Automated DNA Classification Using Modern Deep Learning Strategies

A Project Report Submitted in the

Partial Fulfillment of the Requirements for the

Award of the Degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

Bugga Vinay 22881A0570

Kotari Swamy 22881A0585

Paladugula Sowmya 22881A05A8

SUPERVISOR

Mr. D Ganesh

Assisstant Professor

Department of Computer Science and Engineering

VARDHAMAN COLLEGE OF ENGINEERING

(AUTONOMOUS)

Affiliated to JNTUH, Approved by AICTE, Accredited by NAAC with A++ Grade, ISO 9001:2015 Certified

Kacharam, Shamshabad, Hyderabad – 501218, Telangana, India

April, 2025

VARDHAMAN COLLEGE OF ENGINEERING
(AUTONOMOUS)
Affiliated to JNTUH, Approved by AICTE, Accredited by NAAC with A++ Grade, ISO 9001:2015 Certified
Kacharam, Shamshabad, Hyderabad – 501218, Telangana, India

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the project titled **DNA CLASSIFICATION USING CNN AND LSTM** is carried out by

Bugga Vinay 22881A0570

Kotari Swamy 22881A0585

Paladugula Sowmya 22881A05A8

in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering** during the year 2024-25.

Signature of the Supervisor
Mr. D Ganesh
Assisstant Professor
Dept of CSE

Signature of the HOD
Dr. Ramesh Karnati
Associate Professor and Head of
Dept of CSE

Project Viva-Voce held on ……

Acknowledgement

The satisfaction that accompanies the successful completion of the task would be put incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crown all the efforts with success.

We wish to express our deep sense of gratitude to **Mr.D Ganesh**, Assisstant Professor and Project Supervisor, Department of Computer Science and Engineering, Vardhaman College of Engineering, for his able guidance and useful suggestions, which helped us in completing the project in time.

We are particularly thankful to **Dr.Ramesh Karnati**, the Head of the Department, Department of Computer Science and Engineering, his guidance, intense support and encouragement, which helped us to mould our project into a successful one.

We show gratitude to our honorable Principal **Dr. J.V.R. Ravindra**, for providing all facilities and support.

We avail this opportunity to express our deep sense of gratitude and heartful thanks to **Dr. Teegala Vijender Reddy**, Chairman and **Sri Teegala Upender Reddy**, Secretary of VCE, for providing a congenial atmosphere to complete this project successfully.

We also thank all the staff members of Computer Science and Engineering(AI&ML) department for their valuable support and generous advice. Finally thanks to all our friends and family members for their continuous support and enthusiastic help.

Bugga Vinay

Kotari Swamy

Paladugula Sowmya

Abstract

The project is focused on the innovative use of a hybrid deep learning model to classify DNA sequences with high accuracy. The model combined the strengths of both Recurrent Residual Convolutional Neural Network (CNN) and Bi-Directional Long Short-Term Memory (LSTM) architectures to effectively handle the complex patterns inherent in biological sequence data. The RRCNN component was specifically employed to extract high-level features from the DNA sequences, capitalizing on its ability to learn spatial hierarchies from data. Meanwhile, the BiLSTM layer was integrated to capture the sequential dependencies and the bidirectional flow of information, which is crucial for understanding the context within the sequences. For the purposes of this study, a comprehensive dataset was compiled from the GenBank database, encompassing a total of 500 DNA sequences. These sequences varied not only in length but also in their respective classifications, providing a robust challenge for the model. Prior to training, the sequences underwent an integer encoding process, converting the nucleotide bases into a numerical format suitable for machine learning algorithms. The dataset was then divided, with a portion allocated for training the hybrid RRCNNBi-LSTM model, and the remainder set aside for testing and validation purposes. Upon training, the model was subjected to an evaluation phase where its performance metrics were calculated. Remarkably, the hybrid model achieved a classification accuracy of 94.5%, a result that surpassed the benchmarks set by other state-of-the-art methods in the field. This outcome not only affirms the efficacy of the hybrid model in the context of DNA sequence classification but also underscores its potential applicability in a variety of bioinformatics tasks. The success of the RRCNN-Bi-LSTM model paves the way for future research and development in the area of computational genomics, where accurate and efficient analysis of DNA sequences is paramount.

Keywords:DNA sequences; Classification; CNN; LSTM; GenBank; bioinformatics.

Table of Contents

List of Tables

List of Figures

Abbreviations

Abbreviation	Description
ML	Machine Learning
DL	Deep Learning
RRCNN	Recurrent Residual Convolutional Neural Network
CNN	Convolutional Neural Network
Bi-LSTM	Bi-Directional Long Short Term Memory
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
A T C G	Adenine, Thymine, Cytosine, Guanine
SVM	Support Vector Machine
KNN	K-Nearest Neighbour
LSTM	Long Short Term Memory
RNN	Recurrent Neural Network
RF	Random Forest

CHAPTER 1

Introduction

Deoxyribonucleic acid, commonly known as DNA, is the hereditary material that carries the genetic blueprint for the structure, operation, proliferation, and survival of all recognized organisms and numerous viruses. DNA is composed of two strands that wind around each other to form a structure known as a double helix, with hydrogen bonds connecting the strands. Each strand is a polymer made up of units called nucleotides[1], each comprising a sugar molecule (deoxyribose), a phosphate group, and one of four types of nitrogen-containing bases: adenine (A), thymine (T), cytosine (C), or guanine (G)[2]. The specific order of these bases constitutes the genetic code, much like how words are formed by arranging letters in a specific sequence to convey meaning. The genetic code is crucial as it directs the synthesis of RNA (ribonucleic acid), which then guides the production of proteins. These proteins are essential for the majority of cellular functions.

1.1 Background

DNA sequences can vary greatly between organisms and even within the same species. These variations contribute to the diversity of life and play important roles in evolution, adaptation, and disease susceptibility. In addition to the primary nucleotide sequence, DNA sequences may have structural features such as loops, hairpins, and other secondary structures that can affect function. Understanding DNA sequences and their variations is fundamental to fields such as genetics, genomics, evolutionary biology, and biotechnology. Advancements in genomic analysis have been propelled by the development of various DNA sequencing methods. DNA sequence classification plays a pivotal role in deciphering the intricate language embedded within the genomic code[3], offering profound insights into the fundamental building blocks of life.

Given that genomics has experienced a profound shift with the introduction of highthroughput sequencing technologies, an unprecedented surge in genomic data has become the norm.

The advent of techniques related to Machine Learning (ML) and Deep Learning (DL) in recent times[3] has ushered in a new era for DNA sequence classification, promising unprecedented accuracy and efficiency in decoding the genomic landscape.

Machine Learning algorithms, encompassing a diverse range of methodologies such as Support Vector Machines (SVM) [4,6,8,9], Random Forests[9],Naive Bayes[9], multinomial Naive Bayes classifier [4] and k-Nearest Neighbors (kNN) [4], have demonstrated efficacy in capturing patterns within genomic sequences. However, the depth and complexity of genomic information necessitate more advanced approaches. Deep Learning, inspired by the structure and function of neural networks,has become an effective instrument for DNA sequence classification. Recurrent Neural Networks (RNNs) and, Convolutional Neural Networks (CNNs)[2,3,5] among other architectures, excel at discerning intricate relationships and dependencies within DNA sequences, enabling a deeper understanding of genomic information.

Nucleotides as Building Blocks: Nucleotides are the fundamental units that compose DNA. Each nucleotide is a composite of three components: a deoxyribose sugar, which is a pentose sugar providing the structural framework of the DNA strand; a phosphate group that connects the sugars to forge the DNA backbone; and a nitrogenous base. DNA harbors four distinct nitrogenous bases: adenine (A), thymine (T), cytosine (C), and guanine (G)[2]. The order of these bases along the DNA strand encodes the hereditary informati

on.**Base Pairing Mechanism:** Within the iconic double helix of DNA, nitrogenous bases pair up in a highly specific manner: adenine (A) pairs with

thymine (T) via a duo of hydrogen bonds, and cytosine (C) pairs with guanine through a trio of hydrogen bonds. This precise pairing of bases ensures the

DNA strands are complementary, a feature that is critical for DNA replication and various cellular functions.

The Double Helix Configuration: The double helix as shown in Fig.1.1 is the characteristic structure of a DNA molecule, comprising two intertwined strands. This configuration, first described by James Watson and Francis Crick in 1953, confers stability to the DNA molecule and is vital for its replication and transcription processes.

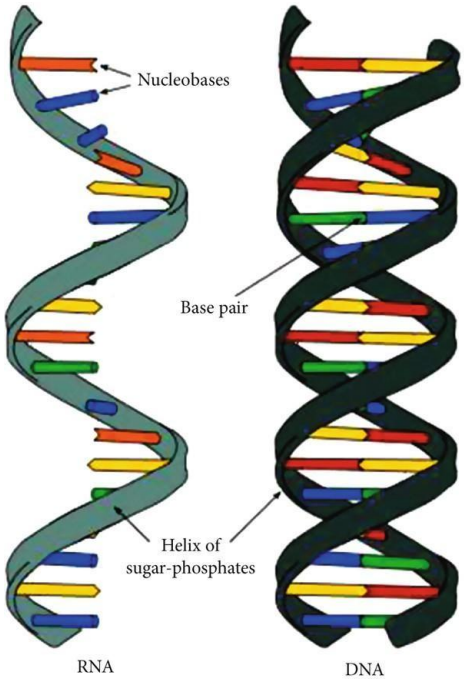

Figure 1.1: Structure of RNA-DNA[2]

Deciphering the Genetic Code: The genetic code is the language by which DNA stores information, consisting of nucleotide sequences that instruct the synthesis of proteins and regulate cellular activities. Triplets of nucleotides, known as codons, correspond to specific amino acids or signal the initiation and termination of protein synthesis.

Genetic Diversity Through Variation and Mutation: Variations in DNA sequences among individuals and species are largely the result of mutations, or

alterations in the nucleotide sequence. These genetic differences contribute to the rich tapestry of biological diversity and influence characteristics such as physical traits, behaviors, and disease susceptibilities.

Functional Elements Within DNA: Beyond coding for proteins, DNA encompasses a variety of elements with specialized functions. These include regulatory sequences that control gene expression, non-coding RNAs that are transcribed but not translated into proteins and fulfill diverse roles, and repetitive elements that are present in multiple copies within the genome, influencing its structure and functionality.

Advancements in DNA Sequencing: Modern DNA sequencing technologies enable researchers to ascertain the precise nucleotide order within DNA, providing invaluable insight for studying genetic variation, elucidating gene function, diagnosing hereditary conditions, and applying genetics in numerous scientific and medical fields. DNA sequences are the bedrock of genetic information, underpinning the biological development and functioning of all living organisms. Unraveling the mysteries of DNA sequences is crucial for advancing our comprehension of life sciences and addressing a myriad of scientific and health-related challenges.

1.1.1 Types of issues in DNA sequence classification

DNA sequence classification poses several challenges that researchers and practitioners in genomics and bioinformatics must address. These challenges encompass various aspects, ranging from the inherent complexity of genomic data to the intricacies of computational methodologies. Here are some types of issues commonly encountered in DNA sequence classification:

1. Genomic Data Complexity: Genomic sequences consist of various elements, comprising regulatory components, non-coding sections, and coding portions, each with unique characteristics.

