International Journal of Engineering Research and Science & Technology

IJERST

ISSN : 2319-5991

www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 1, 2025

OPTIMIZING VIRAL DNA SEQUENCE CLASSIFICATION WITH DEEP LEARNING AND GENETIC ALGORITHMS

[1]S. Bavankumar, [2]Dr. V. Rathikarani, [3]Dr. R. Santhoshkumar

[1]Research Scholar, [2]Assistant Professor, [3]Associate Professor
[1,2,3]Department of Computer Science and Engineering,
[1,2]Annamalai University
[3]St. Martin's Engineering College, Secunderabad, Telangana, India

*Corresponding Author

E-mail : sbavankumarcse@smec.ac.in

ABSTRACT

DNA sequence classification plays a vital role in biological data analysis, especially in identifying and categorizing novel viral genomes. Accurate classification of these sequences is essential for mitigating the risks of viral outbreaks, such as COVID-19, by expediting vaccine development. This study introduces a hybrid deep learning model designed to improve the efficiency and accuracy of viral DNA sequence classification. The proposed model combines Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) and Bidirectional CNN-LSTM architectures. To further enhance performance, a Genetic Algorithm (GA) was employed for optimizing the weights of the CNN. GA was selected due to its capability to navigate complex search spaces effectively, boosting the model's feature extraction capabilities. Three encoding techniques were investigated to transform DNA sequences into numerical formats suitable for model input: k-mer encoding, label encoding, and one-hot vector encoding. Additionally, an advanced oversampling method was applied to address the issue of imbalanced datasets. Among the tested configurations, the GA-optimized CNN hybrid model using label encoding achieved the highest classification accuracy of 94.88%, outperforming other encoding methods.

Keywords: Deep learning, viral genome classification, convolutional neural networks, genetic algorithm, DNA sequence encoding.

I. INTRODUCTION

Machine learning (ML) approaches have had a profound impact on bioinformatics, which has contributed significantly to its success. Deep learning (DL) techniques emerged as a branch of machine learning (ML) with a vast amount of data being generated. These techniques are thought to be more successful and efficient when we deal with big amounts of data [1]. In recent times deep learning has demonstrated remarkable progress in various domains such as natural language processing and computer vision now the preferred method for genomics modeling tasks, including predicting genetic variation's impact on gene regulatory mechanisms [3]. Many fields including genomics have been transformed by deep learning a subset of machine learning. Next- generation sequencing (NGS) methods are crucial in biological and medical research, requiring processing and analysis methods for variant calling, metagenomic classification, genomic feature detection, and downstream analysis. Machine-learning techniques, particularly deep learning, have gained traction for these tasks.[2] It refers to the large-scale analysis and interpretation of biological data in genomics through the use of neural networks and other techniques. DL models are capable of recognizing complex patterns and producing precise predictions in a range of genomics tasks.

For DNA sequence analysis tasks, including gene expression analysis, DNA-protein binding site prediction, and protein structure prediction, several CNN models have been put forth. Researchers can find complex genetic insights by using deep learning models which could transform biological research, disease understanding, and personalized medicine.

The purpose of this paper is to examine the challenges that arise and the applications of deep learning in DNA sequence classification. Our aim is to investigate how deep learning models can address the complexities and high dimensionality of DNA sequences, and how they can improve the accuracy and efficiency of classification. The study explores deep learning techniques' potential to improve genetic data understanding and advance biomedical endeavors like disease diagnosis, drug discovery, and personalized treatment strategies. With the continuous evolution of deep learning methodologies, the integration of CNNs into genomic research promises further advancements in deciphering the intricacies of genetic data and its implications for human health.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

II. LITERATURE SURVEY

The intention of this literature review is to offer an extensive understanding of the current state of knowledge in the field by analyzing the existing research on DNA sequence classification using deep learning.

In this paper [10] the authors have explored a novel hybrid deep learning approach that uses a genetic algorithm (GA) for optimizing weight within the framework of convolutional neural networks (CNNs) to effectively classify viral DNA sequences. additionally uses three different encoding techniques such as k-mer and label encoding to investigate the combination of Bidirectional CNN-LSTM and Long Short- Term Memory (LSTM) model architectures. and vector encoding in one- hot mode. With label encoding, the authors GA-optimized CNN hybrid model—which attains the highest classification accuracy of 96. 88 percent.

