REVIEW ARTICLE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
Vol. 05, Issue 11 November
www.rspsciencehub.com

Check for updates

http://dx.doi.org/10.47392/IRJASH.2023.072

A Survey on Deep Learning Approaches Used in Genomics

Rohit Kumar Gupta[1], Dr. Sweeti Sah[1], Dr B. Surendiran [1], Dr. Shankar Narayan[2], Dr Arunkumar P [1]

[1]Department of Computer Science and Engineering, National Institute of Technology Puducherry, Karaikal India.

[2]Sr. Analyst/Sr.Scientist, Forensic Science Laboratory, Govt of Puducherry, Puducherry- 607403.

Emails: rohitguptababai@gmail.com, sweetisah3@gmail.com, surendiran@nitpy.ac.in, shankarnarayan.v@gmail.com, arunkumar.pselvam@gmail.com

Article History

Received: 16 August 2023
Accepted: 12 November 2023
Published: 16 November 2023

Keywords:

Common stage in DNA sequencing;
DNA sequencing method;
Deep learning;
Genomics

Abstract

Deep learning (DL) methods have shown remarkable success in addressing various problems across different domains. Classifying DNA sequences presents a formidable challenge in the field of bioinformatics. This review delves into various technologies centered around Alignment methods and Deep Learning for the purpose of classification. The aim is to achieve accurate and scalable predictions for DNA sequence classification. DL methods have proven effective in overcoming the primary challenges faced during the training process. The paper delves into previous classification methods like alignment methods and highlights their limitations. Subsequently, we delve into the application of deep learning, specifically using CNN and RNN models, for DNA sequence classification. We evaluate their respective accuracies and discuss the differences and drawbacks associated with these methods.

1. Introduction

Genomic information can be characterized as the digital repository containing an organism's genetic data, primarily composed of nucleotides. These nucleotides are portrayed through finite sequences of letters derived from a genetic alphabet, such as those found in DNA and proteins. Mathematically, DNA has been represented using character strings, where each character corresponds to a letter in the genetic alphabet (Anastassiou). Nucleotides play a fundamental role as the basic building blocks of nucleic acids, which consist of a sugar molecule (such as deoxyribose) and a nitrogen-containing base (A, C, T, G) (Henderson, Frank, and Paterson). Both DNA and RNA are comprised of polymer chains featuring sequences of nucleotides. DNA is distinguished by its deoxyribose sugar, while RNA contains ribose sugar. The DNA structure is characterized by a lengthy sequence of bases, depicted as a finite string composed of four letters (A, C, T, G) (Hartwell et al.).

An illustrative example is ATCGATCG....ATCGCTGAAGT.

DNA is comprised of two strands that coil into a double-helix structure, resembling a spiral ladder. Consisting of four nucleotide types—adenine (A), cytosine (C), guanine (G), and thymine (T), DNA establishes chemical bonds that link the two DNA strands. Adenine (A) invariably pairs with thymine (T), while guanine (G) consistently pairs with cytosine (C), as depicted in Fig. 1. It is noteworthy that each cell in the body houses a complete copy of about 3 billion DNA base pairs (Travers and Muskhelishvili Yang et al.). Genes, which are small DNA segments, are responsible for storing genetic information.

Bioinformatics involves employing computational techniques to analyze extensive biological

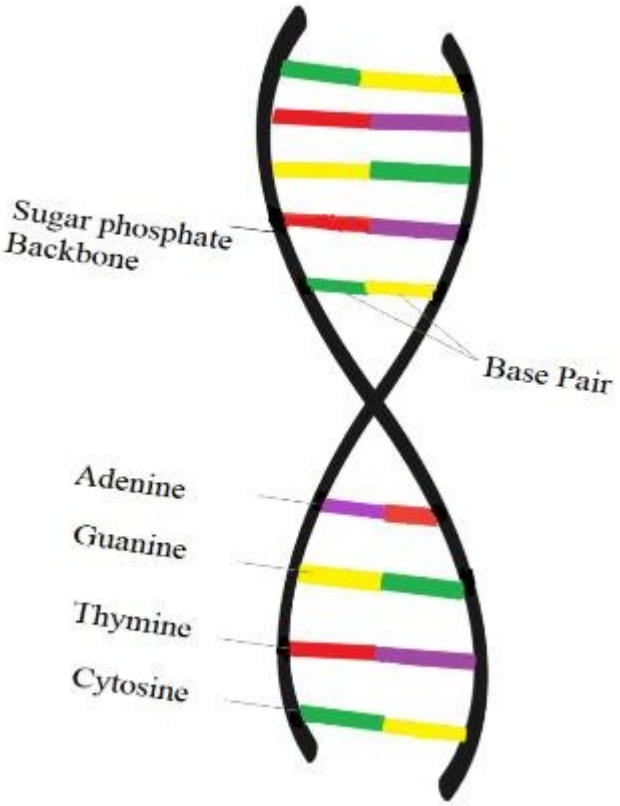

FIGURE 1. Structure of DNA (Hartwell et al.)

datasets, encompassing genetic information. Within this field, biological data is derived through the amalgamation of biological data, informatics data, and mathematical data. This combined dataset is then referred to as biological data, and it is analyzed within the framework of bioinformatics (Luscombe, Greenbaum, and Gerstein). This approach involves the summarization and comprehension of the characteristics and hierarchy of biological information. Bioinformatics utilizes computational analysis of biological data, specifically DNA sequences, to achieve various goals. A primary task involves comparing new DNA sequences with a well-known database using a similarity function to predict their respective groups. (Xiong) EBWS offers diverse tools for DNA sequence operations on a new PHP-based server (Kaloudas, Pavlova, and Penchovsky). Taxonomy (Dunn and Everitt Godfray and Charles) involves classifying Biological Organisms based on comprehend them.