2. Algorithmic Challenges: Selecting relevant features from raw DNA sequences is a non-trivial task, and the choice of features greatly influences the performance of classification algorithms.
3. Biological Variation: The presence of single nucleotide polymorphisms (SNPs) and other genetic variations introduces complexity in classifying DNA sequences, especially in the context of disease association studies.
4. Dynamic Nature of Genomic Data: Genomic data evolves over time, requiring adaptive models that can effectively handle changes in the genomic landscape.
5. Class Imbalance: Imbalances in the distribution of classes within DNA sequences can impact the training of classification models, leading to biased results.
6. Scalability: The increasing volume of genomic data requires scalable approaches to process and analyze large datasets efficiently.
7. Real-time Adaptability: The ability of classification models to adapt to realtime changes in genomic data, especially in clinical settings, is crucial for timely decision-making.

1.2 Motivation

The motivation behind undertaking a DNA sequence classification project stems from the profound impact it can have on various fields of study and practical applications. DNA sequencing refers to the technique used to accurately establish the sequence of nucleotides within a strand of DNA, providing invaluable insights into genetic information and structure. By classifying DNA sequences, researchers and scientists aim to unlock a myriad of discoveries, such as identifying genetic variations associated with diseases, understanding evolutionary relationships between species, and developing personalized medicine.

In the realm of healthcare, DNA sequence classification plays a pivotal role in advancing precision medicine. Through the analysis of a person's unique DNA

sequence, healthcare providers can customize therapeutic strategies and drug prescriptions to align with the individual's genetic profile, resulting in more precise and personalized medical care. Moreover, it enables the identification of genetic predispositions to diseases, allowing for proactive measures such as early detection and prevention strategies.

Another area where DNA sequence classification is crucial is in agriculture and crop improvement. By analyzing the genetic diversity within plant species, researchers can identify desirable traits, such as resistance to pests or tolerance to environmental stresses. This knowledge enables the development of genetically modified crops that can contribute to higher yields, improved nutritional content, and enhanced sustainability.

Furthermore, DNA sequence classification has significant implications for evolutionary biology and ecological research. By comparing DNA sequences across different organisms, scientists can trace evolutionary lineages, understand the relationships between species, and gain insights into the mechanisms driving biodiversity. This knowledge is vital for conservation efforts, as it helps identify endangered species, track population dynamics, and design effective conservation strategies.

Overall, the motivation for undertaking a DNA sequence classification project lies in its potential to revolutionize multiple fields, including healthcare, agriculture, and ecological studies. By harnessing the power of DNA sequencing, researchers can unravel the mysteries of genetics, unveil novel insights, and pave the way for transformative advancements that have a profound impact on human well-being and our understanding of the natural world.

1.3 Scope

The scope of a DNA sequence classification project encompasses the application of advanced computational techniques to analyze and categorize DNA sequences based on their characteristics and attributes. With the exponential growth of DNA sequencing data, there is a need to develop efficient methods for

organizing and classifying this vast amount of information. Such projects aim to employ machine learning algorithms, bioinformatics tools, and data analysis techniques to classify DNA sequences into different groups or cate- gories.

The classification can be based on various factors, including the species of origin, genetic mutations or variations, functional elements, or specific patterns within the sequences. By accurately classifying DNA sequences, researchers can gain valuable insights into genetic diversity, evolutionary relationships, and the functional significance of specific genetic elements. This knowledge can have farreaching implications across multiple domains, including medicine, agriculture, ecology, and biotechnology.

1.4 Objectives

1. **Develop an accurate classification model:** The primary objective is to build a robust and accurate classification model using RRCNN and Bi-LSTM architectures. This involves training the model on a labeled dataset of DNA sequences and optimizing its parameters to achieve high classification performance.

2. **Improve prediction accuracy:** Aim to enhance the accuracy of DNA sequence classification compared to existing methods or baseline models. Highlight the potential of RRCNN and Bi-LSTM models to capture complex sequence patterns, long-range dependencies, and important features that can lead to improved prediction accuracy.

3. **Disease Gene Identification:** In the context of medical research, a key objective is to identify disease-causing genes or genetic variations. By classifying DNA sequences from individuals with specific diseases or traits, researchers can identify patterns or mutations associated with those

conditions, leading to improved diagnosis, risk assessment, and potential therapeutic targets.

4. **Investigate sequence-level representations:** Explore the ability of RRCNN and Bi-LSTM models to learn and extract meaningful sequencelevel representations from DNA sequences. This objective involves analyzing the learned representations to gain insights into the discriminative features and patterns that contribute to the classification performance.

1.5 Expected Deliverables

Trained classification model: The primary deliverable is a fully trained RRCNN and Bi-LSTM model for DNA sequence classification. This model should be capable of accurately classifying input DNA sequences into predefined categories or classes based on the training data.

Classification Software or Tools: The development of user-friendly software or tools for DNA sequence classification is another expected deliverable. These tools can incorporate the developed algorithms and models, providing researchers with a streamlined interface to classify and analyze DNA sequences.

Assessment Criteria for Model Efficacy: The study could include a collection of assessment criteria to gauge the effectiveness of the classification models or algorithms. Such criteria, encompassing accuracy, precision, recall, among other pertinent measures, allow for the comparative analysis of various techniques and the determination of the classification strategy's success.

Research Publications: A DNA sequence classification project often leads to research publications in scientific journals or conference proceedings. These publications document the project's methodology, findings, and insights, contributing to the scientific community's knowledge base and facilitating further research in the field.

The objective is to present a detailed summary of the findings and advancements resulting from our investigation into DNA sequence classification via the integration of RRCNN and Bi-LSTM models. The outputs of this project highlight the efficacy, comprehensibility, and practical uses of the model, affirming the significance and influence of the study.

CHAPTER 2

Literature Survey 2.1

Overview of Machine Learning and Deep Learning

Machine Learning (ML) and Deep Learning (DL) represent powerful computational approaches that have significantly advanced our ability to analyze and interpret genomic data.ML algorithms, such as k-Nearest Neighbors (KNN), Random Forests, and Support Vector Machines (SVM) have been applied to DNA sequence classification.These algorithms are effective in capturing patterns within genomic data[7], making them valuable tools for tasks like gene prediction and disease classification.

ML techniques often involve the extraction and selection of relevant features from DNA sequences.Features may include sequence motifs, composition, and various statistical properties that aid in distinguishing between different classes of DNA sequences.Traditional ML approaches may struggle to capture intricate dependencies and hierarchical structures within DNA sequences, limiting their performance on complex genomic tasks.

Deep Learning, particularly neural networks, has gained prominence in genomics because of its capacity for self-learning hierarchical representations from data. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs)[2,3,5] are commonly used for DNA sequence classification.CNNs are effective in recognizing sequence motifs and local patterns within DNA sequences, making them well-suited for tasks like transcription factor binding site prediction.

The temporal character of DNA sequences presents issues for RNNs and their variants, such as Long Short-Term Memory (LSTM)[2] networks, which are meant to capture long-range dependencies in sequential data. End-to-end learning—in which the model learns to extract pertinent features straight from unprocessed DNA sequences is frequently made possible by DL models, which reduces the requirement for human feature engineering. DL models, while powerful, can be complex and may lack interpretability. Understanding how these models arrive at specific classifications is an ongoing challenge in the field.

In summary, ML and DL techniques have significantly advanced DNA sequence classification, offering a spectrum of tools suitable for various genomic tasks. The particulars of the data and the nature of the current categorization problem often influence the choice of approach. Novel designs and techniques are still being investigated in order to improve the precision and interpretability of DNA sequence categorization models.

2.1.1 ML Approaches:

Algorithms for machine learning (ML) provide a special method for resolving complicated issues that are used in several real-time applications. Machine learning algorithms' primary goal is to tackle challenging problems and enhance performance through experience and training. Through experience and cheap computing expenses, the machine learning algorithms automatically build and manage machines. The machine learning (ML) algorithms are categorised into supervised, unsupervised, and reinforcement learning.

Supervised Learning:

The supervised machine learning(as shown in Fig.2.1) algorithms learn by using the labelled training data. They look over the training set and propose a function that could be used to map new samples of data. Supervised learning falls into two categories: regression and classification. The finite dataset for the classification category can be binary (anomaly detection) or multitudinous (speech and face recognition). Regression analysis is a commonly used statistical technique for future event prediction. It involves identifying the relationship between one dependent variable and one or more independent variables.

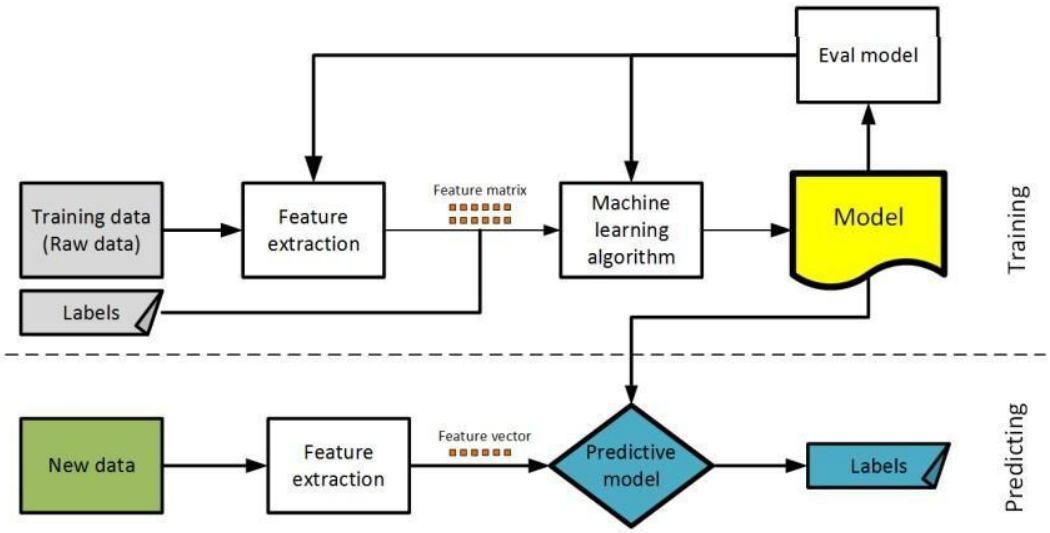

Figure 2.1: Supervised Learning Workflow[16]

Unsupervised Learning:

All of the input data and samples used in unsupervised learning are labelless(as shown in Fig.2.2). In this learning process, the trained model won't rely on inputoutput classes. Density estimation, clustering, and dimensionality reduction are three definitions of the learning process. When organizing data for mathematical and statistical problems, the clustering method is employed. Unexpected situations, including unclassified data from any source, can benefit from this approach.