In this paper [11], the author has discussed a hybrid computational recurrent neural network (CNN/RNN) architecture, CRPTS, is used to predict TFBSs on 66 in vitro datasets by combining DNA sequence and shape features. Here method efficiently extracts features from large-scale genomic sequences, finds common patterns, and captures local structural information without relying on DNA shape data.
In this paper [12], The author has talked about a technique called PDBP-Fusion that combines CNN and Bi-LSTM to predict proteins that bind DNA sequence. A bi-directional long-short-term memory network (Bi- LSTM) is used to store important long-term dependencies in context and CNN is used
to learn local characteristics. The PDBP-Fusion technique can predict DBPs with 86.45% sensitivity, 79.13% specificity, 82.81% accuracy, and 0.661 MCC on the PDB14189 benchmark dataset. In contrast to other advanced prediction models, there has been a minimum 9.1% rise in the suggested approaches.

In this study [13], the author has investigated the usage of CNN CNN-LSTM and CNN-Bidirectional LSTM architectures with encoding approaches for DNA sequence categorization [13]. Testing accuracy for label encoding is lower than training and validation accuracy, and the CNN and CNN Bidirectional LSTM models with k-mer encoding showed good accuracy with test data scores of 93. 16% and
93. 13%, respectively.

The author of this paper [14] has investigated a novel method for feature extraction from DNA sequences based on a hot vector matrix, as well as a machine learning- based classifier. Four different classifiers: Recurrent neural networks (RNN), Decision Trees, K-nearest neighbor (KNN) algorithms, Support Vector Machines (SVM), and Convolution neural networks (CNN) were examined. based on several factors, such as accuracy and precision rate, and the outcome reveals 93.9% accuracy.
The author of this paper [15] has studied methodology used for classifying various functional genome types, such as coding regions (CDS), long noncoding regions (LNC), and pseudogenes (PSD), in genomic data using genomic signal processing techniques to transform nucleotide sequences into graphical representations of the information contained in DNA sequences. Convolutional deep learning models were then employed. The results showed accuracy scores of 83% and 84% when differentiating between CDS vs. CDS and LNC in contrast. PSD, correspondingly.

III. IMPORTANCE OF DNA SEQUENCE ANALYSIS

DNA sequence analysis is significant because many fields, including agriculture, medicine, and evolutionary biology, depend on an understanding of genetic data. It comprises examining how the nucleotides (A, T, C, and G) are arranged in a DNA molecule. DNA sequence analysis provides insights into traits, diseases, and drug responses by enabling researchers to analyze genetic variations within individuals and populations through the use of SNPs, insertions, deletions, and structural variations. utilized in forensic science as well as DNA profiling to identify individuals, solve crimes, and create family ties [16][17]. Researchers can determine the functional components of genomes, deduce relationships between species' evolutionary histories, and learn more about the genetic underpinnings of disease through DNA sequence analysis.
Traditional DNA sequence analysis techniques are growing more and more ineffective because the amount of genomic data is growing exponentially. Consequently, the genomics community is increasingly turning its attention to deep learning techniques. More precise predictions of functional elements, regulatory regions, and protein- binding sites are made possible by deep learning algorithms like convolutional neural networks (CNNs) and RNNs, or recurrent neural networks, which actually adept at deciphering intricate patterns in DNA sequences [18]. Personalized medicine advances when deep learning algorithms make it easier to integrate genomic data with clinical information. Predicting patient outcomes, disease risks, and drug reactions is made possible by this. [19] From DNA sequence data, deep learning models can infer gene regulatory networks, detect regulatory motifs,

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

and predict gene expression levels, which helps to streamline our understanding of gene function and regulation. [20]. Deep learning-based models aid in virtual screening, drug target identification, and drug repurposing by predicting molecular interactions, drug–target binding affinities, and compound bioactivity profiles from genomic and chemical data. [22].

IV. CHALLENGES IN DNA SEQUENCE CLASSIFICATION

Classifying DNA sequences is essential for many genomic applications, including variant calling, gene prediction, and disease detection. Deep learning has revolutionized genomic research by producing sophisticated models for DNA sequence analysis. However, classifying DNA sequences accurately presents several challenges for deep learning algorithms as shown in Fig. 1. By their very nature, DNA sequences are discrete and symbolic, consisting of four adenine (A), cytosine (C), guanine (G), and thymine (T) are the nucleotide bases. The following are a few challenges in classifying DNA sequences: DNA sequences differ from one another in terms of length, complex pattern, and biological noise. Capturing the intricate correlations and patterns present in DNA sequences is a significant challenge for deep learning models. Realistic models that can learn and reflect these complex properties are necessary because DNA sequences are non-linear and hierarchical [23]. The existence of mistakes and noise in DNA sequencing data is one of the main problems with DNA sequence categorization. To make sure that the incoming data is accurate and reliable, preprocessing methods including normalization, error correction, data cleaning, and quality control are crucial. Managing ambiguous bases, sequencing artifacts, and missing values is essential to creating clean, standardized datasets for deep learning models.[24]