Short tandem repeats, often abbreviated as STRs, are brief DNA sequences with fewer than 400 base pairs, which makes them particularly advantageous (see Figure 2). Classifying DNA sequences plays a crucial role in genomics, as it enables the prediction of the DNA's classification, which is invaluable in the medical field for identifying viral infections. Given the vast number of viruses, such as HIV and COVID-19, which exceed 1.6 million, it becomes essential to label and determine the virus's name based on the DNA sequence. To accomplish this, the DNA sequence is compared against the Gen Bank database provided by the NCBI. The Gen Bank database is a repository of DNA sequences, encompassing more than 106 billion nucleotide bases (Reid et al.). Short Tandem Repeats, or STRs, serve as repositories for repeat units with lengths ranging from 2 to 6 base pairs and are frequently amplified using the polymerase chain reaction. Owing to their minimal DNA quantity requirements, STRs have gained popularity in forensic laboratories and are widely utilized in biological research due to their polymorphic nature and high mutation rates. These sequences are alternatively known as microsatellites or simple sequence repeats.

Short tandem repeats (STRs) can be categorized into various types according to the repeat unit's characteristics. For instance, they can be categorized according to the major repeat unit length, leading to classifications such as mono, di, tri, tetra, penta and hexa-nucleotide repeats (Reid et al.). Currently, these STR markers are utilized in the development of the FBI Combined DNA Index System, which constitutes an extensive DNA database covering a jurisdiction (Ruitberg et al.).

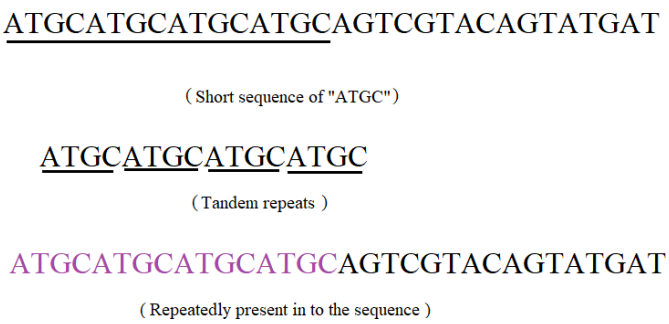

FIGURE 2. Short Tandem Repeats (STR) (Reid et al.)

CODIS, the Combined DNA Index System, stands as one of the paramount tools for locating, identifying, and classifying DNA sequences (Penacino). This system is a versatile software that serves the purpose of constructing a DNA database, consisting of three specific index

types: LDIS, SDIS, and NDIS (Jovanović). LDIS is administered at the local city level, SDIS at the state agency level, and NDIS by the Federal Bureau of Investigation (see Figure 3). Within the CODIS software, numerous databases are integrated for information retrieval concerning missing individuals and convicted offenders. It's crucial to emphasize that CODIS does not store any personally identifiable information associated with the DNA profiles.

CODIS functions primarily as a valuable tool for generating investigative leads in cases where biological evidence is collected from crime scenes. It operates through two distinct indexes: the forensic index, which contains DNA profiles derived from evidence collected at crime scenes, and the offender index, which houses DNA profiles of individuals convicted of sex offenses (and in some states, extended to include other felonies) (Penacino). Within the CODIS software, multiple databases are designed for various information searches, encompassing data on missing persons and convicted offenders.

DNA sequences are archived with a '.DNA' extension (Jovanović). Various file formats have been developed to facilitate the storage, manipulation, analysis, and comparison of nucleotide and protein sequences. Notable formats include Plain sequence format, FASTA format, FASTQ format, GCG format, GCG – RSF (Rich sequence format), and EMBL format (Graham and Faulds). The FASTA format and FASTQ format serve as standardized means for sequence data storage. The FASTA format is used when raw sequence data is required, while the FASTQ format is employed when there is a need for quality-related information about the DNA sequence (Shen et al.).

The Plain sequence format comprises a sequence containing solely nucleotide bases and spaces. This format does not permit the inclusion of numbers or additional information regarding the DNA sequence. One limitation of the plain sequence format is that it can accommodate only a single DNA sequence within a given DNA sequence file.

Example:

ATCTAGTCCTGATAGAT-
GACGATGACGATCAGTT-
AGGTCTAGTCCTGAATCTA-
GTCCTGGATCTAGTCCT-
GACGATACGATCAAGTCCT-
GGATCTAGTCTACGATCAT

FASTA Format: The FASTA Format utilizes a distinct sequence ID known as the FASTA definition line, marked by the greater-than (">") symbol for each sequence. This text-based format is commonly utilized to archive DNA sequences, and a single sequence file may encompass multiple DNA sequences, often using file extensions such as '.FASTA' or 'fna'. Below is an example of a sequence presented in the FASTA format (Pearson).

Example:

>Sequence1-id description ATC-
GATAACGTGCTGAGTAGTGTGACTGTATC-
GATAACGTGCTGAGTAGTGTGAATAAGT-
GCATAACGTGCTTC

>Sequence2-id description GCTGAG-
TAGTGTGAATCGATAACGTCGATAAC-
CTGTATGTGCTGAGTAGTAACGTG-
CATAACGTGCTTCTGAATA

FASTQ format: FASTQ stands as a text-based format employed for the storage of biological sequences, especially those generated by high-throughput sequencing instruments (Wski). Sequences in the FASTQ format are structured across four lines. The initial line commences with the '@' character and incorporates a sequence identifier, optionally accompanied by a description. The second line contains the unprocessed sequence letters. The third line commences with a '+' symbol, accompanied by a description. Lastly, the fourth line contains the quality values corresponding to the sequence presented in second line (Shen et al. Wski).