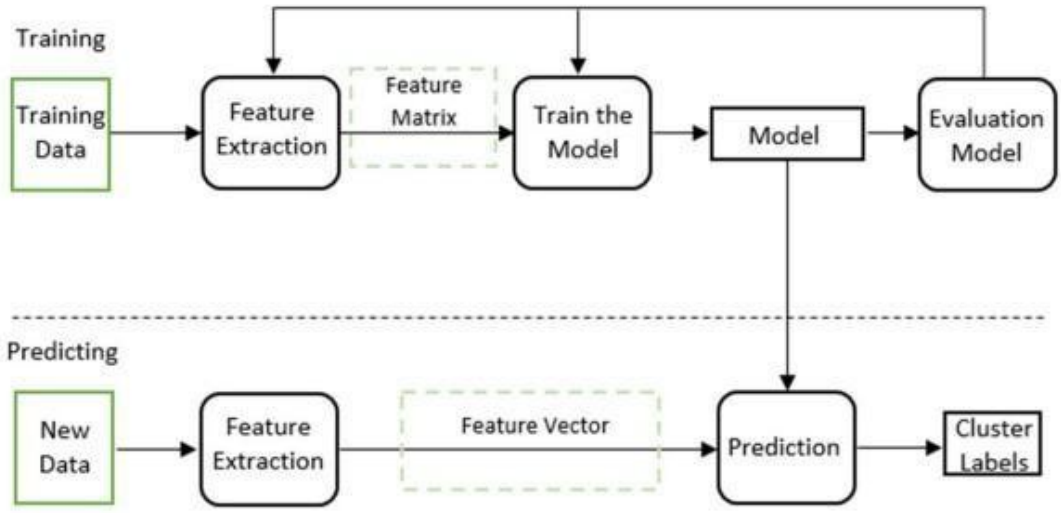

Figure 2.2: Unsupervised Learning Workflow[17]

Reinforcement Learning:

Between supervised learning and unsupervised learning without label information is reinforcement learning (RL)(as shown in Fig.2.3). Reward value for wise decision-making and optimizing rewards is linked to reinforcement learning. Policy search and value function approximation are the two ways to go about this.

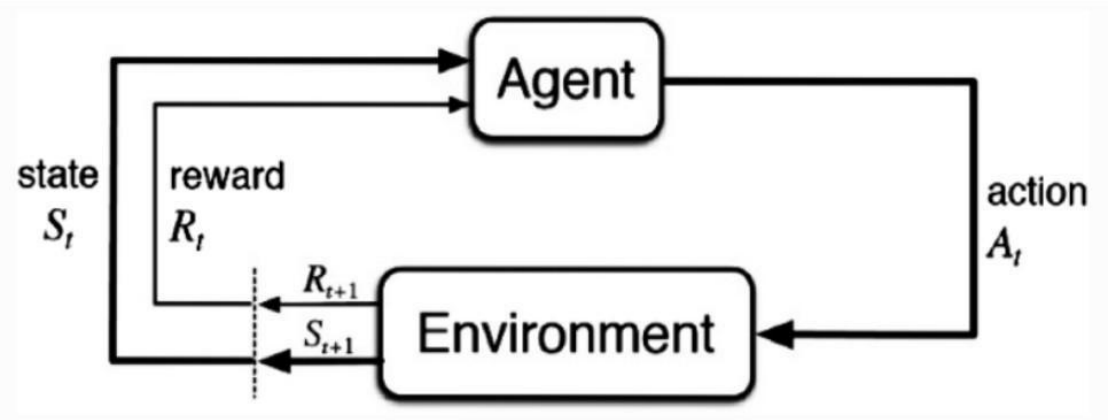

Figure 2.3: Reinforcement Learning Workflow[18]

These detection algorithms frequently gain by combining different strategies, creating hybrid models that combine the best features of deep learning-based and conventional methods. Research on creating more resilient and adaptable detection techniques is being conducted as the field of deep fake detection advances, in an effort to combat the dynamic nature of creating misleading material.

2.1.2 DL Approaches:

Multi-layered neural networks are used in deep learning (DL), a subset of machine learning techniques, to extract data and identify patterns. DL has made significant strides in handling large datasets and complex operations. Recurrent neural networks (RNNs) and Convolutional neural networks (CNNs) are two widely used deep learning architectures.

CNNs perform well in hierarchical feature extraction and are frequently used in image-related applications. Because RNNs are built for sequential data, they can be used for applications such as natural language processing. Despite their strength, deep learning techniques need large amounts of computing power and a large number of training datasets. DL techniques have shown efficacy in domains like

natural language processing, autonomous systems, and image and audio recognition, despite their complexity.

Convolutional Neural Networks (CNNs): Convolutional Neural Networks (CNNs)(as shown in Fig.2.4) distinguish themselves from conventional Artificial Neural Networks (ANNs) by being particularly effective in recognizing patterns. To produce feature maps with the help of image-specific features, CNNs incorporate various neural network layers known as convolution layers and pooling layers. The higher-level performance of CNNs as compared to ANNs can be observed in various computer vision Applications and tasks[20].

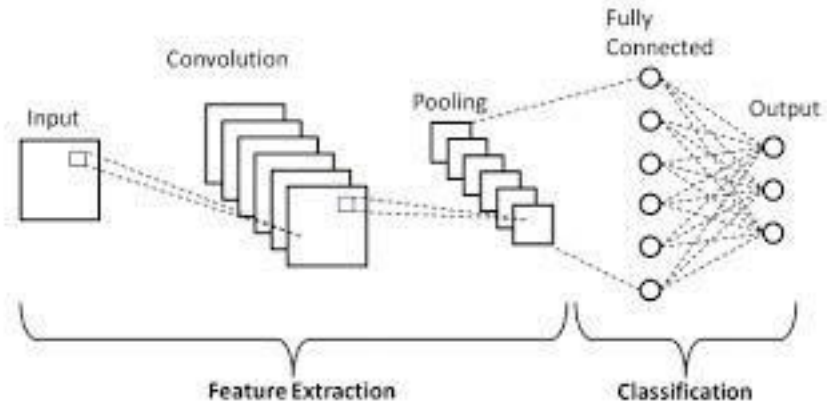

Figure 2.4: Basic CNN Structure[20]

Recurrent neural networks: RNNs are a widely used and well-known method in the field of DL. The main applications of a complicated method like
RNNs can be seen in speech and text processing tasks and Natural Language Processing tasks. The hidden layers of a typical RNN(as shown in Fig.2.5) deals with sequential data as opposed to CNNs which deal with image data. This feature is crucial for many applications as the data sequence's intrinsic structure offers helpful information. With x denoting the input layer, y the output layer, and h the hidden layer, the RNN may therefore be thought of as a unit of short-term memory.[21]

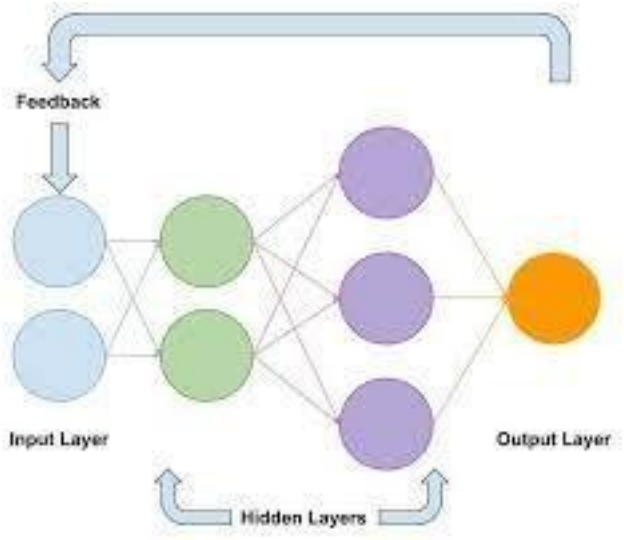

Figure 2.5: Basic RNN Structure[21]

2.2 Literature Review on Existing Methods

The initiation of this project involved an exhaustive review of contemporary literature focusing on DNA sequence classification utilizing Deep Learning and Machine Learning methodologies. Recent years have witnessed a surge in research focused on leveraging deep learning (DL) and machine learning (ML) techniques for DNA sequence classification, revolutionizing our understanding of genomics. In the realm of ML, traditional algorithms likely k-Nearest Neighbors (KNN), Random Forests, and Support Vector Machines (SVM) have been instrumental in discerning patterns within DNA sequences, contributing to tasks like gene prediction and disease classification.

These methods, while effective, often require manual feature engineering and struggle with capturing intricate dependencies present in complex genomic data. The advent of DL has addressed these challenges, with designs of neural networks, such as recurrent and convolutional neural networks proving particularly adept at automatically extracting hierarchical representations from raw DNA sequences. Deep learning models, by learning features directly from data, have significantly enhanced the accuracy and efficiency of DNA sequence classification[10], offering promise for unraveling the intricate code embedded within the genome.

Alejandro et al. introduced the application of deep learning for SARSCoV-2 detection and classification [11]. The author combined AI techniques with deep learning approaches to find the representation of SARS-CoV-2 genomic sequences. 553,00 sequences are used to train the classifier. The model displayed 100% specificity and a promising 98.73% accuracy. Table 2.1 discusses about the Summary of ML Algorithms used for DNA sequence classification.

Table 2.1: Summary of ML Algorithms used for DNA sequence classification

Model	**Reference**	**Summary**
Na¨ıve Bayes (NB)	[9]	The classifier will give zero probability to any category variable in the test set that does not appear in the training dataset, making prediction impossible.
Random Forest	[9]	Random Forest (RF) algorithm is employed as one of the classifiers for handling the complexity of DNA sequences and contributes to tasks such as gene prediction and disease classification.
SVM	[4],[6],[8],[9]	The SVM algorithm was employed for DNA sequence classification to sequence a human dataset's DNA.Drug manufacturers may be able to target particular groupings of people who are related genetically by having an understanding of each patient's distinct genetic profile.
KNN	[4]	KNN algorithm is used for DNA sequence classification as it is faster than other methods as there need no training before generating predictions.
Multinomial Na¨ıve Bayes	[4]	Multinomial Naive Bayes classifier is used to classify DNA sequences using label and k-mer encoding. Achieving good accuracy.

Beinke et al. proposed comparing ML methods with and without feature extraction [12]. The three algorithms that the author suggested are CNN, DNN, and N-gram probabilistic models. The author used randomly generated subsequences of DNA and a technique for feature extraction based on distance in the DNA sequence. Additionally, the models were assessed by the author using a variety of datasets, including COVID-19, AIDS, hepatitis C, and influenza. They found that all three of the models had accuracy levels above 99%.

Khanh Le et al. used the method of virus origin identification [13]. This work demonstrated how to identify the DNA basis of altered viruses. The authors employed an extreme gradient boosting technique as part of a hybrid strategy. This approach is used to identify when a DNA sequence begins to replicate. The program, utilizing the XGBoost classifier, achieved an accuracy of 89.51%.

M S Antony Vigil et al. [14]. The authors of this article assessed a variety of machine learning methods for the DNA sequencing problem using a human dataset. These methods included XGB Classifier, Multilayer Perceptron, LSTM, SVM, Adaboost, Naive Bayes,Random Forest Classifier, and CNN.

Aleshinloye Yusuf Abass et al. [15]. The authors of this study retrieved exons from several prostate gene sequences that were used in tests.A bi-LSTM model was created by combining one-hot encoding for class labels with a k-mer encoding technique for DNA sequences. The model predicts with a 91% validation accuracy and a 95% training accuracy, in that order.

