Automated DNA Classification Using Modern Deep Learning Strategies

[1]Bugga Vinay, [2]Kotari Swamy, [3]Paladugula Sowmya, [4]D Ganesh
[1]Vinaybugga8@gmal.com,[2]kotariswamy1@gmail.com,[3]paladugulasowmya7@gmail.com,[4]dganesh@vardhaman.org
[1 2 3 4]Department Of Computer Science and Engineering
Vardhaman College of Engineering, Hyderabad, India.

Abstract—Advancements in genomics have led to an exponential increase in the availability of DNA sequence data, offering a rich source of information for various biomedical applications, including disease prediction, functional annotation, and evolutionary analysis. Efficient and accurate classification of DNA sequences is paramount to unlocking the hidden knowledge within these vast datasets. This wealth of genetic information presents a remarkable opportunity for advancing our understanding of viral pathogens such as SARS, and MERS. Effective classification of these viral DNA sequences is crucial for epidemiological studies, drug development, and vaccine design. This paper introduces a novel approach, employing Recurrent Residual Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks, for the classification of virus DNA sequences. Our proposed model, RRCNN-LSTM, integrates the strengths of CNNs in extracting local sequence features and LSTMs in capturing long-range dependencies within the sequences. The utilization of residual connections in the CNN component enhances the gradient flow during training, facilitating the modelling of intricate sequence patterns. Additionally, the LSTM component effectively handles variable-length sequences, making the RRCNN-LSTM model versatile for a wide range of DNA sequence classification tasks. This proposed approach gave a remarkable accuracy of 96.27%, which helps in classifying the viruses.

Keywords—DNA sequence classification, genomics, bioinformatics, Viral pathogens, convolutional neural networks, LSTM

I. Introduction

DNA sequence classification is a fundamental task in bioinformatics and computational biology. It involves the categorization of DNA sequences into different classes based on certain patterns within the sequences. This classification can serve various purposes, such as identifying genes, understanding genetic variations, or predicting the function of a given DNA sequence.

At present, there are 10 million individual viruses present on our planet. They are COVID-19, SARS, MERS, Dengue etc. Due to the increase in the DNA sequences, machine learning is used for the classification of DNA. Any living organism is made up of DNA. DNA is made up of four nucleotides they are Adenine (A), thymine (T), cytosine (C), and guanine (G). DNA is unique for every organism.

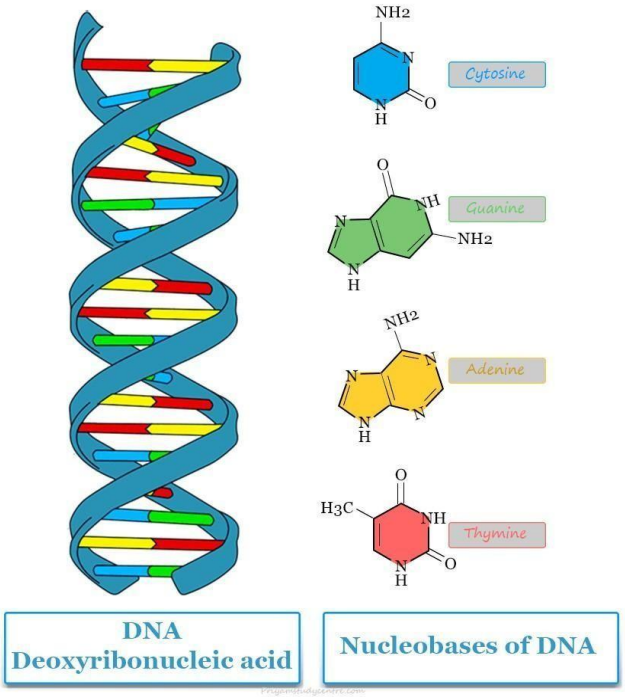

Fig. 1. DNA structure

(https://www.priyamstudycentre.com/2023/08/deoxyribonucleic-acid-dna.html)

DNA is present in a double-helical structure(as shown in Fig. 1). Each base in one strand forms bond with its complementary base on the another strand in a double-stranded DNA. Adenine is paired with thymine and cytosine paired with guanine.