Example: @SEQ_ID

AGTAGTGTG...TGTGCTGA+(The quality values for the sequence are encoded as in Line2)

GCG Format: The term GCG refers to Genetics Computer Group format, and it is designed to hold precisely one sequence within a single file (Womble). The sequence initiates with annotation lines encompassing information such as the sequence identifier, sequence length, and a checksum. Following the annotation lines, there are two dots ("..") indicating the sequence's commencement. It is important to note that GCG-format sequence files are exclusively generated using the GCG package (Shen et al. Dölz).

Example: ID (Description) AGTAGT-
GTGAATCGATAACGTCGTGTGT-

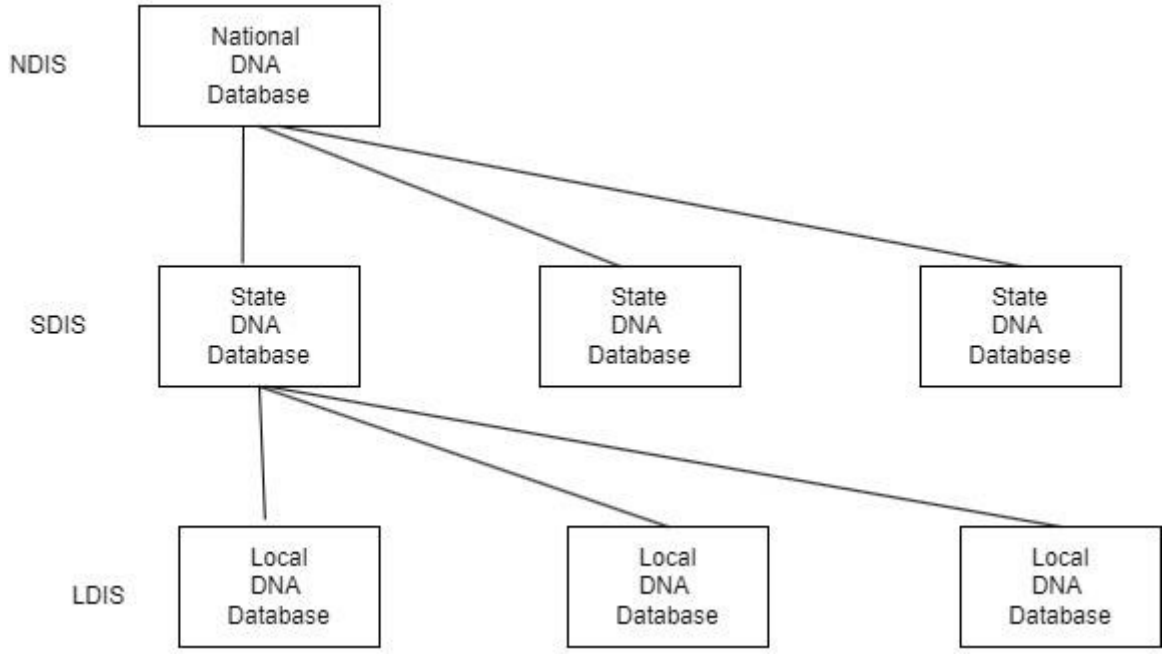

FIGURE 3. CODIS (Jovanović)

GAATCGATAACGTCGAGTTGTGAATC-GATAACGTCGCGTCG

The Nucleotide Archival Format (NAF) is a unique file format designed specifically for the lossless compression of nucleotide sequences presented in FASTA and FASTQ, without the need for any references (Kryukov et al.).

1.1. Motivation:

Classification methods for DNA using Machine Learning and Deep Learning techniques exhibit distinct characteristics. Machine Learning may encounter challenges in handling intricate patterns, while Deep Learning, employing Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN), excels in managing extensive DNA datasets and capturing intricate features. This efficiency disparity primarily arises from architectural differences, with CNN being suitable for image-driven tasks and RNN for handling sequential data like DNA sequences.

The important objective of this study is to unfold the application of Deep Learning-based methods for sequence classification and comprehend their functioning. The research aims to assess existing DNA classification techniques, which include alignment based approaches, alignment free methods, and combinations involving DSP and machine learning, with a focus on identifying critical limitations. The introduction of Deep Learning methods for DNA classification is expected to enhance accuracy compared to other methodologies. Additionally, the study delves into DNA sequence preprocessing for classification models.

The research investigates two Deep Learning architectures, namely CNN (Convolutional Neural Networks) and RNN (Recurrent Neural Networks), each based on distinct computational models designed for DNA sequence classification. The 16S rRNA dataset from RDP 11 is employed for classification purposes. DNA sequences are stored with the '.DNA' extension and can be selected from a range of formats, including Plain sequence, FASTA, GCG, GCG - RSF, and EMBL, for DNA database storage.

Lastly, the study examines the efficiency and challenges associated with deep neural architectures in the classification of DNA sequences in the field of bio informatics.

2. Literature Review:

[Anastassiou et al., 2001] Genomic information is inherently digital, typically manifested as finite sequences containing the bases A, T, G, and C, as observed in DNA. Analogous to DNA and proteins, these sequences are often represented mathematically as character strings, where each character corresponds to a letter in an alphabet.

[Henderson et al., 2014] Nucleotides play a foundational role as the building blocks of nucleic

acids, incorporating sugar molecules like ribose and deoxyribose.

[Hartwell et al., 2018] Polymers take shape through extended chains of nucleotides, and both DNA and RNA, as such polymers, are constructed from these prolonged sequences of nucleotides. These nucleotides are composed of letters within an alphabet.

[Travers et al., 2015] DNA exhibits a unique structure, characterized by two strands that wind together into a spiral ladder known as a helix. Nucleotides connect to each other in what is referred to as a base pair, forming a connection between double DNA strands.

[Yang et al., 2020] In this work the DNA sequence is constructed using 4 types of deoxyribose nucleotides, which are commonly denoted as bases. These nucleotides bond together through chemical interactions, with adenine pairing with thymine and guanine with cytosine.