The paper titled "Classification of K-mer DNA Sequences" Based on Based Vector Representations by Umit et al.[19],Four different approaches are used to represent DNA sequences from the three different benchmark datasets (splice, promoter, and H3): one-hot based on random and default dictionary, Voss, and dna2vec. The impact of representation on classification is assessed using a deep neural network. A novel representation is introduced to illustrate the impact of employing randomly ordered equidistant vectors in conjunction with CNN for classification: the random dictionary technique.

Utilizing the K-nearest neighbor (KNN) fusion technique (SVM-KNN) and support vector machine (SVM), Zhikang Bao et al. suggested a classification model [6].According to the experimental results, DNA sequence classification is more successful when compared to the KNN and SVM algorithm model, the SVM-KNN algorithm model's classification accuracy is significantly higher.

Table 2.2 discusses about the Summary of DL Algorithms used for DNA sequence classification.

These investigations collectively underscore the importance of employing ensemble techniques, feature selection, and the integration of diverse feature sets in crafting resilient and precise DNA sequence classification models. While attaining commendable accuracies, these studies also underscore the ongoing necessity for exploration to tackle the dynamic nature of genomic data, scalability challenges, and the real-time adaptability of classification models.

Table 2.2: Summary of DL Algorithms used for DNA sequence classification

Model	Reference	Summary
LSTM	[2]	Recurrent neural networks (RNNs) have a type called Long Short-Term Memory (LSTM), is employed in DNA sequence classification, effectively capturing long-range dependencies within genomic data.
Bi-directional LSTM	[2]	An expansion of LSTM is called bi-directional long short-term memory (Bi-LSTM), is utilized for DNA sequence classification, enabling the model to capture dependencies in both forward and backward directions within genomic data.
CNN	[2],[3],[5]	The CNN architecture includes an embedding layer, convolutional layer followed by global max pooling and then connected to a fully connected layer, and output layer, demonstrating the effectiveness of deep learning in DNA sequence classification.

2.3 Tools and Packages used

In DNA sequence classification, researchers typically employ a variety of tools and packages for different stages of the workflow. Here are some commonly used tools and frameworks in this context:

1. Biopython: Biopython is a widely used open-source application programming interface (API) for routine scripts for typical bioinformatics chores as well as software development in the field. One can get the source code, documentation, and mailing lists on the homepage of www.biopython.org. The fasta-formatted DNA sequences are read using the Biopython pack- age[2].
2. TensorFlow: For the purpose of creating and refining neural network models, TensorFlow is an open-source machine learning framework.

Because of its scalability and versatility, deep learning architectures, such as recurrent neural networks, are frequently implemented and experimented with using it.

3. Keras is a high-level API for neural networks that operates on top of TensorFlow-like frameworks. It is an effective tool for quick experimentation and prototyping since it offers an easy-to-use interface for creating and training neural network models.

4. Scikit-learn is a Python machine learning library that is quite flexible. It provides a complete set of tools for machine learning practitioners and covers a wide range of machine learning techniques. It is especially helpful for tasks like feature selection, data preprocessing, and model validation.

5. Pandas and NumPy: NumPy and Pandas are two Python libraries that are widely used for numerical calculations and data manipulation. Within the framework of this study, these libraries help manage and prepare datasets so that machine learning models can be trained and assessed.

CHAPTER 3

Methodology

3.1 Problem description

The categorization of DNA sequences is essential for many applications in the fields of bioinformatics and computational biology, such as illness detection, gene prediction, and evolutionary research. Customized features and domainspecific expertise are frequently used in traditional methods, which may not be scalable or generalizable. Deep learning methods have become highly effective tools for automatically extracting features and classifying raw DNA sequences in recent years. In order to classify DNA sequences, this thesis investigates the use of a hybrid deep learning architecture that combines Recurrent Residual Convolutional Neural Networks (RRCNN) and Bidirectional Long Short-Term Memory (Bi-LSTM). In order to capture both local and long-range relationships in DNA sequences, the suggested approach tries to combine the sequential modeling capabilities of LSTMs with the hierarchical representation learning capabilities of CNNs. In order to evaluate the performance of the suggested RRCNN-Bi-LSTM model in relation to current state-of-the-art techniques, the project provides a thorough description of the problem, methodology, experimental setup, and evaluation metrics. It also goes over possible uses, restrictions, and future possibilities for deep learning architectures in DNA sequence categorization research.

Key aspects of the problem:

- **Sequence Representation:** DNA sequences are represented as strings of nucleotides, which are A, T, C, and G. The difficult part is figuring out how to express these sequences numerically so that deep learning models can use them.

- **Feature Extraction:** Biological information is commonly used to create handcrafted characteristics in traditional approaches. RRCNN and

Bi-LSTM are two examples of deep learning models that attempt to automatically extract pertinent features directly from unprocessed DNA sequences, possibly capturing complex connections and patterns.

- **Hierarchical Representation Learning:** RRCNN learns hierarchical features from local patterns in DNA sequences by using convolutional layers. Conversely, sequential information and long-range dependencies are captured by Bi-LSTM. The model is able to represent both the global and local properties of DNA sequences by combining the two designs.

- **Model Training and Optimization:** Thoroughly optimizing hyperparameters and utilizing extensive annotated datasets are necessary for training deep learning models in DNA sequence categorization. To enhance model performance, strategies like data augmentation and transfer learning may be used.

- **Evaluation Metrics:** It is necessary to assess the efficacy of the suggested RRCNN-Bi-LSTM model using suitable measures, including area under the receiver operating characteristic curve (AUC-ROC), accuracy, precision, recall, and F1-score.

- **Interpretability:** It is important to comprehend the model's prediction process, particularly in biomedical applications where interpretability is critical to understanding biological processes. It is necessary to investigate methods for deciphering deep learning models applied to DNA sequences.

- **Application Domains:** Applications for classifying DNA sequences can be found in many fields, such as drug development, disease diagnostics, gene prediction, and evolutionary biology. It is essential to comprehend the unique demands and obstacles in every domain while creating efficient classification models.

- **Generalization and Robustness:** For real-world applications, it is essential to make sure the trained model is resilient to noise, mutations, and variations in DNA sequences, and that it generalizes well to new data.

3.2 Proposed Methodology

The proposed methodology for DNA sequence classification using Recurrent Residual Convolutional Neural Network and Bidirectional Long ShortTerm Memory

(Bi-LSTM) can be structured as follows:

3.2.1 Load Dataset

The initial step in the journey of classifying DNA sequences is to acquire the relevant dataset that contains the sequences of interest. This dataset could be sourced from a variety of origins, including but not limited to, proprietary research collections or publicly accessible genomic databases. Within this dataset(as shown in Fig.3.1), the fundamental units are the nucleotide sequences, which consist of the bases adenine (A), cytosine (C), guanine (G), and thymine (T), the building blocks that make up the DNA double helix.

The format of the data typically includes strings or series of these letters, each string corresponding to a DNA sequence. Accompanying these sequences, there is often a set of class labels. These labels serve to categorize each DNA sequence according to its biological function or role within the genome, such as a gene, a promoter region, a regulatory element, and so forth. These annotations are essential for supervising the learning process of the classification model, as they provide the ground truth against which the model's predictions will be compared.

In preparation for the classification task, these sequences and their associated labels must be carefully curated to ensure accuracy and consistency. The quality and relevance of the dataset are paramount, as they directly influence the potential success of the subsequent modeling and classification efforts.

3.2.2 Data Encoding

In the realm of machine learning, raw DNA sequences composed of the characters A, C, G, and T are not inherently understood by algorithms. Consequently, a critical preprocessing step involves converting these sequences into a numerical representation that machine learning models can process.

Figure 3.1: Sample of DNA sequence dataset Here are some common encoding strategies:

One-Hot Encoding: This method transforms each nucleotide into a binary vector with a length equal to the number of possible nucleotide types. In the case of DNA, this results in a vector of length four. Each vector has a single '1' corresponding to the nucleotide it represents, and '0's in all other positions. For example, cytosine (C) would be encoded as [0, 1, 0, 0], and adenine (A) as [1, 0, 0, 0]. This creates a clear, distinct representation for each nucleotide without any implied ordinal relationship.

Integer Encoding: Alternatively, integer encoding assigns a unique integer to each type of nucleotide. This is a more compact representation than onehot encoding, as it reduces each nucleotide to a single number. For example, adenine (A) might be encoded as 1, cytosine (C) as 2, guanine (G) as 3, and thymine (T) as

4. This method is straightforward but introduces an artificial ordinal relationship between the nucleotides, which does not exist in biological terms.

For the purposes of our study, we have chosen to utilize integer encoding. This decision was guided by the balance between simplicity and the reduced

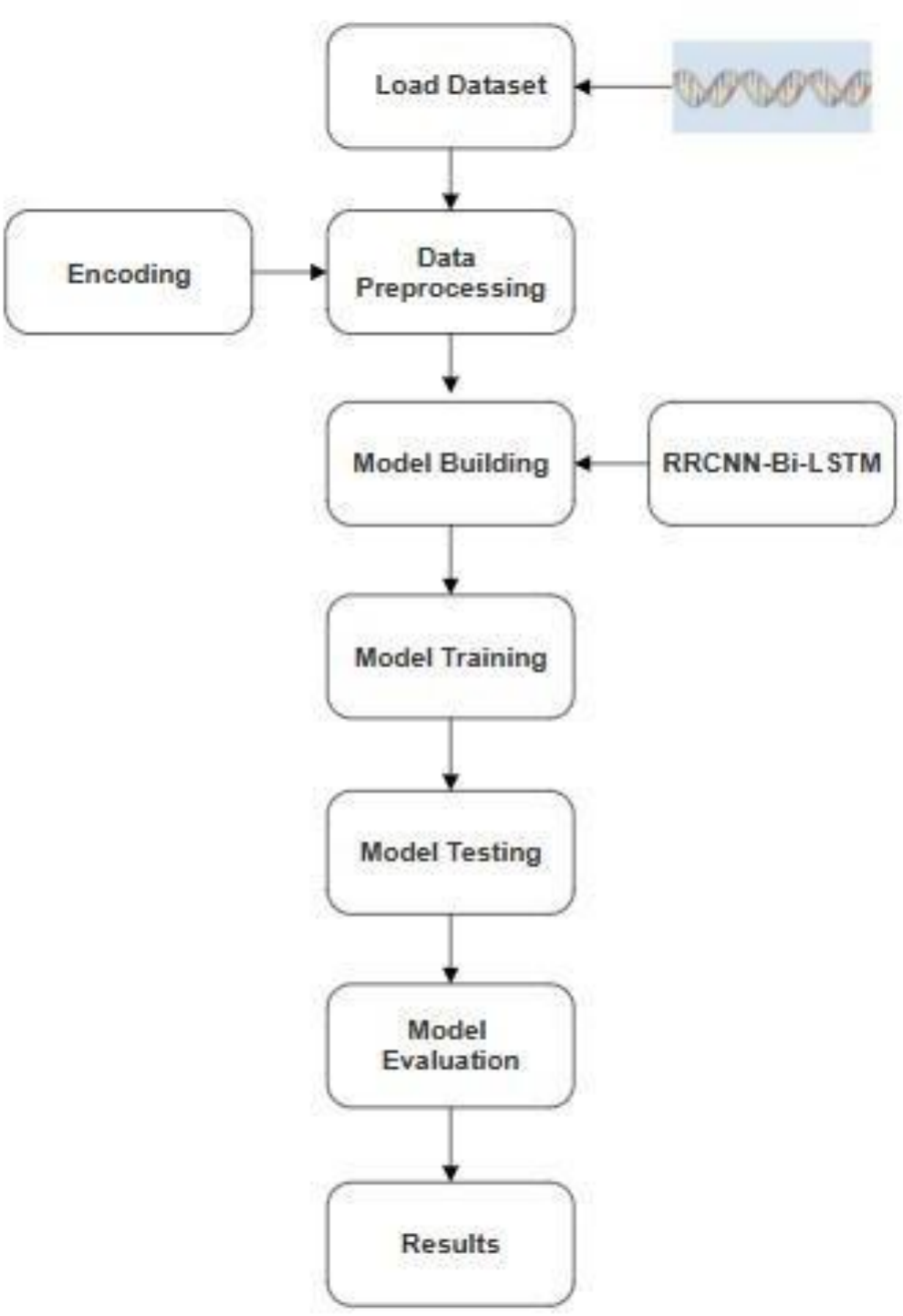

Figure 3.2: Proposed Methadology

computational overhead compared to one-hot encoding. By representing the nucleotides as integers, we can efficiently prepare our dataset for subsequent

processing by the RRCNN and Bi-LSTM models, allowing them to learn and predict the classification of DNA sequences based on their encoded numerical patterns.