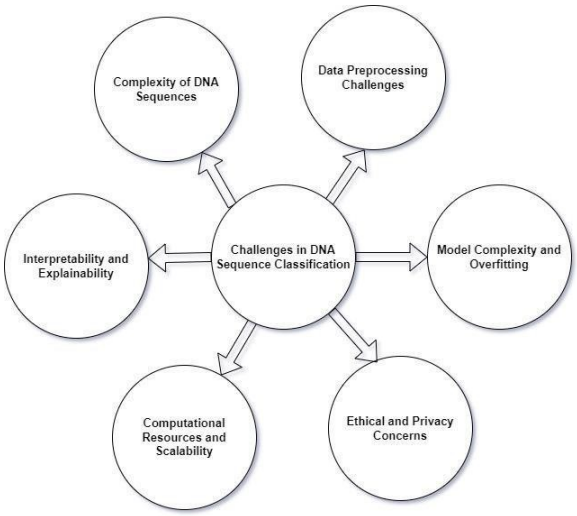

Fig 1: Challenges in DNA sequence classification

It is imperative to refine models and employ domain adaptation strategies to fully utilize the insights gained from current datasets while tackling novel genomic classification jobs. Concerns about permission, data security, and privacy are among the ethical issues brought up by the classification.

V. FEATURE SELECTION TECHNIQUES

To extract the most pertinent and instructive features from a high-dimensional and large-scale dataset, feature selection is an essential stage in the categorization of structural and functional characteristics of DNA sequences. There are four typical methods for feature selection in DNA sequence classification. The first one is Frequency-Based Methods which assess how frequently particular patterns or motifs appear in DNA sequences and also help to discover traits that are either strongly represented or exhibit notable differences in their occurrence across various sequence classes. The second method used for feature selection is Information theory- based methods, like entropy and mutual information, which measure the amount of information that each feature adds to the target variable. These methods select the features that provide the most discriminating information by weighing each one's relevance and redundancy.

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

VI. DATA PREPROCESSING TECHNIQUES

Data preprocessing is essential in the field of DNA sequence classification because it converts unprocessed and raw data into a format that machine learning models can easily parse and analyze. There are a few techniques used for preparing data before classifying DNA sequences as shown in Fig. 2. They are as follows:

1. Data Cleaning

Data cleaning is a process of fixing and correcting errors or inconsistencies in the data, such as duplicates, outliers, and missing numbers and values. There are a few techniques used for data cleaning, such as imputation, noise removal, and transformation

2. Data Integration

The process of merging data from various sources to produce a single, cohesive dataset is known as data integration. This can be challenging when managing data with different formats, structures, and meanings.

3. Feature Selection

An essential and significant phase in the categorization of DNA sequences is feature selection. It entails deciding which characteristics or aspects are most pertinent to the categorization process. By doing this, the dimensionality of the data is decreased, and the classification models' effectiveness and accuracy are increased.

4. Encoding Methods

DNA sequences are represented using different encoding methods in a way that machine learning models such as deep learning models can understand. Label encoding, One-hot encoding, k-mer encoding are frequently utilized encoding techniques. Label encoding maintains the positional information of a DNA sequence by giving each nucleotide a distinct index value. For the sequences, K-mers are produced, which stand for subsequences of length k.

5. Imbalanced Datasets

It describes datasets in which the number of samples in one class is much lower than in the others; classifying DNA sequences in these kinds of datasets can be difficult. Methods such as the Synthetic Minority Over-sampling Technique (SMOTE) can be used to generate synthetic samples for the minority class to address this issue.