[Luscombe et al., 2001] Bioinformatics is the application of computational techniques for analyzing vast datasets pertaining to biomolecular information. It proves to be a valuable tool for unravelling intricate biological data, encompassing genetic codes.

[Xiong et al., 2006] Biological data is meticulously generated for molecular analysis by leveraging the combined forces of Mathematics, informatics, and biology.

[Kaloudas et al., 2018] The Essential Bioinformatics Web Services (EBWS) implementation caters to the analysis of polymers such as RNA and DNA sequences. The server offers a range of web-based applets for conducting various types of analysis.

[Dunn et al., 2004] Taxonomy is the systematic process of organizing and classifying groups of organisms based on their distinctive characteristics.

[Godfray et al., 2002] The utility of taxonomy extends to organizing and indexing knowledge stored in various forms, such as documents and photos.

[Reid et al., 2013] Short Tandem Repeat (STR) is a straightforward molecular technique utilized for DNA profiling in human cells. Its abbreviated sequence length makes it an ideal candidate for Polymerase Chain Reaction (PCR). The incorporation of STR in cell culture management can significantly improve the detection of cellular cross-contamination, leading to more precise assays.

[Ruitberg et al., 2001] Short Tandem Repeats consist of repeat units ranging from 2 to 6 base pairs in length. Due to their low DNA quantity requirement, STRs can be easily amplified using the polymerase chain reaction, making them a preferred choice in forensic laboratories.

[Fan et al., 2007] Short Tandem Repeats are concise, tandemly repeated DNA sequences with a repetitive unit typically spanning 1–6 base pairs. Also known as microsatellites or simple sequence repeats, STRs play versatile roles in molecular biology.

[Penacino et al., 2006] CODIS is employed in criminal investigations involving biological evidence retrieved from crime scenes, utilizing two indexes: the forensic and offender indexes.

[Jovanović, 2021] CODIS, functioning as a DNA database system, emerges as a pivotal tool for identifying perpetrators through comprehensive DNA analysis.

[Jeffery et al., 2012] DNA extension is employed for the efficient storage of DNA sequence files.

[Shen et al., 2016] The recognized standards for storing sequence data are FASTA and FASTQ.

[Graham et al., 2008] This study outlines various sequence storage formats of DNA, including FASTA format, Plain sequence format, EMBL format, GCG format, and GCG-RSF (Rich sequence format).

[Pearson et al., 1994] A sequence file formatted in FASTA can accommodate multiple sequences, with each sequence initiating with a single-line description marked by the symbol '>'.

[Deorowicz et al., 2011] The FASTQ format, a text-based structure for nucleotide sequences like DNA and their corresponding quality scores, supports multiple files and sequences. Each entry begins with a single-line description starting with the symbol '@'.

[Womble et al., 2000] The GCG format functions as a tool for manipulating, analyzing, and comparing nucleotide and protein sequences, with GCG representing Genetics Computer Group.

[Dölz et al., 1994] and [Kryukov et al., 2019] These papers elaborate on the Nucleotide Archival Format (NAF), a novel file format designed for the lossless, reference-free compression of nucleotide sequences formatted in FASTA and FASTQ.

[Anderson et al., 1998] This paper introduces a tool for searching DNA sequences.

[Schmieder et al., 2011] Describing an application that provides graphical guidance and performs filtering, reformatting, and preprocessing of DNA sequences, this paper explores practical aspects of DNA sequence manipulation.

[Blankenberg et al., 2010] This paper outlines a pipeline for manipulating next-generation sequencing data extracted from a sequencing machine, encompassing all steps through quality filtering.

[Saima et al., 2021] This work compares alignment based methods, alignment free methods and Deep Learning (DL) methods (such as RNN and CNN) for DNA classification. DL proves to be effective but demands a substantial amount of data.

[Armstrong et al., 2019] This paper provides a succinct survey of genome alignment and multiple alignment methods for classification, offering insights into the current state of the genome alignment and comparative annotation fields.

[Eisenhofer et al., 2019] The study assesses four distinct reference databases, demonstrating that nucleotide-to-nucleotide alignments using MAL Tn can faithfully replicate simulated metagenomes, even when dealing with short reads and elevated deamination levels. The research emphasizes the crucial role of database selection, underlining its significance for emerging researchers in the field of paleo microbiology.

[Altschul et al., 1990] This research introduces the fundamental alignment search tool, highlighting BLAST's significant speed advantage over existing sequence comparison tools while maintaining comparable sensitivity.

[Lipman et al., 1985] The study outlines an algorithm for identifying similarities between newly determined amino acid sequences and those already available in databases.

[Thompson et al., 1995] This paper delves into the improvement of the multiple sequence alignment method for aligning divergent nucleotide sequences. Individual weights are assigned to each sequence, and amino acid substitution matrices are adjusted at different alignment stages based on the sequences' divergence.

[Nagla et al., 2020] This research explores the advantages and limitations of the alignment-based method for DNA sequencing.

[Domazet-Lošo et al., 2011] The study presents a method based on matches between two DNA sequences, where the number of matches indicates close homology. The authors implemented a program called alfy (Alignment-Free local homology), demonstrating its efficiency in accurately detecting recombination breakpoints in simulated DNA sequences. The application of alfy to Escherichia coli genomes reveals new evidence supporting a hypothesis.

[Remita et al., 2017] This Paper introduces CASTOR, a classification platform based on machine learning. The research assesses CASTOR's performance in classifying diverse datasets, and the CASTOR web platform provides an open-access, collaborative, and reproducible environment for machine learning classifiers.

[G. E. Sims et al., 2011] This study presents the Feature Frequency Profiles (FFPs) alignment-free method, utilizing the frequencies of I-mer features in entire genomes to infer phylogenetic distances. It identifies distinctive features for clade classification, offering valuable insights into group evolution.