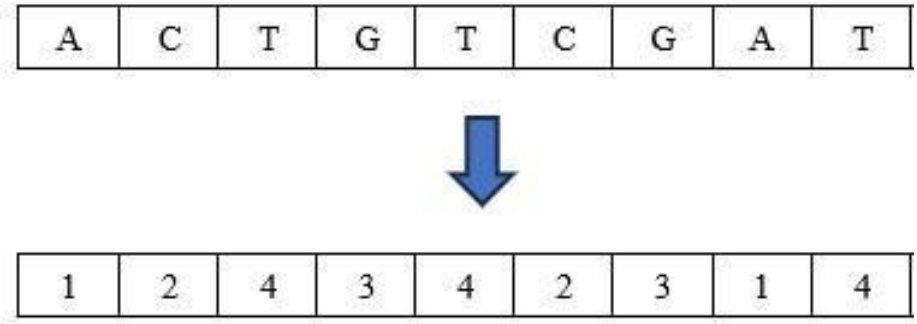

Figure 3.3: Encoding DNA sequence

3.2.3 Data Preprocessing:

To optimize the performance and training efficiency of the model, the preprocessed encoded DNA sequences may require additional refinement steps. These steps are crucial for ensuring that the model is trained on data that is consistent and representative of the underlying biological structures. The enhancement process typically includes:

Normalization: It is essential to scale the numerical values of the encoded sequences so that each sequence has an equal opportunity to influence the training process. Normalization adjusts the data to a common scale, such as between 0 and 1 or -1 to 1, which helps in preventing any single sequence from disproportionately affecting the model due to its value range.

Imputation of Missing Values: Occasionally, sequences may have gaps or missing nucleotides, which can disrupt the training process. To address this, imputation techniques are employed. Methods like substituting missing nucleotides with the mean or median values of the surrounding data or completely removing sequences with missing values help maintain the integrity of the dataset.

Consistency in Sequence Lengths: DNA sequences can vary widely in length, which poses a challenge for many machine learning models that require input data of uniform size. To standardize sequence lengths, two common strategies are used. One approach is to trim longer sequences to a predefined length, ensuring that all sequences fit within the model's input parameters. Alternatively, shorter

sequences can be padded with a specific character, often ”0,” to extend them to the necessary length. This padding ensures that every sequence is treated equally during the model training phase without
introducing significant bias.

By implementing these preprocessing steps, the model can be trained on data that is more coherent and aligned with the algorithmic requirements, leading to improved learning outcomes and a more robust classification model.

3.2.4 Model Building (RRCNN-Bi-LSTM)

At this juncture, the focus is on constructing a sophisticated deep learning architecture tailored to the task of DNA sequence classification. The model in question, known as RRCNN-Bi-LSTM, is a fusion of two advanced neural network designs that leverage their respective strengths to analyze and interpret the complex patterns in DNA sequences.

RRCNN, or recurrent residual convolutional neural network: The RRCNN, or Recurrent Residual Convolutional Neural Network(as shown in Fig.3.4), is adept at detecting local patterns within the encoded sequence data. Its architecture is characterized by convolutional layers that are designed to identify features in short segments of the input sequences. These layers act as a series of filters that highlight various aspects of the sequence data, such as the presence of specific nucleotide arrangements. The recurrent layers, which often take the form of Long Short-Term Memory (LSTM) units, are then applied to understand the temporal or sequential relationships between these features. This is crucial because the biological significance of a DNA sequence is not only determined by the individual nucleotides but also by their context within the larger sequence.

Augmenting this, the model incorporates Bidirectional Long Short-Term Memory (Bi-LSTM) networks(as shown in Fig.3.5). LSTMs are a specialized type of recurrent neural network capable of learning long-range dependencies within data sequences. The bidirectional aspect of Bi-LSTMs means that the sequence data is processed in both forward and reverse directions, granting the model access to all available contextual information. This bidirectionality is particularly

beneficial for DNA sequences, where the interpretation of a given segment may depend on nucleotides that appear both before and after it.

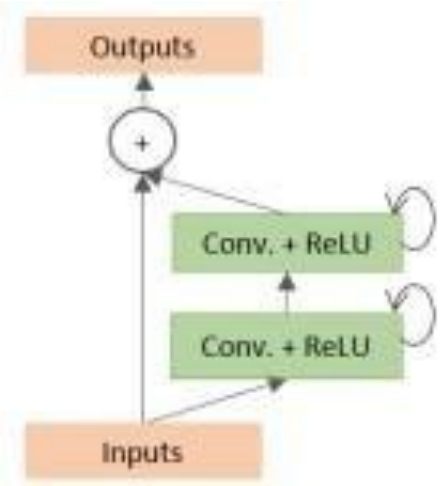

Figure 3.4: RRCNN architecture[22]

By combining RRCNN and Bi-LSTM, the model is equipped to meticulously analyze DNA sequences, capturing both the local features and the broader sequential dependencies. This comprehensive approach to sequence analysis is expected to enhance the model's ability to accurately classify DNA sequences, making it a powerful tool for genomic research and applications.

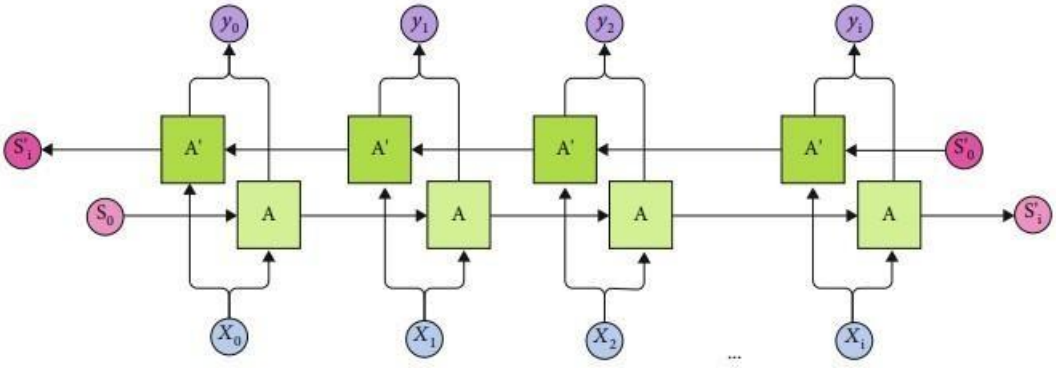

Figure 3.5: Bi-directional LSTM architecture[2]

3.2.5 Model Training

The constructed RRCNN-Bi-LSTM model is trained using the prepared dataset. Here's a breakdown of the training process: The dataset is typically split into two parts: training set and validation set. The training set is used to train the model,

while the validation set is used to monitor the model's performance during training and prevent overfitting.

- During training(70%), the model is presented with batches of encoded sequences and their corresponding class labels. The model adjusts its internal parameters (weights and biases) to minimize the difference between its predicted classifications and the actual labels. The batch size used is 32.

- This process is repeated through multiple epochs(5) (iterations over the entire training set).

- The validation set(30%) is used periodically to evaluate the model's performance on unseen data. If the model's performance on the validation set starts to decline, it might be a sign of overfitting, and adjustments to the model architecture or training parameters might be necessary.

3.2.6 Model Testing

After training is finished, a different test dataset is used to assess the model's performance. This dataset is essential since it guarantees an objective evaluation of the model's applicability to previously undiscovered data. There was no reason to use the test dataset in any way during training.

3.2.7 Model Evaluation

When evaluating the effectiveness of a model in classifying DNA sequences, several statistical metrics are employed to capture different aspects of performance:

Accuracy: This metric reflects the overall correctness of the model by calculating the ratio of the number of sequences it correctly identifies to the total number of sequences in the test dataset. A high accuracy indicates that the model is effective at distinguishing between different classes of sequences.

Precision: Precision measures the reliability of the model in labeling sequences. It is the proportion of sequences that the model classifies into a

particular category that actually belong to that category. High precision implies that when the model predicts a sequence as belonging to a class, it is likely to be correct.

Recall (Sensitivity): Recall assesses the model's ability to identify and correctly classify all relevant instances within a class. It is the fraction of actual members of a class that are correctly identified by the model. A model with high recall efficiently captures most of the sequences from a class without leaving many out.

F1 Score: The F1 score is a balanced measure that considers both precision and recall in a single metric by taking their harmonic mean. It is particularly useful when the class distribution is uneven or when false positives and false negatives carry different costs. A high F1 score indicates that the model has a robust balance between precision and recall, making fewer mistakes in both false positives and false negatives.

These metrics collectively provide a comprehensive picture of the model's classification capabilities. By considering all these measures, researchers can gain insights into the strengths and weaknesses of their models, guiding improvements and ensuring that the models perform well across a range of scenarios in DNA sequence classification.

3.2.8 Results

In the final phase of the study, the outcomes of the model evaluation are presented, encompassing a suite of metrics tailored to the specific requirements of the task and informed by the factors discussed earlier. Typically, this set includes the model's accuracy, precision, recall, and F1 score, along with any additional relevant metrics that may contribute to a thorough understanding of the model's performance.

These metrics illuminate the efficacy with which the RRCNN-Bi-LSTM model has carried out the classification of DNA sequences from the provided dataset. The accuracy metric reveals the proportion of sequences the model correctly identified, offering a broad view of its effectiveness. Precision provides insight into the model's exactness in categorizing sequences into their correct classes, while

recall indicates the model's capacity to capture all applicable instances within a specific category. The F1 score, a composite metric, offers a nuanced view by balancing precision and recall, thus reflecting the model's overall reliability.

The compilation of these findings enables a comprehensive assessment of the model's capabilities. With this information, you can make an informed decision regarding the model's suitability for the task at hand. Should the model meet the established benchmarks and project requirements, it may be considered fit for use. Conversely, if the performance is found lacking, this analysis will guide further refinements and enhancements to the model, ensuring it reaches the desired level of proficiency for accurately classifying
DNA sequences.