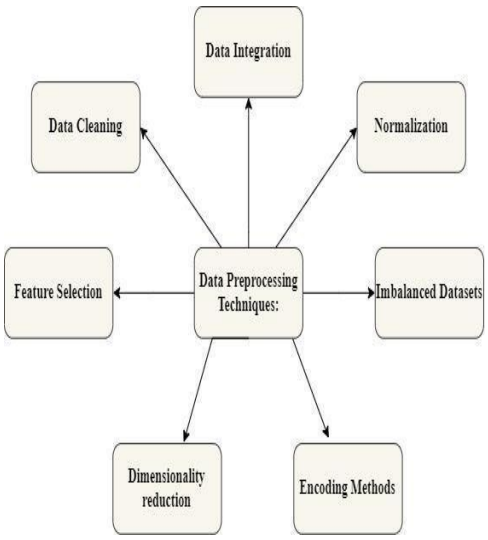

Fig 2: Data Preprocessing Techniques

6. Normalization

Normalization is the process of bringing feature values within a dataset to a common predetermined range [0,1] or any other range. This ensures that characteristics with different scales do not dominate learning. Decimal scaling, z-score normalization, and min- max normalization are examples of common normalization methods.

7. Dimensionality Reduction

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

Dimensionality reduction is to decrease a dataset's feature count without compromising the most important information and characteristics. This helps improve the performance of the classification model and reduce computational complexity. To minimize dimensionality, methods like Principal Component Analysis (PCA) and t-SNE (t-distributed Stochastic Neighbour Embedding) are frequently employed.

VII. DEEP LEARNING MODELS FOR DNA SEQUENCE CLASSIFICATION

Because deep learning algorithms can automatically extract significant features from the input data, they have shown great promise in DNA sequence classification problems. The following are a few popular deep-learning models used for classifying DNA sequences:

1.Convolutional Neural Networks (CNN) Convolutional neural networks or ConvNet perform better when applied to analyze image, speech, or audio signal inputs than other types of neural networks. In CNN, 1D CNNs are used for text classification, 2D CNNs are used for image and audio classification, and for video classification 3D CNNs are commonly used. CNNs employ several layers like convolutional layer, pooling layer, fully connected layer as shown in Fig-3, each of which picks up unique characteristics from an input. A CNN may have hundreds, thousands, or even more layers, depending on how complicated the task for which it is designed is. Each layer builds on the outputs of the one before it to identify intricate patterns. CNN is extensively utilized in the classification of DNA sequences. By using convolutional filters to extract hierarchical structures, functions, and local patterns from the data, they can extract abstract characteristics. Tasks like chromatin effects prediction and protein binding prediction have proven successful for CNNs.

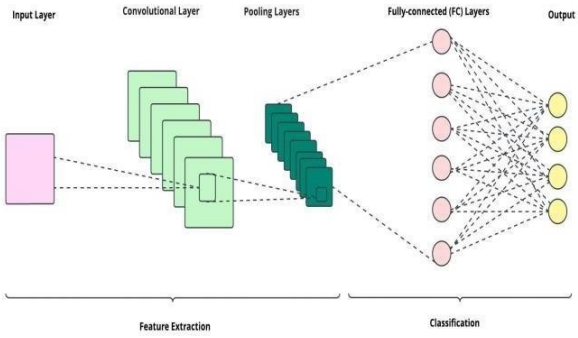

Fig 3: Convolutional Neural Network (CNN)[26]

2. Recurrent Neural Network (RNN)

RNN is a type of neural network in which the output of the previous step is fed into the current step. In a traditional

neural network model, all input data and output data are independent of each other. However, if you need to guess the next word in the sentence, the previous word is necessary and therefore should be kept. Therefore, RNN appeared, which uses a secret algorithm to solve this problem

3. Hybrid Models

To take advantage of each deep learning architecture's advantages, hybrid models blend them. For instance, a CNN and LSTM combination can identify long-range as well as local relationships in DNA sequences. To increase classification performance, hybrid models have been applied to the classification of DNA sequences.

VIII. COMPARATIVE ANALYSIS: CNNS WITH ENCODING TECHNIQUES VS.HYBRID MODELS OR RNNS

For DNA sequence classification, it is hypothesized that CNNs with encoding techniques can perform better than hybrid models or RNNs based on the data that is currently available. CNN and RNN models for DNA sequence classification frequently use encoding techniques, like one-hot encoding [25].

It has been demonstrated that CNNs using encoding techniques can achieve high accuracy in DNA sequence classification tasks. For instance, in one study, a CNN model classified DNA sequences with an accuracy of 93.16 percent [13]. According to a different study, CNN-based models outperformed alternative techniques in the prediction of transcription factor binding sites. The ability of CNNs to capture local patterns and dependencies in DNA sequences makes them well-suited for this task. However, it is important to note that

International Journal of Engineering Research and Science & Technology
ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

CNNs alone are effective in extracting features for classification tasks in DNA sequence analysis. Hybrid models that combine CNNs and RNNs can capture both local and long-term dependencies.
So, encoding-based CNNs have demonstrated encouraging performance in DNA sequence classification tasks. They are a good fit for this task because of their capacity to identify local patterns and extract features from DNA sequences. While RNNs and hybrid models each have advantages, CNNs that use encoding techniques have been proven to perform better in particular situations.