Accurate classification of DNA sequences helps scientists to identify genetic variations, derive meaningful patterns, and

provide insights into biology. In this context, this paper presents an approach to classify DNA sequences by using the capabilities of Recurrent Residual Convolutional Neural Networks (RRCNN) and Long Short-Term Memory (LSTM) networks. Our proposed model, named RRCNN-LSTM, leverages the strengths of both architectures to address the challenges posed by DNA sequences.

The distinctive architecture of RRCNN-LSTM integrates residual connections within the CNN component, which aids in the extraction of fine-grained sequence features while mitigating the vanishing gradient problem. Simultaneously, the LSTM component effectively captures both long and shortterm dependencies, making the model adaptable to DNA sequences of varying lengths. This amalgamation of deep learning techniques not only enhances classification performance but also contributes to the interpretability of the model.

II. Related Work

This section discusses various pieces of literature on the classification of DNA sequences. Many Machine Learning models and Deep learning models are used for clasification.

A statistical liner model was employed to classify hepatitis C viruses proposed by Diallo et al.[1] using linear classification models like Logistic Regression, Markov, Linear SVM, and Multinomial Bayes. The author used these models to classify the HCV dataset's partial and complete genomic sequences. The author also compared the results with various k-mer sizes.

Comparing ML algorithms with or without feature extraction was proposed by Beinke et al.[2]. The author proposed three algorithms namely CNN, DNN, and N-gram probabilistic models. The author employed a feature extraction method based on distance in the DNA sequence and randomly generated sub-sequences of the DNA. The author also evaluated the models using vaious datasets such as influenza, AIDS,hepatitis C, and COVID-19. They observed an accuracy of above 99% for the three models.

Identification of the origin of viruses was employed by Khanh Le et al.[3]. This study showcased about finding the base for mutated viruses DNA. The authors used a hybrid approach namely, an extreme gradient boosting algorithm. This algorithm is used to point out the start of replication of a DNA sequence. The algorithm got an accuracy of 89.51% using XGBoost classifier.

Classification and detection of SARS-CoV-2 by the use of deep learning was introduced by Alejandro et al.[4]. The author coupled deep learning methods with techniques of AI to discover the representation of SARS-CoV-2 genomic sequences. The classifier is trained on 553 sequences. The model showed a promising accuracy of 98.73% with 100% specificity.

S. Aswath et al.[5], the researchers created a machinelearning model designed to enhance the categorization of DNA sequences. Various Nature-Inspired Algorithms (NIA), including the Firefly Algorithm (FA), Grey Wolf Optimization (GWO) Algorithm, Bat Algorithm (BA), and Hybrid Bat Algorithm (HBA), were utilized in this study.

Yiren Wang et al.[6], authors showcased that the accuracy of the gradient boosted tree classifier model is remarkable, achieving high performance levels ranging from approximately 96% for molecules with a single mismatch to 99.5% for those without such variations.

Hemalatha Gunasekaran et al.[7], the study utilized CNN, CNN-LSTM, and CNN-Bidirectional LSTM architectures, employing both Label and K-Mer encoding for the classification of DNA sequences. The maximum accuracy achieved was 93.16%, accomplished by CNN with K-mer encoding.

Iis Setiawan Mangkunegara et al.[8], authors applied the grid search cross-validation optimization technique to fine-tune the SVM classification model.. The proposed accuracy is 77% before optimization and 90% after optimization.

Yusuf Aleshinloye Abass et al.[9]. In this study, authors extracted exons from multiple prostate gene sequences utilized in experiments.Employing a k-mer encoding strategy for DNA sequences alongside one-hot encoding for class labels, a model of bi-LSTM was constructed. The model's predictions yield a 95% training accuracy and a 91% validation accuracy respectively.