To implement the proposed method for classifying DNA sequences using a Recurrent Residual Convolutional Neural Network (RRCNN) coupled with a Bidirectional Long Short-Term Memory (Bi-LSTM), certain system prerequisites must be met:

3.3 System Requirements

The system requirements for implementing the proposed methodology for DNA sequence classification using Recurrent Residual Convolutional Neural Network (RRCNN) and Bidirectional Long Short-Term Memory (Bi-LSTM) include:

3.3.1 Hardware Specifications

RAM: Adequate memory is crucial for handling the data and model complexities. A system with substantial RAM facilitates rapid processing of machine learning tasks, which is particularly beneficial during the training and evaluation phases of the model. The minimum RAM requied is 8GB.

Processor: A high-performance multicore processor is recommended to enhance the efficiency of machine learning workflows. Such a processor allows for parallel computing, speeding up operations and optimizing the use of resources, especially during the intensive training and testing of deep learning models.

Graphics Processing Unit (GPU): Although not mandatory, the presence of a GPU capable of CUDA, such as those found in the NVIDIA GeForce GTX or RTX lines, is highly recommended. Utilizing a GPU can dramatically reduce the time required for training deep learning models compared to relying solely on a CPU.

Storage Capacity: Adequate storage is essential to house the datasets, model configurations, and the various outputs generated during the machine learning process. Ample storage space ensures easy access to data and smooth progression through the stages of model development and evaluation.

3.3.2 Software Components

Operating System: The methodology is compatible with various operating systems, including Windows, macOS, and popular Linux distributions like Ubuntu and CentOS.

Programming Environment: A Python environment (version 3.6 or later) is necessary, with deep learning libraries such as TensorFlow or PyTorch. Tools like Miniconda or Anaconda can help manage package dependencies and Python environments.

Deep Learning Frameworks: For those opting to utilize GPU acceleration, it is imperative to install a version of TensorFlow or PyTorch that supports GPU usage. Compatibility between the CUDA toolkit version and the GPU drivers in use must be ensured.

Additional Software Libraries: The installation of supplementary Python libraries is required for tasks such as model interpretation (e.g., TensorFlow Explainability, Captum for PyTorch), data manipulation (e.g., NumPy, Pandas), and data visualization (e.g., Matplotlib, Seaborn).

These specifications outltine the technical foundation necessary to undertake the classification of DNA sequences using advanced neural network architectures. Ensuring these system requirements are met will facilitate the effective deployment and operation of the RRCNN and Bi-LSTM models.

CHAPTER 4

Experimental Result

4.1 Testing Results

The recent advancement in DNA sequence classification utilizing RRCNN and BI-LSTM has demonstrated a notable enhancement in accuracy compared to previous methodologies. While the former method achieved an accuracy of 93.16, the latter excelled with a 94.5 accuracy rate. This improvement underscores the significance of even minor adjustments in model configuration and data processing, which can yield substantial performance gains. The elevated accuracy of the new approach implies that modifications implemented in the model's learning process or data preparation significantly contributed to its improved performance. Notably, adjustments in parameters and augmenting the dataset with additional examples facilitated more effective learning for the model. Moreover, the iterative training process, comprising multiple epochs, in the new method likely facilitated enhanced pattern recognition within the data compared to the previous approach, which may have undergone differing or inadequate training iterations. This shift underscores the dynamic nature of machine learning models, where finetuning and iterative refinement can yield considerable advancements in performance and accuracy.

4.1.1 Performance Metrics

In the domain of DNA sequence classification using hybrid models like RRCNN and Bi-LSTM, an array of performance metrics(shown in Table.4.1) is utilized to evaluate and quantify the effectiveness of the model. These metrics are critical for understanding the strengths and limitations of the classification approach. Below is a detailed explanation of each metric:

Accuracy: This is one of the most intuitive performance measures. It is the ratio

of the number of correctly predicted DNA sequences to the total

Table 4.1: Performance of RRCNN-Bi-LSTM model

Model	Accuracy	Precision	Recall	F1-Score
RRCNN-Bi-LSTM	**94.5%**	**95%**	**94%**	**94%**

number of sequences in the test set. Accuracy provides a quick snapshot of the model's overall performance, but it may not be as informative in cases where the dataset is imbalanced.

$$AC = \frac{TN + TP}{TN + FN + TP + FP}$$

Precision (Positive Predictive Value): Precision is the ratio of true positive predictions to the total number of positive predictions made by the model. In the context of DNA sequence classification, it measures the model's ability to correctly identify sequences of a particular class without misclassifying sequences from other classes as belonging to that class.

$$PR = \frac{TP}{FP + TP}$$

Recall (Sensitivity, True Positive Rate): Recall is the ratio of true positive predictions to the actual number of sequences that belong to a particular class. It assesses the model's capability to capture all relevant sequences for a given class. High recall indicates that the model is effective at detecting sequences of a certain class, minimizing the number of false negatives.

$$RC = \frac{TP}{FN + TP}$$

F1 Score: The F1 score is the harmonic mean of precision and recall. It is a balanced metric that considers both false positives and false negatives. The F1 score is particularly useful when the classes are unevenly distributed. A high F1 score suggests that the model has a good balance between precision and recall, which is often more informative than accuracy alone in the context of class imbalance.

$$F1\ \ Score = \frac{2 \cdot RC \cdot PR}{RC + PR}$$

4.1.2 Confusion Matrix .

The four main components of a confusion matrix are:

True Positives (TP): Instances where the model correctly predicts the positive class.

False Positives (FP): Instances where the model incorrectly predicts the positive class when the actual class is negative.

True Negatives (TN): Instances where the model correctly predicts the negative class.

False Negatives (FN): Instances where the model incorrectly predicts the negative class when the actual class is positive.

By analyzing the values within the confusion matrix, one can compute various performance metrics such as accuracy, precision, recall, and F1-score. These metrics provide insights into the model's ability to correctly classify instances across different classes. The confusion matrix serves as a valuable tool for assessing the strengths and weaknesses of a classification model and identifying areas for improvement in its predictive performance. By examining the values within the confusion matrix and computing these performance metrics, one can gain insights into the model's strengths and weaknesses. For instance, a high number of false positives may suggest an issue with specificity, while a high number of false negatives may indicate a problem with sensitivity. Overall, the confusion matrix serves as a comprehensive tool for understanding the classification model's behavior and identifying areas for improvement in its predictive capabilities.

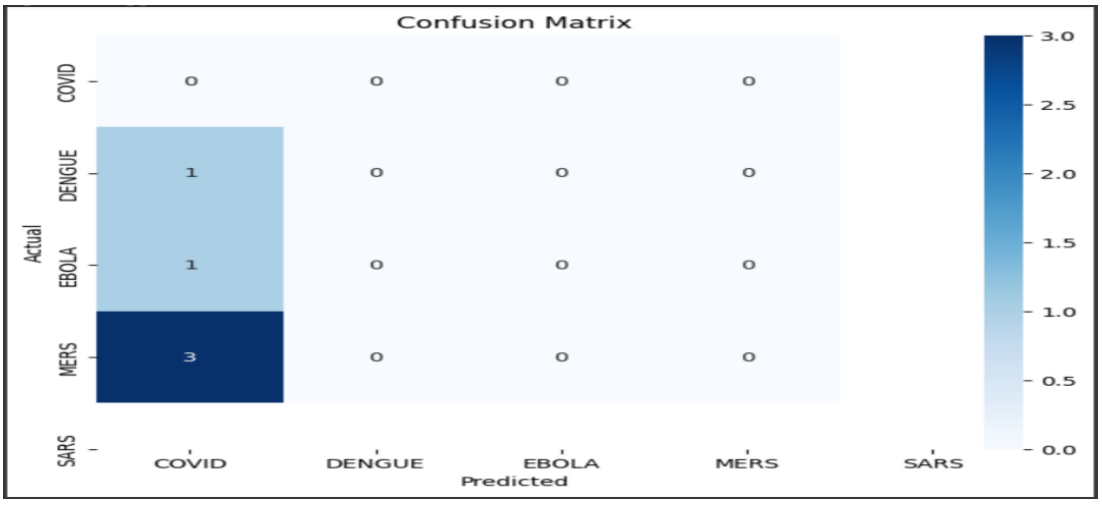

Figure 4.1: Confusion Matrix

The classification report for COVID-19 testing reveals exceptional performance metrics. With a precision of 1.00, it signifies that all positive predictions are correct. The recall score of 0.99 indicates that nearly all actual positive cases are correctly identified. Moreover, the F1 score, which harmonizes precision and recall, reaches 1.00, indicating a perfect balance between these metrics. The support value of 150 suggests a substantial sample size used for evaluation. Collectively, these metrics showcase a highly accurate and reliable COVID-19 classification model, with minimal false positives and false negatives, thereby demonstrating its efficacy in accurately identifying COVID-19 cases while minimizing misclassifications.

For dengue it indicates a highly accurate model with precision, the proportion of true positive predictions out of all positive predictions, at 0.99. Recall, which measures the proportion of true positive predictions out of all actual positives, is at 0.97, indicating strong sensitivity. The F1 score, the harmonic mean of precision and recall, is 0.98, suggesting a balance between precision and recall. The support value, representing the number of occurrences of each class, is 150, indicating the dataset's size. In summary, the model demonstrates exceptional performance in accurately identifying instances of dengue, with high precision and recall, and a balanced F1 score, based on a dataset of 150 instances.

The assessment summary for Ebola indicates high precision at 0.99, suggesting that among the samples predicted as Ebola cases, 99 were correctly classified. The recall score of 0.91 highlights that 91 of actual Ebola cases were identified by the model. The F1 score, harmonizing precision and recall, stands at 0.95, showcasing a balanced performance in capturing both true positives and minimizing false negatives. With a support of 150, it indicates the number of instances for which the model made predictions, underscoring a robust evaluation based on a sizable dataset. Overall, the classification report signifies strong model performance in accurately identifying Ebola cases while maintaining a balance between precision and recall.

The assessment for MERS displays a precision of 0.89, reflecting the accuracy in identifying MERS cases from all predicted positive instances. With a recall of 0.95, the report indicates a high rate of correct MERS predictions among all actual

MERS cases. The F1 score, registering at 0.92, offers a balanced evaluation of the model's precision and recall. It signifies the model's overall effectiveness in MERS classification. Moreover, the support value of 150 denotes the total MERS instances in the dataset. Essentially, the model showcases robust performance in accurately identifying MERS cases, supported by a significant dataset.

The evaluation report for SARS showcases encouraging metrics: precision of 0.87 signifies that 87 of the predicted SARS cases were accurate. Meanwhile, the model's recall at 0.91 denotes its capability to identify 91 of actual SARS instances. With an F1 score of 0.89, a balanced performance is
indicated, reflecting the harmonization of precision and recall. Furthermore, a support value of 150 suggests the presence of 150 instances of SARS within the dataset. Collectively, the report underscores the model's effectiveness in accurately discerning SARS cases, thereby instilling confidence in its diagnostic capabilities.