[Fan et al., 2021] This study outlines the fundamentals of deep learning, a subset of machine learning that focuses on autonomous learning and improvement. Deep learning employs artificial neural networks designed to simulate human thinking and learning, in contrast to the simpler concepts relied upon by traditional machine learning.

[Chauhan et al., 2018] The work here introduces the distinction between conventional machine learning and deep learning, presenting a comparative analysis of various machine learning approaches such as SVM, PCA, LDA, and decision trees.

[Lauzon et al., 2012] This study briefly describes deep learning approaches and provides evidence that a deep learning approach allows for better classification compared to popular classifiers based on handcrafted features.

[Tabar et al., 2016] The study introduces deep learning methods to enhance the classification performance of EEG motor imagery signals. It investigates the use of convolutional neural networks (CNN) and Stacked Autoencoders (SAE) for classifying EEG Motor Imagery signals.

[Hussain et al., 2018] In this work, a CNN architecture model is established to assess its effectiveness in terms of accuracy and efficiency with new

image datasets through Transfer Learning. The retrained model is evaluated, and its performance is compared to some state-of-the-art approaches.

[Albawi et al., 2017] This paper delves into the issues and impacts associated with each parameter of Convolutional Neural Network (CNN). The performance of CNN is contingent on the number of levels, and as this number increases, the output time of levels also rises.

[Zewen et al., 2021] This review introduces the history of CNN, providing an overview of various convolution techniques. It explores advanced CNNs that achieve state-of-the-art results, highlighting essential considerations for function and hyper-parameter selection.

[Yamashita et al., 2018] Convolutional Neural Networks (CNN) stand out as powerful models employed in computer vision tasks, particularly in medical image analysis within radiology. They exhibit capabilities in classifying images, detecting objects, and segmenting regions of interest. Challenges in radiology encompass limited data, interpretability, and data privacy.

[O'Shea et al., 2015] This study provides a concise introduction to CNNs, discussing the architecture of CNN networks. CNNs consist of neurons that self-optimize through learning, with each neuron receiving an input and performing a specific operation.

[Mathew et al., 2020] This work describes the evolution of deep learning, exploring various approaches and architectures of deep learning along with their applications.

[Yong et al., 2019] This study explains the state of the art in deep learning, offering both current insights and a historical perspective. [Lo Bosco et al., 2016] The study compares two deep learning architectures for automatically classifying bacterial species. CNNs demonstrate proficiency in simpler tasks but exhibit lower performance in more complex tasks, where Long Short-Term Memory (LSTM) networks prove more effective.

[Rizzo et al., 2015] This work explores a classifier designed for 16S bacterial genomic sequences, integrating spectral representation and Convolutional Neural Networks (CNN). CNNs demonstrate strong performance on small datasets, achieving high accuracy levels ranging between 95% and 99% for full-length data.

[Nguyen et al., 2016] This study introduces an innovative approach to classifying DNA sequences utilizing Convolutional Neural Networks. One-hot vectors are utilized to represent sequences as input to the model, treating the sequences as textual data.

3. Common Stages of DNA Sequencing:

Downloaded DNA sequences from sources often manifest low-quality DNA sequences, sequence artifacts, and contamination issues, resulting in inaccurate sequencing outcomes. Therefore, it is crucial to preprocess DNA sequence data before inputting it into the classification model to determine the class level of the DNA sequence [25].

3.1. Preprocessing:

Various applications offer features specifically designed for preprocessing sequence datasets. Each application is tailored to handle short read data and can accommodate longer read sequences, providing a range of additional features and functions. During this stage, the DNA sequences obtained from the source undergo preprocessing before being fed into the classification model. Preprocessing involves data-related tasks such as cleaning, transforming, and reducing the sequences [26, 27].

3.2. Feature Extraction:

The next stage in DNA classification is feature extraction, which entails identifying the most pertinent information from the DNA sequence utilizing diverse DNA storage formats. These features serve to construct the most appropriate representation of the input data, encapsulating essential characteristics derived from the raw DNA data.

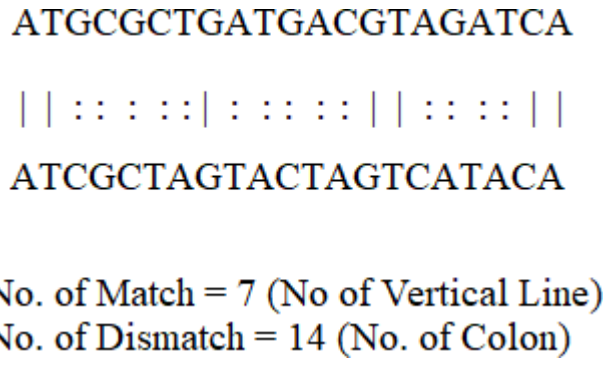

FIGURE 4. Alignment-based DNA sequence analysis (Eisenhofer and Weyrich)

3.3. Training:

In this phase, the DNA classification undergoes the training process to generate the most accurate predictions possible based on the DNA sequence.

3.4. Classification:

During this stage, the DNA sequence data is prepared to be input into the classification model for predicting the classified data.

4. Sequencing Methods for DNA Analysis:

4.1. Alignment Method:

Various alignment methods with distinct characteristics will be explored, encompassing both Alignment-based and Alignment-free approaches. The alignment of DNA sequences plays a pivotal role in numerous molecular biological analyses, posing a fundamental challenge in genomics. These methods involve mapping the letters of a set of sequences to approximate specific relationships. At the sequence level, the comparison of new sequences involves genome alignment and comparative annotation at the gene level. Presently, there are hundreds of publicly available vertebrate genome assemblies (Armstrong et al.).

