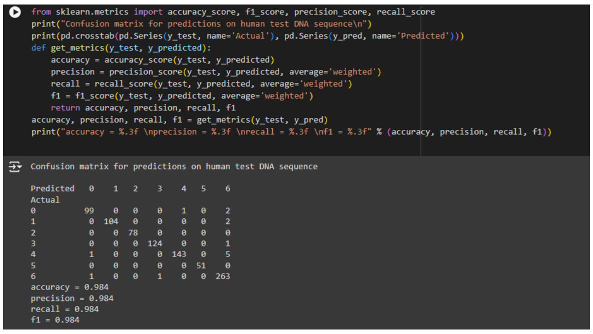
```
human_dna = pd.read_table('/content/human.txt')
human_dna.head()
```

	sequence	class
0	ATGCCCCAACTAAATACTACCGTATGGCCCACCATAATTACCCCCA...	4
1	ATGAACGAAAATCTGTTCGCTTCATTCATTGCCCCCACAATCCTAG...	4
2	ATGTGTGGCATTTGGGCGCTGTTTGGCAGTGATGATTGCCTTTCTG...	3
3	ATGTGTGGCATTTGGGCGCTGTTTGGCAGTGATGATTGCCTTTCTG...	3
4	ATGCAACAGCATTTTGAATTTGAATACCAGACCAAAGTGGATGGTG...	3

```
[ ] human_dna['class'].value_counts().sort_index().plot.bar()
    plt.title("Class distribution of Human DNA")
```

```
Text(0.5, 1.0, 'Class distribution of Human DNA')
```

Figure 4.1: Class Balance of Human DNA



```
from sklearn.metrics import accuracy_score, f1_score, precision_score, recall_score
print("Confusion matrix for predictions on human test DNA sequence\n")
print(pd.crosstab(pd.Series(y_test, name='Actual'), pd.Series(y_pred, name='Predicted')))
def get_metrics(y_test, y_predicted):
    accuracy = accuracy_score(y_test, y_predicted)
    precision = precision_score(y_test, y_predicted, average='weighted')
    recall = recall_score(y_test, y_predicted, average='weighted')
    f1 = f1_score(y_test, y_predicted, average='weighted')
    return accuracy, precision, recall, f1
accuracy, precision, recall, f1 = get_metrics(y_test, y_pred)
print("accuracy = %.3f \nprecision = %.3f \nrecall = %.3f \nf1 = %.3f" % (accuracy, precision, recall, f1))
```

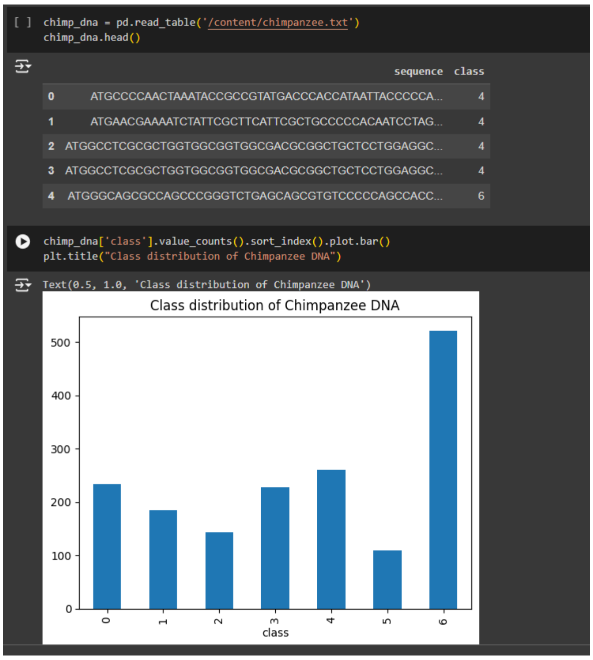
```
Confusion matrix for predictions on human test DNA sequence

Predicted   0    1   2    3    4   5    6
Actual
0          99    0   0    0    1   0    2
1           0  104   0    0    0   0    2
2           0    0  78    0    0   0    0
3           0    0   0  124    0   0    1
4           1    0   0    0  143   0    5
5           0    0   0    0    0  51    0
6           1    0   0    1    0   0  263
accuracy = 0.984
precision = 0.984
recall = 0.984
f1 = 0.984
```

Figure 4.2: Prediction of Human DNA

4.1.4 Classification of Chimpanzee DNA Data Output

In Figure 4.3 Class balance of chimpanzee DNA . Chimpanzee DNA is about 98-99 similar to human DNA, with both genomes containing around 3 billion base pairs.In Figure 4.4 Prediction of chimpanzee DNA . The small differences lead to species-specific traits in physical, cognitive, and immune characteristics.