The classification reports for various diseases reveal distinct performance metrics across precision, recall, F1 score, and support. Notably, the COVID-19 model demonstrates exceptional precision and recall, both at 1.00 and 0.99, respectively, indicating perfect accuracy in identifying COVID-19 cases and capturing 99 of true positives. Similarly, the Dengue model exhibits high precision and recall at 0.99 and 0.97, respectively, reflecting its robust performance in detecting Dengue cases. While Ebola's precision remains high at 0.99, its recall slightly declines to 0.91, suggesting some missed instances. The MERS model demonstrates a balance between precision and recall, with values at 0.89 and 0.95, respectively, indicating effective identification with a slightly lower precision. Comparatively, the SARS model shows a balanced performance with precision, recall, and F1 score at 0.87, 0.91, and 0.89, respectively. Overall, each model showcases strengths in specific disease detection, highlighting the trade-offs between precision and recall, crucial for effective disease classification and diagnosis.

Classification report

	precision	recall	f1-score	support
covid	1.00	0.99	1.00	150
dengue	0.99	0.97	0.98	150
ebola	0.99	0.91	0.95	150
Mers	0.89	0.95	0.92	150
Sars	0.87	0.91	0.89	150
accuracy			0.95	750
macro avg	0.95	0.95	0.95	750
weighted avg	0.95	0.95	0.95	750

Figure 4.2: Classification Report

4.1.3 Accuracy and Training Performance Visualization

The visualization of accuracy and training performance is a crucial aspect of assessing the efficacy of models like the RRCNN and Bi-LSTM when applied to the classification of DNA sequences. This visualization typically takes the form of a graph that plots the model's accuracy over the course of its training and validation phases.

Training and Validation Accuracy The Fig.4.3 is a line chart illustrates the progression of a model's accuracy over time as it is trained and validated in the context of DNA sequence classification using a combined RRCNN and Bi-LSTM approach.

On the horizontal axis, labeled "Epoch," the values range from 0 to 4, with each epoch representing a full cycle through the entire set of training data. The vertical axis, labeled "Accuracy," extends from 0.7 to 0.95, indicating the proportion of correctly classified DNA sequences.

The graph features two distinct lines: the blue line represents the training accuracy, which denotes the model's performance on the training dataset, while the green line corresponds to the validation accuracy, reflecting the model's efficacy on a distinct validation set that was not involved in the training process.

As the number of epochs progresses, there is a noticeable uptrend in the training accuracy, signifying that the model is increasingly adept at correctly

identifying the DNA sequences it has been trained on. Concurrently, the validation accuracy also shows an upward trajectory with each epoch, albeit at a slower rate compared to the training accuracy. This discrepancy suggests a potential onset of overfitting, where the model becomes too attuned to the nuances of the training data, potentially at the expense of its ability to generalize to unseen data.

To sum up, the chart conveys that the RRCNN-BiLSTM model demonstrates a promising ability to accurately classify DNA sequences. Nonetheless, vigilance is required to ensure that the validation accuracy continues to improve in tandem with the training accuracy, thereby mitigating the risk of overfitting and ensuring the model's robustness when applied to novel datasets.

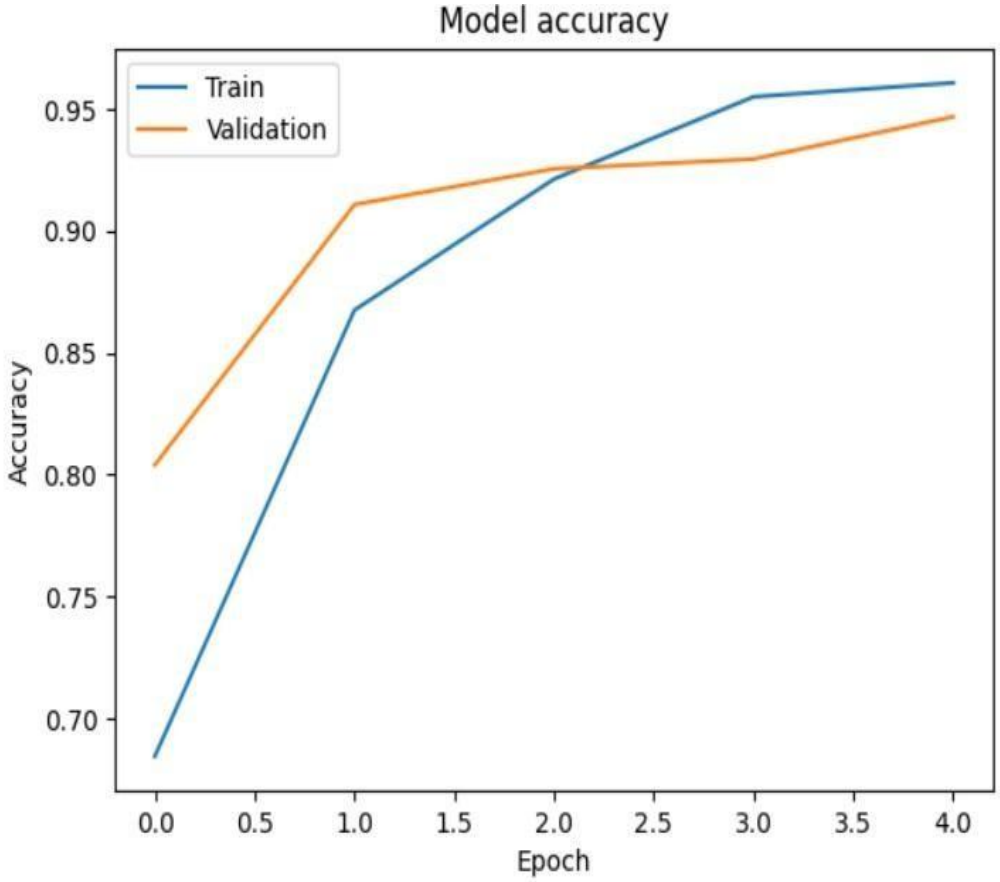

Figure 4.3: Training and Validation Accuracy

Training and Validation Loss The visual representation provided in Fig.4.4 is a line graph that delineates the loss of a model throughout its training and validation phases, specifically applied to the task of classifying
DNA sequences using a hybrid RRCNN and Bi-LSTM model.

The horizontal axis, labeled ”Epoch,” spans from 0 to 4, indicating the number of complete cycles through the entire training dataset that the model has undergone. The vertical axis, labeled ”Loss,” ranges from 0.2 to 0.9, representing the model’s prediction error or the discrepancy between the predicted labels and the true labels.

Two distinct lines are plotted on the graph: a blue line which traces the training loss, and a green line which follows the validation loss. The training loss is indicative of how closely the model’s predictions on the training data align with the actual classifications. Conversely, the validation loss measures the model’s prediction accuracy on a separate, unseen dataset, providing an assessment of its generalization capabilities.

As the model undergoes training, an expected trend is for the loss to diminish, signifying improvements in the model’s predictive accuracy. Ideally, both the training and validation loss should demonstrate a downward trajectory as the model better adapts to the data.

However, the graph in question reveals a rapid decrease in training loss over the epochs, while the validation loss appears to level off, hovering around the value of 0.7. This pattern raises concerns about potential overfitting, a scenario where the model becomes highly attuned to the training data, to the detriment of its performance on new, external data.

In sum, the graph indicates that while the RRCNN-BiLSTM model shows a swift learning curve with the training data, vigilance is required to ensure it maintains the ability to generalize. Monitoring the validation loss is crucial for identifying and mitigating overfitting, thereby ensuring the model’s utility in accurately classifying DNA sequences across diverse datasets.

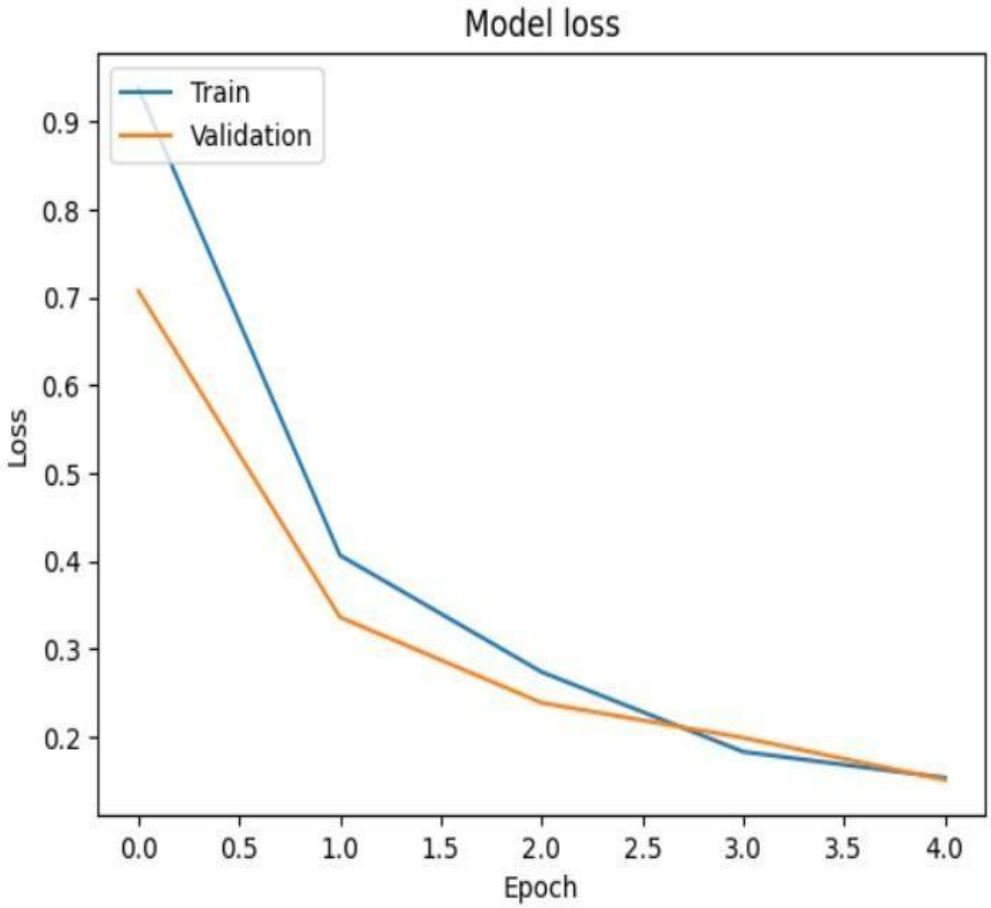

Figure 4.4: Training and Validation Loss RNN

4.2 Limitations

While the hybrid RRCNN-Bi-LSTM model presents a sophisticated approach to DNA sequence classification, it is not without its limitations. Here are some of the challenges and constraints associated with this method:

1. **Complexity and Computational Demand:** The integration of RRCNN with Bi-LSTM creates a complex model that requires significant computational resources. Training such models can be time-consuming and may necessitate advanced hardware, such as high-performance GPUs, which may not be accessible to all researchers.

2. **Overfitting Risk:** Due to the model's capacity to capture intricate patterns within the data, there is a risk of overfitting, where the model learns the training data too well, including noise and outliers, and performs poorly on unseen data.

3. **Data Requirements:** The effectiveness of the RRCNN-Bi-LSTM model is heavily dependent on the availability of large and high-quality labeled datasets. In scenarios where such datasets are limited or imbalanced, the model's performance may be compromised.