M S Antony Vigil et al.[10]. In this study, authors evaluated various machine learning methods including a variety of models including LSTM, SVM, Adaboost, Naive Bayes, Multilayer Perceptron, CNN, Random Forest Classifier, and XGB Classifier for the DNA sequencing task using a human dataset.

III. Proposed Methodology

The methodology for DNA sequence classification using RRCNN-LSTM involves several steps. Firstly, the SARS and MERS dataset is collected. Since the dataset is highly imbalanced SMOTE(Synthetic Minority Oversampling Technique) is performed. This generates the synthetic samples of minority classes like SARS to match with the majority class like MERS. Then the dataset is encoded using a label encoder. Then the dataset is then divided into training set and testing set. Next, a Recurrent Residual Convolution Neural Network(RRCNN) is trained to capture both sequential dependencies and local patterns in DNA sequences. Next, LSTM is employed to capture dependencies over extended periods and sequential patterns within the data, rendering it particularly suitable for the analysis of DNA sequences. The predictions from the RRCNNLSTM are used to classify the new data. The model is evaluated employing metrics such as accuracy, precision, recall, and F1score. Fine-tuning and optimization methods are employed for optimization of the model. This involves exploring parametertuning methods to enhance the RRCNN-LSTM model performance. Overall, this methodology

combines the power of RRCNN and LSTM algorithms to create a robust system for the classification of DNA sequences.

A. Dataset

The DNA sequences of SARS and MERS are obtained from the database specifically, "The National Centre for Biotech-

	Attribute	Class
0	cgttctcctgcagaactttgattttaacgaacttaaataaaagccc...	0
1	aactttgattttaacgaacttaaataaaagccctgttgtttagcgt...	0
2	aagtgaatagcttggctatctcacttcccctcgttctcttgcagaa...	0
3	gatttaagtgaatagcttggctatctcacttcccctcgttctcttg...	0
4	aagtgaatagcttggctatctcacttcccctcgttctcttgcagaa...	0

Fig. 2. sample of dataset

nology Information (NCBI)" (https://www.ncbi.nlm.nih.gov/) . The genomic sequences are in the FASTA format. The sample of the dataset is shown in Fig. 2.

B. Model Implementation

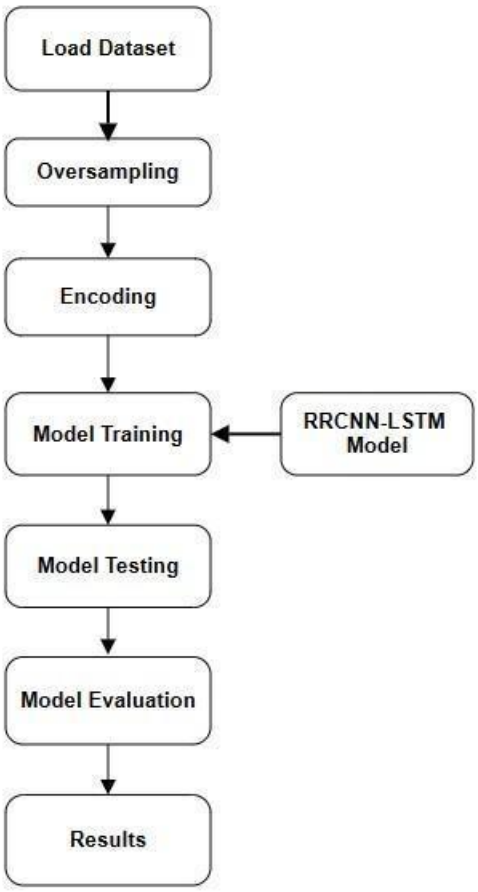

Fig. 3. Block Diagram for DNA sequence classification

- Load Dataset: Acquire and load the DNA Sequence dataset. It contains the DNA sequences of SARS and MERS. There are 1453 sequences of MERS and 674 sequences of SARS.
- Oversampling: Perform Oversampling on the dataset as the DNA sequences are imbalanced. As the DNA sequences of SARS are less compared to MERS, oversampling is performed. SMOTE is applied to SARS as it is the minority class. SMOTE generates synthetic sequences of SARS to match with the majority class i.e. MERS.
- Encoding: Encoding is performed on the dataset to convert the DNA sequences into a format suitable for input into the neural network. Commonly used encoding techniques like one-hot encoding or embedding representations. This step converts A, C, G, and T into numerical representation as shown in Fig. 3. After Encoding, the dataset undergoes partitioning into training and testing sets.