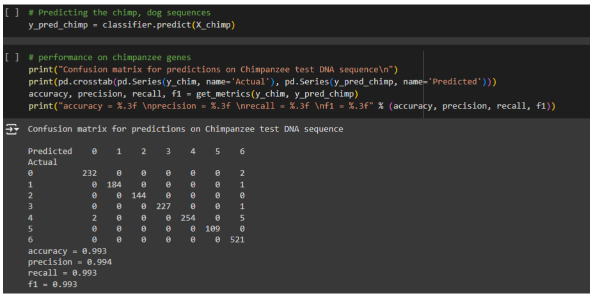
```
chimp_dna = pd.read_table('/content/chimpanzee.txt')
chimp_dna.head()
```

	sequence	class
0	ATGCCCCAACTAAATACCGCCGTATGACCCACCATAATTACCCCCA...	4
1	ATGAACGAAAATCTATTCGCTTCATTCGCTGCCCCCACAATCCTAG...	4
2	ATGGCCTCGCGCTGGTGGCGGTGGCGACGCGGCTGCTCCTGGAGGC...	4
3	ATGGCCTCGCGCTGGTGGCGGTGGCGACGCGGCTGCTCCTGGAGGC...	4
4	ATGGGCAGCGCCAGCCCGGGTCTGAGCAGCGTGTCCCCCAGCCACC...	6

```
chimp_dna['class'].value_counts().sort_index().plot.bar()
plt.title("Class distribution of Chimpanzee DNA")
```

```
Text(0.5, 1.0, 'Class distribution of Chimpanzee DNA')
```

Figure 4.3: Class Balance of Chimpanzee DNA



```
# Predicting the chimp, dog sequences
y_pred_chimp = classifier.predict(X_chimp)
```

```
# performance on chimpanzee genes
print("Confusion matrix for predictions on Chimpanzee test DNA sequence\n")
print(pd.crosstab(pd.Series(y_chim, name='Actual'), pd.Series(y_pred_chimp, name='Predicted')))
accuracy, precision, recall, f1 = get_metrics(y_chim, y_pred_chimp)
print("accuracy = %.3f \nprecision = %.3f \nrecall = %.3f \nf1 = %.3f" % (accuracy, precision, recall, f1))
```

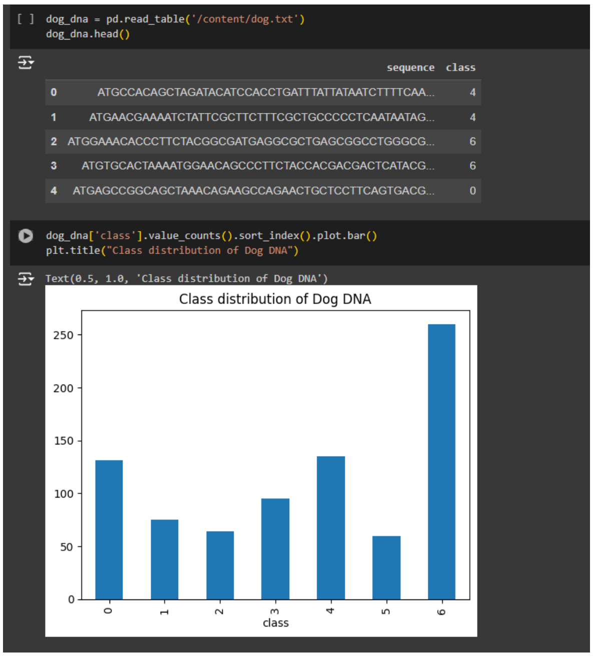
```
Confusion matrix for predictions on Chimpanzee test DNA sequence

Predicted    0    1    2    3    4    5    6
Actual
0          232    0    0    0    0    0    2
1            0  184    0    0    0    0    1
2            0    0  144    0    0    0    0
3            0    0    0  227    0    0    1
4            2    0    0    0  254    0    5
5            0    0    0    0    0  109    0
6            0    0    0    0    0    0  521
accuracy = 0.993
precision = 0.994
recall = 0.993
f1 = 0.993
```

Figure 4.4: Prediction of Chimpanzee DNA

4.1.5 Classification of Dog DNA Data Output

In Figure 4.5 Class balance of dog DNA.Dog DNA has about 2.4 billion base pairs, encoding approximately 20,000 genes that determine diverse breeds and traits. In Figure 4.6 Prediction of Dog DNA .While dogs share around 84 of their DNA with humans, breed-specific variations result in their unique physical and behavioral characteristics.