4.2. Alignment–Based Method:

Alignment based techniques are utilized for DNA sequence classification. This method depends on identifying base-to-base matches in two or more sequences, calculating a score based on the number of matches and mismatches between sequences (Fig. 4). By doing so, they ascertain the class of a given query sequence by identifying the most similar sequence in the known set (Eisenhofer and Weyrich). Several alignment-based tools, such as BLAST (Altschul et al.), FASTA (Lipman and Pearson), and MUSCLE (Thompson, Higgins, and Gibson), are accessible. However, Alignment-based methods have limitations as they do not leverage location information for a sequence within the genome. These methods entail high time-memory computational costs, relying on each nucleotide and necessitating continuous sequences of nucleotides for matching. Additionally, sequence comparisons based solely on alignments can be challenging, particularly when aligning numerous brief reads from different components of the genomes (–alhalem et al.).

4.3. Alignment- Free Method:

The alignment based method does not yield satisfactory results for the ever-expanding volume of genomic data. This is attributed to the increasing diversity of genomic data types. Recognizing the need for location information in DNA sequencing, the alignment-free method addresses the challenges associated with substring alignment. This review paper explores diverse alignment-free methods utilized in DNA sequence analysis, encompassing k-Mer/word frequency, common substring length, spaced word matches, micro-alignments, information theory related methods and graphic representation. These approaches find applications in clustering and classifying sequences, with examples of alignment-free software applications like CASTOR and FFP provided [34, 35, 36]. However, it is crucial to acknowledge that alignment-free methods face limitations, including issues with memory overlapping and a scarcity of implementation software.

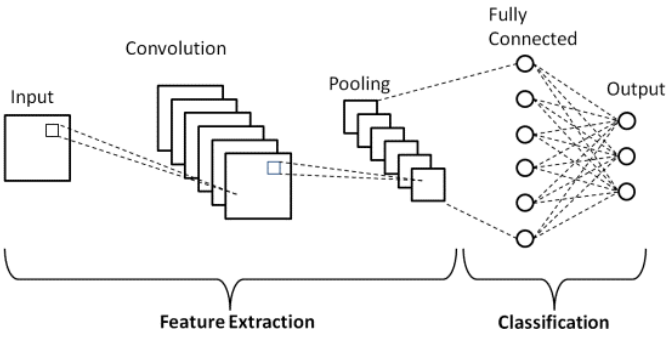

FIGURE 5. Convolution Neural Network Architecture (Yamashita et al.).

5. Deep Learning(DL) :

Deep learning, as a subset of machine learning, emulates the functioning of the human brain by employing neural networks. These networks extract significant features from raw data to perform classification tasks. Deep learning possesses the capability to enhance itself by scrutinizing computer algorithms (Fan, Ma, and Zhong).

In various computer vision applications, architectures such as deep neural networks, reinforcement learning, and recurrent neural networks have demonstrated remarkable results, occasionally surpassing human expert performance (Chauhan and Singh Lauzon and Quintal). Domains like image processing, including speech recognition, face

recognition, object detection, and biomedical applications, have experienced enhanced accuracy due to deep learning methods. In comparison to other state-of-the-art approaches, deep learning consistently provides superior classification performance (Tabar and Halici).

5.1. Convolution Neural Network:

The Convolutional Neural Network (CNN) emerges as an impressive form of Artificial Neural Network architecture, particularly adept at addressing image-driven pattern recognition tasks. With numerous connections and layers, CNN operates in a manner akin to the human brain. It falls under the category of Deep Neural Networks, excelling in recognizing and classifying features from images (Yamashita et al.). CNN takes images as input, finding applications ranging from image and video recognition to image classification, medical image analysis, computer vision, and natural language processing (Hussain, Bird, and Faria).

The CNN architecture comprises two main stages: Feature Extraction and Classification. In the Feature Extraction stage, diverse features are derived from input images through convolution mathematical operations utilizing a specific filter size M X M (O'shea and Nash). As the filter traverses the input image, the dot product is computed between the filter and corresponding parts of the input image, matching the filter's size M X M. The resulting convolution output is then directed to the subsequent Fully Connected (FC) layer, integrating weights, biases, and neurons that establish connections between different layers, preparing the output for classification (Li et al.). The convolution layer plays a pivotal role in CNN by executing the essential task of feature extraction, consuming a significant portion of the network's processing time. The network's performance is contingent on the number of layers, with an increase potentially leading to prolonged training and testing times. The activation function also stands out as a critical parameter in the CNN model (Albawi, Mohammed, and Al-Zawi Yamashita et al.).

5.2. Recurrent Neural network:

Recurrent neural networks (RNN) represent a distinctive architecture in Deep Learning, characterized by feedforward neural networks augmented with edges to capture sequence dynamics through cycles in the network of nodes. RNNs are commonly employed in tasks involving sound or sequential data, such as speech recognition, natural language processing, and sentiment analysis. These networks are dynamic, with their state continuously evolving until equilibrium is reached. Distinguishing themselves from Feedforward Neural Networks (FFNNs), Recurrent Neural Networks (RNNs) allow for feedback between nodes. The computation of each hidden node involves a joint calculation of the input value and the information generated in preceding nodes [46, 47].

Various versions of RNN exist, and one of them is the Long Short-Term Memory (LSTM), designed to address the vanishing gradient problem commonly encountered in basic RNNs (Sherstinsky). LSTMs incorporate components known as "gates" to mitigate the vanishing gradient issue. Another noteworthy variant, the Hopfield network, utilizes bidirectional connections between nodes and is applied to solve various mathematical problems, such as the traveling salesman problem. LSTM networks can be broadly classified into two types: LSTM-dominated networks and integrated LSTM networks (Yu et al.).