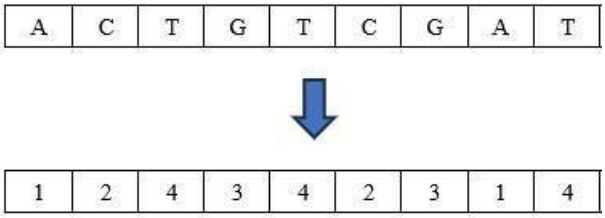

Fig. 4. Encoding of DNA sequence

- Model Training: The RRCNN-LSTM model is chosen due to its ability to capture long-range dependencies. The RRCNN is a hybrid model that combines convolution layers with recurrent layers. LSTM is used to capture sequential patterns. RRCNN is implemented first to capture local features and patterns in the sequences. The output from the RRCNN is fed into LSTM to capture longrange dependencies and sequential patterns. The model undergoes training using the preprocessed training data obtained in the preceding step.
- Model Testing: Test the trained model using the training dataset to evaluate the classification performance of DNA sequences.
- Model Evaluation: Evaluate the Hybrid RRCNN-LSTM using the performance metrics like precision, accuracy, F1 score and recall to assess their performance in classifying the DNA sequences.

C. Models

- Recurrent Residual Convolution Neural Network:: RRCNN, which stands for Recurrent Residual Convolutional Neural Network, is a specific architecture used for DNA sequence classification tasks. It combines the strengths of recurrent neural networks (RNNs) and residual convolutional neural networks (RCNNs) to

capture both sequential dependencies and local patterns in DNA sequences. The architecture of RRCNN shown in Fig. 5, consists of multiple layers, including convolutional layers, Residual, recurrent layers, and fully connected layers.

- Long Short-Term Memory:LSTM (Long Short-Term Memory) shown in Fig. 6, is a type of recurrent neural network (RNN) architecture that is commonly used in DNA sequence classification in combination with

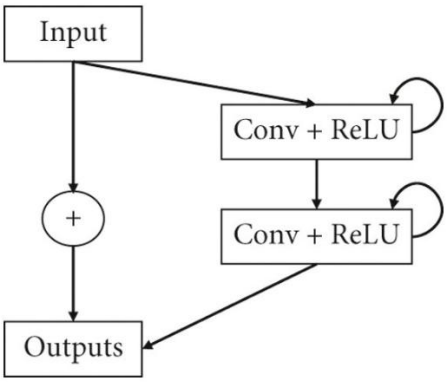

Fig. 5. RRCNN Model

RCNN (Recurrent Convolutional Neural Network). LSTM is specifically crafted to capture prolonged dependencies and sequential patterns in data, rendering it well-suited for the analysis of DNA sequences.. The combination of RRCNN and LSTM leverages the strengths of both models to improve the performance of DNA sequence classification tasks.

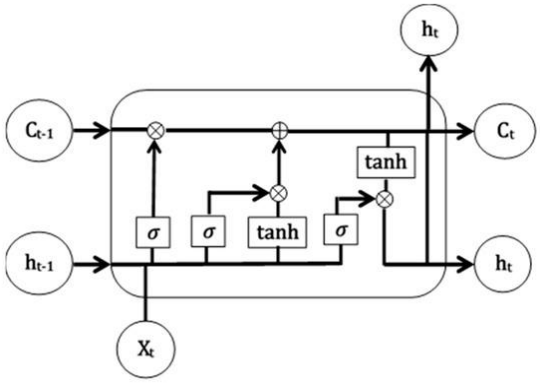

Fig. 6. LSTM Model

Algorithm: Input:

– D: A training dataset with labeled DNA sequences. – C: The set of possible classes (SARS or MERS).