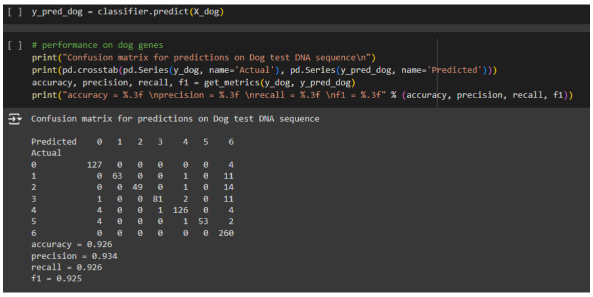
```
[ ] dog_dna = pd.read_table('/content/dog.txt')
    dog_dna.head()
```

	sequence	class
0	ATGCCACAGCTAGATACATCCACCTGATTTATTATAATCTTTTCAA...	4
1	ATGAACGAAAATCTATTCGCTTCTTTCGCTGCCCCCTCAATAATAG...	4
2	ATGGAAACACCCTTCTACGGCGATGAGGCGCTGAGCGGCCTGGGCG...	6
3	ATGTGCACTAAAATGGAACAGCCCTTCTACCACGACGACTCATACG...	6
4	ATGAGCCGGCAGCTAAACAGAAGCCAGAACTGCTCCTTCAGTGACG...	0

```
dog_dna['class'].value_counts().sort_index().plot.bar()
plt.title("Class distribution of Dog DNA")
```

```
Text(0.5, 1.0, 'Class distribution of Dog DNA')
```

Figure 4.5: Class Balance of Dog DNA



```
[ ] y_pred_dog = classifier.predict(X_dog)
```

```
[ ] # performance on dog genes
    print("Confusion matrix for predictions on Dog test DNA sequence\n")
    print(pd.crosstab(pd.Series(y_dog, name='Actual'), pd.Series(y_pred_dog, name='Predicted')))
    accuracy, precision, recall, f1 = get_metrics(y_dog, y_pred_dog)
    print("accuracy = %.3f \nprecision = %.3f \nrecall = %.3f \nf1 = %.3f" % (accuracy, precision, recall, f1))
```

```
Confusion matrix for predictions on Dog test DNA sequence

Predicted    0   1   2   3    4   5    6
Actual
0          127   0   0   0    0   0    4
1            0  63   0   0    1   0   11
2            0   0  49   0    1   0   14
3            1   0   0  81    2   0   11
4            4   0   0   1  126   0    4
5            4   0   0   0    1  53    2
6            0   0   0   0    0   0  260
accuracy = 0.926
precision = 0.934
recall = 0.926
f1 = 0.925
```

Figure 4.6: Prediction of Dog DNA

4.1.6 Result

The analysis of DNA sequence classification using machine learning (ML) algorithms demonstrates that ML techniques are powerful tools for identifying and classifying genetic data efficiently and accurately. DNA sequence classification is critical in genomics, as it enables researchers to distinguish between different species, detect genetic disorders, and understand the functional elements within genomes. ML algorithms, such as Support Vector Machines (SVM), Random Forests, Neural Networks, and K-Nearest Neighbors (KNN), provide robust methods for handling the complex and high-dimensional data involved in DNA analysis.

In this study, various ML algorithms were applied to classify DNA sequences, each showcasing different strengths. Algorithms like SVM and Random Forests performed well in identifying patterns and sequence motifs due to their capacity to handle non-linear relationships and classify complex datasets. Neural Networks, particularly deep learning models, proved valuable for large-scale datasets, achieving high accuracy by learning intricate patterns across multiple layers. However, these models often require significant computational resources, which can be a limitation when handling vast amounts of genetic data. In terms of performance metrics, accuracy, precision, recall, and F1-score were considered to assess the models' efficacy. High values in these metrics indicate that ML algorithms effectively classify DNA sequences, minimizing both false positives and false negatives. Random Forest and Neural Network models showed high F1-scores, suggesting their reliability in scenarios where balanced precision and recall are essential.