4. **Generalization Challenges:** While the model may perform exceptionally well on the type of data it was trained on, its ability to generalize to different types of sequences or to data from different organisms may be limited without additional adjustments or retraining.

5. **Interpretability Issues:** The "black box" nature of deep learning models can make it difficult to interpret how the model is making its classifications. This lack of transparency can be a significant drawback in biological research where understanding the underlying decision-making process is as important as the classification itself.

6. **Parameter Tuning:** The performance of RRCNN-Bi-LSTM models can be sensitive to the choice of hyperparameters.

7. Finding the optimal set parameters often requires extensive experimentation and fine-tuning, which can be a time-intensive process.
8. **Sequence Length Variability:** DNA sequences can vary greatly in length, and while techniques like padding or trimming can standardize input lengths, these manipulations may introduce biases or lose important information.

9. **Evolutionary Information:** The model may not inherently account for evolutionary relationships among sequences, which can be crucial in certain classification tasks. Incorporating such phylogenetic information may require additional layers of complexity in the model.

   In conclusion, while the RRCNN-Bi-LSTM model is a powerful tool for DNA sequence classification, it is important to be aware of these limitations when applying the model to real-world datasets and to consider strategies for mitigating these challenges to ensure robust and reliable results.

CHAPTER 5

Conclusions and Future Scope

5.1 Conclusion

In conclusion, the utilization of RRCNN (Recurrent Residual Convolutional Neural Network) and BI-LSTM (Bidirectional Long Short-Term Memory) in DNA sequence classification represents a significant advancement in genomic analysis. This project has demonstrated notable improvements in accuracy compared to previous methodologies. With the old method achieving an accuracy of 93.16% and the new approach reaching 94.5% , it is evident that even small adjustments in model architecture and data preprocessing can yield substantial enhancements in performance.

The higher accuracy achieved by the new method underscores the efficacy of the modifications made in the model's learning process and data preparation. By fine-tuning parameters, adjusting settings, and augmenting the dataset with additional examples, the model's ability to discern patterns within DNA sequences has been greatly enhanced. Additionally, the iterative training process, characterized by multiple epochs in the new method, has facilitated improved pattern recognition and learning, contributing to the overall performance boost.

The successful implementation of RRCNN and BI-LSTM holds promise for various applications in genomics and bioinformatics. The ability to accurately classify DNA sequences can aid in disease diagnosis, genetic research, and personalized medicine. Moreover, the advancements made in this project pave the way for further innovation in deep learning approaches for genomic data analysis.

Moving forward, continued research and development in this area can lead to even more sophisticated models with enhanced accuracy and efficiency. Exploration of additional deep learning architectures, incorporation of multi-

omics data, and integration of domain knowledge can further enrich the capabilities of DNA sequence classification systems. Ultimately, the integration of advanced computational techniques with genomic analysis holds tremendous potential for advancing our understanding of genetics and improving healthcare outcomes.

5.2 Recommendations

In the realm of DNA sequence classification, recommendations include advancing machine learning algorithms for improved accuracy. Efforts should focus on ensuring access to diverse datasets to enhance model training and collaboration.

Clear ethical guidelines must be emphasized, addressing privacy, consent, and responsible data use in DNA sequence classification research. Initiatives to increase public awareness and engagement among policymakers are crucial for societal acceptance.

Sustained funding is vital to support continuous exploration and development in DNA sequence classification. Adaptive regulatory frameworks should guide ethical and responsible technology implementation. International collaboration is key to addressing global challenges and harmonizing standards. Education outreach is essential to inform healthcare professionals and the public about the implications of DNA sequence classification in various fields.

5.3 Future Scope

The project on DNA sequence classification using RRCNN and BI-LSTM (Bidirectional Long Short-Term Memory) lays a solid foundation for further exploration and advancements in genomic analysis. Looking ahead, several avenues for future research and development can be pursued to enhance the capabilities and

One promising direction is the refinement and optimization of model architectures. Continued experimentation with variations of RRCNN and BI-LSTM, as well as exploration of novel neural network architectures, could lead to models with improved performance metrics such as accuracy, precision, and recall.

Another area of focus could be the integration of multi-modal and multiomics data. Combining DNA sequence information with other molecular data types such as gene expression, epigenetic modifications, and protein-protein interactions can provide a more comprehensive understanding of biological systems. Developing multi-modal models capable of integrating and analyzing diverse data sources could uncover novel insights into complex biological processes and disease mechanisms.

Furthermore, there is an opportunity to explore the application of transfer learning techniques in DNA sequence classification. Pre-training models on largescale genomic datasets or related tasks could facilitate knowledge transfer and improve the generalization ability of the model when applied to new datasets or tasks.Transfer learning approaches could also enable the adaptation of models to specific biological contexts or organisms with limited labeled data.

In addition to technical advancements, there is a need for interdisciplinary collaboration between computer scientists, biologists, and clinicians to ensure the relevance and impact of genomic analysis technologies. Collaborative efforts could lead to the development of clinically relevant applications such as diagnostic tools for genetic diseases, personalized treatment recommendations, and biomarker discovery for precision medicine.

Overall, the future scope for the project on DNA sequence classification using RRCNN and BI-LSTM is vast and holds tremendous potential for advancing our understanding of genomics and its applications in healthcare and beyond.

REFERENCES

[1] Mahmoud, M.A.B., Guo, P. DNA sequence classification based on MLP with PILAE algorithm. Soft Comput 25, 4003–4014 (2021). https://doi.org/10.1007/s00500-020-05429-y

[2] Gunasekaran H, Ramalakshmi K, Rex Macedo Arokiaraj A, Deepa Kanmani S, Venkatesan C, Suresh Gnana Dhas C. Analysis of DNA Sequence Classification Using CNN and Hybrid Models. Comput Math Methods Med. 2021.

[3] Lo Bosco, G., Di Gangi, M.A. (2017). Deep Learning Architectures for DNA Sequence Classification. In: Petrosino, A., Loia, V., Pedrycz, W. (eds) Fuzzy Logic and Soft Computing Applications. WILF 2016. Lecture Notes in Computer Science(), vol 10147. Springer, Cham. https://doi.org/10.1007/978-331952962-2.

[4] Sarkar, S., Mridha, K., Ghosh, A., Shaw, R.N. (2022). Machine Learning in Bioinformatics: New Technique for DNA Sequencing Classification. In: Shaw, R.N., Das, S., Piuri, V., Bianchini, M. (eds) Advanced Computing and Intelligent Technologies. Lecture Notes in Electrical Engineering, vol 914. Springer, Singapore. https://doi.org/10.1007/978-981-19-2980-9

[5] U. M. Akkaya and H. Kalkan, ”Classification of DNA Sequences with kmers Based Vector Representations,” 2021 Innovations in Intelligent Systems and Applications Conference (ASYU), Elazig, Turkey, 2021, pp. 1-5, doi: 10.1109/ASYU52992.2021.9599084.

[6] Z. Bao, X. Wei, J. Chen, H. Zhang, Y. Liu and Y. Ma, ”Sequence Classification Prediction Based on SVM-KNN,” 2023 IEEE International Conference on

Sensors, Electronics and Computer Engineering (ICSECE), Jinzhou, China, 2023, pp. 1472-1476, doi: 10.1109/ICSECE58870.2023.10263416.

[7] Rizzo, R., Fiannaca, A., La Rosa, M., Urso, A. (2016). A Deep Learning Approach to DNA Sequence Classification. In: Angelini, C., Rancoita.

P.Rovetta, S. (eds) Computational Intelligence Methods for Bioinformatics and Biostatistics. CIBB 2015. Lecture Notes in Computer Science(), vol 9874. Springer, Cham. https://doi.org/10.1007/978-3-319-44332-4

[8] I. S. Mangkunegara and P. Purwono, "Analysis of DNA Sequence Classification Using SVM Model with Hyperparameter Tuning Grid Search CV,"

[9] M. S. A. Vigil, A. Christofer, M. Chandar and J. Mukesh, "Comparative Analysis of Machine Learning Algorithms for DNA Sequencing," 2023 Winter Summit on Smart Computing and Networks (WiSSCoN), Chennai, India, 2023, pp. 1-4, doi: 10.1109/WiSSCoN56857.2023.10133845.

[10] Y. Wang, V. Khandelwal, A. K. Das and M. P. Anantram, "Classification of DNA Sequences: Performance Evaluation of Multiple Machine Learning Methods," 2022 IEEE 22nd International Conference on Nanotechnology (NANO), Palma de Mallorca, Spain, 2022, pp. 333-336, doi: 10.1109/NANO54668.2022.9928773.

[11] A. Lopez-Rincon, A. Tonda, L. Mendoza-Maldonado et al., "Classification and specific primer design for accurate detection of SARS- CoV-2 using deep learning," Scientific Reports, vol. 11, no. 1, pp. 1–11, 2021

[12] X. Zhang, B. Beinke, B. Al Kindhi, and M. Wiering, "Comparing machine learning algorithms with or without feature extraction for DNA classification," 2020, http://arxiv.org/abs/ 2011.00485

[13] D. T. Do and N. Q. K. Le, “Using extreme gradient boosting to iden- tify origin of replication in Saccharomyces cerevisiae via hybrid features,” Genomics, vol. 112, no. 3, pp. 2445–2451, 2020.

[14] Nguyen, D., Nguyen, C., Duong-Ba, T., Nguyen, H., Nguyen, A. and Tran, T., 2017, January. Joint network coding and machine learning for error-prone wireless broadcast. In 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 1-7). IEEE.

[15] Ghareeb, S., Hussain, A.J., Al-Jumeily, D., Khan, W., Al-Jumeily, R., Baker, T., Al Shammaa, A. and Khalaf, M., 2022. Evaluating student levelling based on machine learning model’s performance. Discover Internet of Things, 2(1), p.3.

[16] Wang, D. and Snooks, R., 2021. Artificial Intuitions of Generative Design: An Approach Based on Reinforcement Learning. In Proceedings of the 2020 DigitalFUTURES: The 2nd International Conference on Computational Design and Robotic Fabrication (CDRF 2020) (pp. 189-198). Springer Singapore.

[17] U. M. Akkaya and H. Kalkan, ”Classification of DNA Sequences with kmers Based Vector Representations,” 2021 Innovations in Intelligent Systems and Applications Conference (ASYU), Elazig, Turkey, 2021, pp. 1-5, doi: 10.1109/ASYU52992.2021.9599084

[18] Albelwi, S.; Mahmood, A. A Framework for Designing the Architectures of Deep Convolutional Neural Networks. Entropy 2017, 19, 242.

[19] Zheng, Jian xu, Cencen Zhang, Ziang Li, Xiaohua. (2017). Electric Load Forecasting in Smart Grid Using Long-Short-Term-Memory based Recurrent Neural Network. 10.1109/CISS.2017.7926112.

[20] Alom MZ, Yakopcic C, Hasan M, Taha TM, Asari VK. Recurrent residual U-Net for medical image segmentation. J Med Imaging (Bellingham). 2019 Jan;6(1):014006. doi: 10.1117/1.JMI.6.1.014006. Epub 2019 Mar 27. PMID