Output:

– RRCNN-LSTM model for DNA sequence classification.

1) Initialize the RRCNN-LSTM model architecture, including layers such as Conv1D, LSTM, and Dense layers.
2) Encode DNA sequences into numerical representations suitable for input D_i in D to the model (e.g., one-hot encoding), and C is used to represent the set of possible class labels.
3) Construct the RRCNN-LSTM model with Conv1D layers for local feature extraction, LSTM layers for capturing sequential patterns, and Dense layers for classification.
4) Define the input layer with the appropriate input shape for DNA sequences.
5) Indicate the loss function (e.g., categorical crossentropy), optimizer (e.g., Adam), and evaluation metric.
6) Compile the RRCNN-LSTM model.
7) Train the RRCNN-LSTM model on the labeled DNA sequence dataset(D) with classes(C).
8) Use the training data to learn the model parameters.
9) Assess the model's performance on a validation set to monitor for overfitting and adjust hyperparameters if necessary.

IV. Results and Discussion

This section provides an in-depth analysis of the outcomes of our DNA sequence classification experiments conducted using the innovative Recurrent Residual CNN-LSTM (RRCNNLSTM) model. The experiments encompassed a diverse and extensive dataset of DNA sequences, embracing a wide range of genetic information, including sequences from viral pathogens like SARS, and MERS. Our primary objective was to rigorously evaluate the model's efficacy in accurately classifying these diverse DNA sequences, and to assess its adaptability and robustness in handling the inherent variability in genomic data, specifically the variable-length nature of genetic sequences.

A. Confusion Matrix

A confusion matrix is a tabular representation that displays the performance of a classification model. It displays the counts or proportions of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. Widely employed in machine learning, it serves as a valuable tool for evaluating the performance of classification tasks. The Confusion matrix for RRCNN-LSTM is:

TABLE I
CONFUSION MATRIX FOR RRCNN-LSTM

	Predicted MERS	Predicted SARS
True MERS	423	13
True SARS	20	430

B. Performance Metrics

When evaluating the performance of DNA sequence classification, several metrics are commonly used to assess its effectiveness. Here are some important performance metrics:

1) Accuracy:
$$AC = \frac{TN + TP}{TP + TN + FN + FP}$$
2) Precision (PR):
$$PR = \frac{TP}{TP + FP}$$
3) Recall (RC):
$$RC = \frac{TP}{TP + FN}$$
4) F1 Score:
$$F1\ Score = \frac{2 \cdot PR \cdot RC}{PR + RC}$$

TABLE II
PERFORMANCE METRICS

	Accuracy	Precision	F1score	Recall
CNN-LSTM	96.27	95	96	97

C. Comparision for Other Models

The models referenced exhibited varying levels of precision and effectiveness, showcasing promising results in classifying DNA sequences. For example, Hemalatha Gunasekaran et al. [11] demonstrated a comparative analysis of four different models which showcased excellent results. Using SVM, Iis Setiawan Mangkunegara and Purwono Purwono[12] showed good results. Umit Murat Akkaya and Habil Kalkan[13] used four different to represent DNA sequences. They proposed a random dictionary approach which showed promising results. Satya Prakash et al.[14], presented a paper about forecasting COVID-19 Pandemic using various approaches. Among them, the best result was obtained by Bidirectional LSTM, Prophet, and Vanilla LSTM.

By contrast, the RRCNN-LSTM algorithm has shown strong performance in classifying DNA sequences. This approach uses recurrent connections and recurrent blocks along with LSTM. The amalgamation of RRCNN-LSTM led to the development of a system for DNA sequence classification, enabling scalable and real-time classification. This system not only consistently recognizes specific sequence patterns but also adeptly manages intricate genetic patterns.