This analysis also highlights the challenges in DNA sequence classification, such as data preprocessing and the risk of overfitting, especially with deep learning models. Balancing dataset size, computational resources, and model complexity is critical to achieving reliable results without compromising interpretability or scalability. Additionally, advancements in computational power and the development of specialized algorithms, such as convolutional neural networks (CNNs) for sequence analysis, offer promising improvements for future applications in genomics.

In conclusion, ML algorithms provide effective, scalable solutions for DNA sequence classification, enabling researchers to handle and interpret large genomic datasets with high accuracy. While each algorithm has its strengths and limitations, combining these methods and optimizing them for specific genomic tasks can maximize classification accuracy and efficiency. The integration of ML in DNA sequence analysis holds significant potential for advancing our understanding of genetics, supporting applications in personalized medicine, evolutionary studies, and bioinformatics. Continued research in this area, particularly with more efficient models and feature engineering techniques, will further enhance the role of ML in genomics and open new pathways for exploring the complexities of DNA.

REFERENCES

[1] Aimin Yang et al , Review on the application of machine learning algorithms in the sequence data mining of DNA, published in Frontiers in Bioengineering and biotechnology on 4th september 2020.

[2] Hemalatha Gunasekaran et al , Analysis of DNA sequence classification using CNN and hybrid models , published in Hindawi Computational and Mathematical Methods in Medicine , Volume 2021 on 16th July 2021.

[3] Amr Ezz El-Din Rashed et al , Sequence Alignment using machine learning using Needleman–Wunsch Algorithm, published in IEEE in 2021.

[4] Firoz Khan1, Cornelius N et al , A Digital DNA Sequencing Engine for Ransomware Detection Using Machine Learning published at IEEE in 2017.

[5] Stephen W. Davies et al , Optimal Structure for Automatic Processing of DNA Sequences published at IEEE on 9th september 1999. Yong-Joon Son et al , Local Alignment of DNA Sequence Based on Deep

[6] Reinforcement Learning published at IEEE EMB in 2021

[7] Robert S. Piecyk et al, Predicting 3D chromatin interactions from DNA sequence using Deep Learning published at Elsevier in 2022.

[8]] H. Gunasekaran, "CNN deep-learning technique to detect Covid-19 using chest X-ray,"Journal of Mechanics of Continua and Mathematical Sciences, vol. 15, 2020

[9] Bosco, G. L., and Di Gangi, M. A. . "Deep learning architectures for DNA sequence classification," in Proceedings of the International Workshop on Fuzzy Logic and Applications in 2016. [10] Chen, L., and Liu, W. "An algorithm for mining frequent patterns in biological sequence," in Proceedings of the 2011 IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences,in 2011

[10] Ernest Bonat, Ph.D., Bishes Rayamajhi, M.S., Apply Machine Learning Algorithms for Genomics Data Classification in the Medium website

[11] Tanisha Pattnaik M.S.Antony Vigil, Aashna Chib, Ayushi Vashisth , Detection of Cloud shadows using deep CNN utilizing spatial and spectral features of landsat imagery

[12] Antony Vigil, Subbiah BharathiDiagnosis of Pulpitis from Dental Panoramic Radiograph Using Histogram of Gradients with Discrete Wavelet Transform and Multilevel Neural Network Techniques.

[13] S. Sri Ram Reddy M.S Antony Vigil, Maneesh Pudhota, Sridhar Gunnam, Kaushik Kumar GP,Predicting CT image from MRI Image using convolutional neural networks.

[14] Sushmita Ghosh Rishikesh Kumar, Dharmin Sodwadia, M.S. Antony Vigil, Increasing Efficiency and Prediction of Heart Disease Using Machine Learning Algorithms.

[15] Harika K and Avula Hanumayamma M.S.Antony Vigil, Logitha Anandan,A Dynamic pricing system for E-scooter based on demand prediction using Neural network.

[16] E.Bhargava Reddy M.S Antony Vigil, K.Vijay Sriram Charan, D.Sai Santosh, Automatic Driver Drowsiness Detection and Accident Prevention System using Image Processing.

[17] Harrish P M.S.Antony Vigil, Selva J,Time Series Modelling and Domain specific predicting air passenger flow traffic using Neural Network

[18] MS Antony Vigil, SR Manoj Raj, Venkata Jagadeesh Reddy, Road Sign Alert and Driver's Drowsiness Alert using Convolutional Neural Network

[19] MS Antony Vigil, PS Meena Kumari, U Soumiya, J Abinaya, P Bhargav Akash, A Robust Approach for Iris and Fingerprint Authentication.