6. Deep Learning Networks for DNA Sequence Classification:

6.1. CNN DNA Sequence Classification

In the architecture of the Convolutional Neural Network (CNN), the initial layer incorporates an embedding layer that takes DNA sequence data as input and undergoes a transformation, rendering it into a 2D image via one-hot encoding (Nguyen et al.). Above this input layer, two subsequent layers consist of convolutional layers, each followed by a max-pooling layer. These convolutional layers employ filters of size 5, with a progressive increase in the number of filters—specifically 10 and 20, respectively. The width and stride of the pooling layers both extend across 5 timesteps, treating a one-hot vector as a single unit. Stacked on these convolutional layers are two fully connected layers. The first layer comprises 500 units and utilizes the tanh activation function, in harmony with the activation employed in the lower levels. The second layer functions as the classification layer, employing the Soft Max activation (Lobosco and Gangi Rizzo et al.).

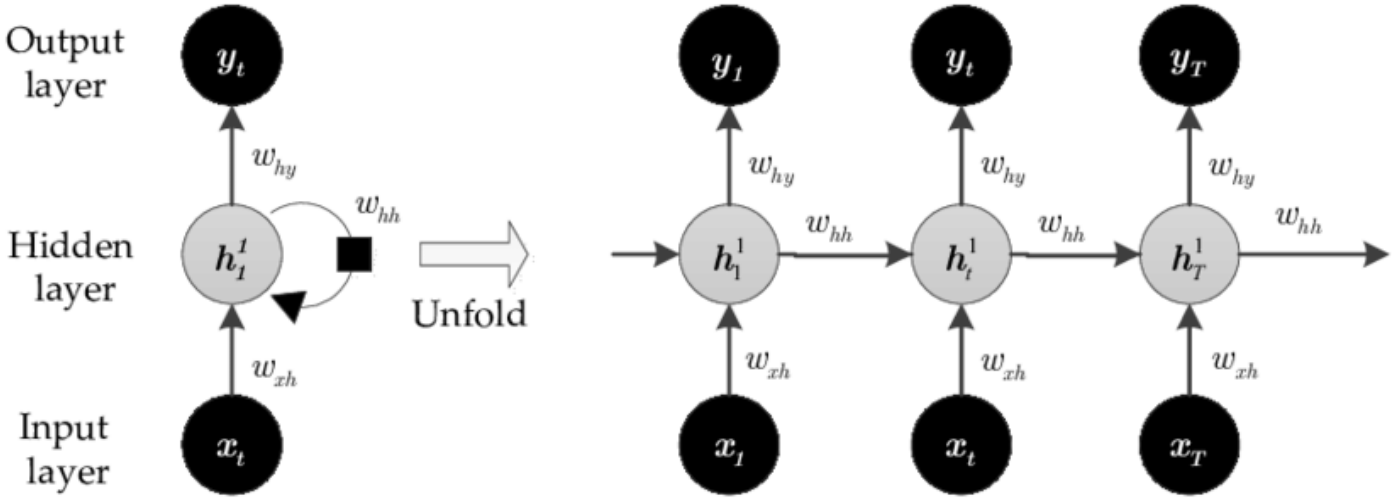

FIGURE 6. RNN -Recurrent neural networks (Mathew, Amudha, and Sivakumari)

6.2. RNN DNA Sequence Classification:

The Recurrent Neural Network comprises six layers, taking input in the form of one-hot encoding vectors. The initial layer includes an embedding layer, succeeded by a max-pooling layer with a size of 2. This max-pooling operation reduces computational load and introduces translational invariance to the network. Following max-pooling, a Long Short-Term Memory (LSTM) model serves as a recurrent layer. The LSTM processes input from left to right, generating a 20-sized output vector at each time step. Subsequent to the LSTM layer, another max-pooling layer is incorporated to extract class-level information. In the subsequent level, bidirectional Long Short-Term Memory is employed to enable the neural network to process sequence information in both directions, backward and forward.

7. Observation and Conclusion:

In this extensive investigation, a meticulous selection process was employed to categorize a total number of sequences into five taxonomic ranks: Phylum (the most coarse-grained), Class, Order, Family, and Genus (finer-grained). The resulting classes exhibited variability, ranging from a minimum of 3 classes (Phylum) to a maximum of 393 classes (Genus). The intermediate levels of Class, Order, and Family encompassed 6, 22, and 65 classes, respectively.

In the context of alignment methods, alignment-based approaches are found to be straightforward for implementation but are suitable primarily for small datasets, demanding considerable processing time for limited data. In the realm of deep learning-based models, in single-task scenarios, CNN outperforms RNN. However, in multi-task situations, RNN demonstrates superior performance and reduced training time. CNN achieves optimal results for small, closely related datasets, with accuracy levels ranging between 95–99% for entire length data.

Multitask learning significantly influences models in terms of both performance and training time, making LSTM a preferred choice. Among CNN, CNN-LSTM, and CNN-bidirectional LSTM, CNN and CNN-Bidirectional LSTM with K-Mer encoding achieve high accuracy, standing at 93.16% and 93.13%, respectively. Evaluation metrics such as precision, recall, sensitivity, and specificity were taken into consideration.

It is crucial to note a limitation of deep learning, emphasizing its reliance on substantial data availability for effective training, shining most brightly in scenarios with abundant training data.

.

References

Albawi, Saad, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network". *2017 International Conference on Engineering and Technology (ICET)* (2017): 1–6.

–alhalem, Samia M Abd, et al. "DNA Sequences Classification with Deep Learning: A Survey". *Menoufia Journal of Electronic Engineering Research* 30.1 (2021): 41–51.

Altschul, Stephen F, et al. "Basic local alignment search tool". *Journal of Molecular Biology* 215.3 (1990): 403–410.

Anastassiou, Dimitris. "Genomic signal processing". *IEEE Signal Processing Magazine* 18.4 (2001): 8–20.