IX. APPLICATIONS OF DEEP LEARNING IN DNA SEQUENCE CLASSIFICATION

Because Computational models and deep learning models offer strong tools for analyzing and interpreting DNA sequences, the fields of genomics and bioinformatics have completely changed. Deep learning has a wide range of significant applications in DNA sequence categorization that help scientists forecast genetic changes, answer intricate biological problems, and find patterns in genomic data that may otherwise go unnoticed. Let's examine a few important uses of deep learning for the classification of DNA sequences:

1. DNA Sequences Classification

Deep learning models have been applied to DNA sequence classification problems, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs). DNA sequences can be accurately classified into several groups or classes. This classification can be beneficial for both structural and functional analysis which are used for motif detection, genomic variant detection, gene function prediction, disease association studies, and drug target identification [7]. Because of these models' multilayer features, they can be able to automatically learn and extract information from the given input sequences.

2. Variant Calling

The process of locating genetic variants in DNA sequences is known as variant calling. Deep learning algorithms have been used for this purpose. Google developed DeepVariant, a deep learning variant caller that uses techniques for image categorization in variant calling and treats mapped sequencing information as images. This method has demonstrated increased precision in identifying indels and single-nucleotide variations.

3. Prediction of Transcription Factor Binding Sites

DNA sequences have been subjected to transcription factor binding site predictions using deep learning models, most notably convolutional neural networks (CNNs). DNA sequences have local patterns and dependencies that CNNs may identify, making binding sites more precisely identified.

4. Classification of Genomic Signals

Deep learning models have been used to classify and analyze genomic signals. Recurrent neural networks (RNNs) and other deep learning architectures, for instance, have been used to categorize genomic signals and find complex patterns connected to certain genomic events using the input sequences [8].

5. Taxonomic Classification

DNA sequences are categorized into several taxonomic categories using deep learning models, which have been utilized for this purpose. To enhance classification performance, hybrid models that blend deep learning architectures with additional methods, including BERT embeddings, have been applied [9].

6. Viral Categorization

Based on DNA or cDNA sequences, deep learning techniques have been applied to the categorization of viruses. Deep learning models have been trained on visual representations of genome sequences, such as k-mer images, to perform viral classification tasks.

X. CONCLUSION

In conclusion, there are many obstacles to overcome in the process of classifying DNA sequences using deep learning techniques, which call for careful thought and creative solutions. Robust model design and optimization are crucial due to the variability of sequence lengths, the intricacy of data representation, and the challenge of extracting and selecting meaningful features. Furthermore, creating precise and trustworthy classifiers requires addressing problems like overfitting, interpretability, and unbalanced data. Some of these difficulties may be overcome by utilizing transfer learning and incorporating domain expertise from genomics, but more study is needed to progress the field and fully realize the promise of deep learning for DNA sequence classification. Deep learning in genomics has the potential to completely change our understanding of genetic information and its role in health, disease, and biological processes, despite these obstacles. By tackling these obstacles head-on, scientists can open the door for revolutionary discoveries in genomic medicine and other fields.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 1, 2025

REFERENCES

1. Koumakis, Lefteris. "Deep learning models in genomics; are we there yet?." Computational and Structural Biotechnology Journal 18 (2020): 1466-1473.
2. Schmidt, Bertil, and Andreas Hildebrandt. "Deep learning in next-generation sequencing." Drug discovery today
26.1 (2021): 173-180.
3. Eraslan, Gökcen, et al. "Deep learning: new computational modelling techniques for genomics." Nature Reviews Genetics 20.7 (2019): 389-403.
4. Nguyen, Ngoc G., et al. "DNA sequence classification by convolutional neural network." Journal Biomedical Science and Engineering 9.5 (2016): 280-286.
5. Meharunnisa, M., M. Sornam, and B. Ramesh. "An Optimized Hybrid Model for Classifying Bacterial Genus using an Integrated CNN-RF Approach on 16S rDNA Sequences." (2024).
6. Alharbi, Wardah S., and Mamoon Rashid. "A review of deep learning applications in human genomics using