TABLE III
COMPARISION WITH OTHER MODELS

Ref No.	Models	Accuracy
[7]	RCNN	93.16%
[8]	SVM	90%
[9]	Bi-Lstm	91%
	Proposed Method (CNN-LSTM)	96.27%

V. CONCLUSION AND FUTURE SCOPE

In conclusion, the implementation of the Recurrent Residual Convolutional Neural Network (RRCNN) combined with Long Short-Term Memory (LSTM) networks has demonstrated its efficacy in DNA sequence classification. The model showcased remarkable accuracy in categorizing DNA sequences, emphasizing its potential impact in the field of bioinformatics and genomics research. The amalgamation of convolutional and recurrent layers facilitated the capture of both local and long-range dependencies within the sequences, contributing to the model's proficiency in discerning intricate patterns crucial for precise classification. Looking forward, the project's future scope lies in the exploration of attention mechanisms, transfer learning, and alternative neural network architectures to further refine accuracy and address evolving genomic data challenges. Continuous updates and collaboration with domain experts promise to propel the model's adaptability and enhance its applicability in unraveling complex biological insights.

VI. ACKNOWLEDGMENT

I want to convey my heartfelt appreciation to my research guide, Mr.D Ganesh, whose expert guidance and continuous support have been invaluable throughout the course of this research. His insightful feedback and encouragement have significantly shaped the direction and quality of this work. I also extend my thanks to Dr.Ramesh Karnati , Head of Department, Department of Computer Science and Engineering, for providing a conducive academic environment and facilitating the necessary resources for the completion of this research. Their collective mentorship and leadership have played a pivotal role in the successful completion of this study, and we are truly appreciative of their contributions.

VII. REFERENCES

[1] A. M. Remita and A. B. Diallo, “Statistical linear models in virus genomic alignment-free classification: application to hepatitis C viruses,” in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, November 2019.

[2] 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.

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

[4] 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.

[5] S. Aswath, CH.Mohan Sai Kumar, V.Hima Deepthi, S.Imran Javeed, and SVN. Rupesh. “DNA Sequence Classification with Improved Performance of Supervised Classifiers using Nature Inspired Algorithms”. In: 2022 2nd International Conference on Intelligent Technologies (CONIT). 2022.

[6] Yiren Wang, Vikram Khandelwal, Arindam K. Das, and M.P. Anantram. “Classification of DNA Sequences: Performance Evaluation of Multiple Machine Learning Methods”. In: 2022 IEEE 22nd International Conference on Nanotechnology (NANO). 2022.

[7] 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.

[8] Setiawan Mangkunegara and Purwono Purwono. “Analysis of DNA Sequence Classification Using SVM Model with Hyperparameter Tuning Grid Search CV”. In: 2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom). 2022.

[9] Yusuf Aleshinloye Abass, Steve A. Adeshina, Nwojo Nnana Agwu, and Moussa Mahamat Boukar. “Analysis of Prostate Cancer DNA Sequences Using Bi-direction Long Short Term Memory Model”. In: 2021 16th International Conference on Electronics Computer and Computation(ICECCO). 2021.

[10] M S Antony Vigil, Alan Christofer, Mithun Chandar, and Jayna Mukesh. “Comparative Analysis of Machine Learning Algorithms for DNA Sequencing”. In: 2023 Winter Summit on Smart Computing and Networks(WiSSCoN). 2023.

[11] Mohamed Kareem Al-Thiabi and Ali J Dawood Al-Alwani. “The Prediction of COVID-19 Virus Mutation Using Long Short-Term Memory”. In: 2022 8th International Conference on Contemporary Information Technology and Mathematics (ICCITM). 2022.

[12] L. William Mary and S. Albert Antony Raj. “Machine Learning Algorithms for Predicting SARS-CoV-2 (COVID19) – A Comparative Analysis”. In: 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC). 2021.

[13] 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.

[14] Satya Prakash, Anand Singh Jalal, and Pooja Pathak. “Forecasting COVID-19 Pandemic using Prophet, LSTM, hybrid GRU-LSTM, CNNLSTM, Bi-LSTM and Stacked-LSTM for India”. In: 2023 6th International Conference on Information Systems and Computer Networks (ISCON).2023.