Armstrong, Joel, et al. "Whole-Genome Alignment and Comparative Annotation". *Annual Review of Animal Biosciences* 7.1 (2019): 41–64.

Chauhan, Nitin Kumar and Krishna Singh. "A Review on Conventional Machine Learning vs Deep Learning". *2018 International Conference on Computing, Power and Communication Technologies (GUCON)* (2018): 347–352.

Dölz, Reinhard. "GCG: comparison of sequences". *Computer Analysis of Sequence Data* (1994): 65–82.

Dunn, Graham and Brian S Everitt. "An introduction to mathematical taxonomy". *Courier Corporation* (2004).

Eisenhofer, Raphael and Laura Susan Weyrich. "Assessing alignment-based taxonomic classification of ancient microbial DNA". *PeerJ* 7 (2019): e6594–e6594.

Fan, Jianqing, Cong Ma, and Yiqiao Zhong. "A Selective Overview of Deep Learning". *Statistical Science* 36.2 (2021): 264–264.

Godfray, H and J Charles. "Challenges for taxonomy". *Nature* 417.6884 (2002): 17–19.

Graham, Duncan and Karen Faulds. "Quantitative SERRS for DNA sequence analysis". *Chemical Society Reviews* 37.5 (2008): 1042–1042.

Hartwell, Leland, et al. *Genetics: from genes to genomes*. New York, NY, USA: McGraw-Hill Education, 2018.

Hussain, Mahbub, Jordan J Bird, and Diego R Faria. "A Study on CNN Transfer Learning for Image Classification". *Advances in Intelligent Systems and Computing*. Springer International Publishing, 2019. 191–202.

Jovanović, Jelena. "DNA and CODIS in Organized Crime Investigations in Montenegro". *Kriminalističke teme* 2.2 (2021): 1–19.

Kaloudas, Dimitrios, Nikolet Pavlova, and Robert Penchovsky. "EBWS: Essential Bioinformatics Web Services for Sequence Analyses". *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 16.3 (2019): 942–953.

Kryukov, Kirill, et al. "Nucleotide Archival Format (NAF) enables efficient lossless reference-free compression of DNA sequences". *Bioinformatics* 35.19 (2019): 3826–3828.

Lauzon, Francis and Quintal. "An introduction to deep learning". *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)* (2012): 1438–1439.

Li, Zewen, et al. "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects". *IEEE Transactions on Neural Networks and Learning Systems* 33.12 (2022): 6999–7019.

Lipman, David J and William R Pearson. "Rapid and Sensitive Protein Similarity Searches". *Science* 227.4693 (1985): 1435–1441.

Lobosco, Giosuè and Mattia Antonino Di Gangi. "Deep learning architectures for DNA sequence classification". *International Workshop on Fuzzy Logic and Applications*. Springer, 2016. 162–171.

Luscombe, N M, D Greenbaum, and M Gerstein. "What is bioinformatics? An introduction and overview". *Yearbook of Medical Informatics* 10.01 (2001): 83–100.

Mathew, Amitha, P Amudha, and S Sivakumari. "Deep Learning Techniques: An Overview". *Advances in Intelligent Systems and Computing*. Springer Singapore, 2021. 599–608.

Nguyen, Ngoc Giang, et al. "DNA Sequence Classification by Convolutional Neural Network". *Journal of Biomedical Science and Engineering* 09.05 (2016): 280–286.

Pearson, William R. "Using the FASTA Program to Search Protein and DNA Sequence Databases". *Computer Analysis of Sequence Data* (1994): 365–389.

Penacino, G A. "Organizing the Argentinean Combined DNA Index System (CODIS)". *International Congress Series* 1288 (2006): 780–782.

Rizzo, Riccardo, et al. "A Deep Learning Approach to DNA Sequence Classification". *Computational Intelligence Methods for Bioinformatics and Biostatistics*. Springer International Publishing, 2016. 129–140.

Ruitberg, Christian M, et al. "STRBase: a short tandem repeat DNA database for the human identity testing community". *Nucleic Acids Research* 29.1 (2001): 320–322.

Shen, Wei, et al. "SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation". *PLOS ONE* 11.10 (2016): e0163962–e0163962.

Sherstinsky, Alex. "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network". *Physica D: Nonlinear Phenomena* 404 (2020): 132306–132306.

Tabar, Yousef Rezaei and Ugur Halici. "A novel deep learning approach for classification of EEG motor imagery signals". *Journal of Neural Engineering* 14.1 (2017): 016003–016003.

Thompson, Julie D, Desmond G Higgins, and Toby J Gibson. "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice". *Nucleic Acids Research* 22.22 (1994): 4673–4680.

Travers, Andrew and Georgi Muskhelishvili. "DNA structure and promoter function". *Biochemical Society Transactions* 14.2 (1986): 199–200.

Womble, David D. "GCG". *Bioinformatics methods and protocols*. Humana Press, 2000. 3–22.

Wski. "Compression of DNA sequence reads in FASTQ format". *Bioinformatics* 27.6 (2011): 860–862.

Yamashita, Rikiya, et al. "Convolutional neural networks: an overview and application in radiology". *Insights into Imaging* 9.4 (2018): 611–629.

Yang, Aimin, et al. "Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA". *Frontiers in Bioengineering and Biotechnology* 8 (2020): 1032–1032.

Yu, Yong, et al. "A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures". *Neural Computation* 31.7 (2019): 1235–1270.

Embargo period: The article has no embargo period.

To cite this Article: Kumar Gupta, Rohit, Dr. Sweeti Sah, Dr B. Surendiran , Dr. Shankar Narayan, and Dr Arunkumar P . "**A Survey on Deep Learning Approaches Used in Genomics**." International Research Journal on Advanced Science Hub 05.11 November (2023): 397–408. http://dx.doi.org/10.47392/IRJASH.2023.072