Implementation of a novel technique for DNA sequence analysis using Machine Learning, Deep Learning and hybrid frameworks

Kshatrapal SINGH[1,2*], Raja Sarath Kumar BODDU[1,3]

[1] School of Computer Engineering, Lincoln University College, Petaling Jaya, Malaysia

pdf.kshatrapal@lincoln.edu.my, iamrajaboddu@gmail.com

[2] Department of Computer Science and Engineering, KCC Institute of Technology and Management, India

mekpsingh1@gmail.com (*Corresponding author)

[3] Department of AI&ML, Raghu Engineering College, Vishakhapatnam, India

iamrajaboddu@gmail.com

Abstract: Current research relies significantly on DNA sequencing. It enables the advancement of a wide range of disciplines, including genomics, metagenomics, and phylogeny. DNA sequencing involves extracting and interpreting DNA strands. In recent years, algorithms based on machine learning have been effectively employed to identify and classify viruses, which is crucial for preventing outbreaks such as COVID-19. Feature extraction aids in identifying the impact of viruses and designing drugs. The research presented here compares DNA sequencing employing Machine Learning algorithms, Deep Learning methods, and Hybrid algorithms. The objective of our suggested approach is to create an improved framework of prediction for DNA studies and achieve the most accurate findings possible. Machine learning, deep learning, and hybrid approaches are the most widely utilized and well-known techniques in this field. Additionally, deep learning provides higher predictive accuracy, and has demonstrated to perform better in several medical fields. The suggested frameworks are Decision Tree, Random Forest, Naive Bayes, Transform Learning, Convolutional Neural Network (CNN), Convolutional Neural Networks-Long Short-Term Memory (CNN-LSTM), as well as Bidirectional Long Short-Term Memory (Bidirectional LSTM). The CNN approach provided 98.4% accuracy with label encoding during training and 96.98% during testing, whereas the Bidirectional LSTM algorithm achieved 98.13% accuracy using K-mer encoding during testing and 97.7% during training. CNN-LSTM with K-mer encoding achieved 97.69% accuracy on collected data during testing and 96.23% during training.

Keywords: DNA Sequence, Machine Learning, Deep Learning, Decision tree, CNN, CNN-LSTM.

1. Introduction

Computational biology relies heavily on DNA sequence analysis. Once a patient becomes infected with a virus, swabs are taken from the patient, and the genes are classified. Viruses and bacteria can be identified by comparing the sequenced genomes with those stored in GenBank (NCBI). GenBank provides a BLAST server to determine the similarities between genomic sequences (Dharaniya et al., 2024). Assuming that DNA sequences expand rapidly, machine learning approaches are utilized for DNA map analysis. DNA is the blueprint for all organisms that exist. DNA consists of four nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). They constitute the basic components of DNA. Each virus's DNA is distinctive, and the sequence of the nucleotides determines the virus's distinct properties. DNA can be single- or double-stranded (see Figure 1). Within double-stranded DNA, every nucleotide forms a complementary pair with one on the other strand. Adenine and thymine form a pair, as do cytosine and guanine. In RNA, uracil (U) substitutes for thymine (T). As a result, the genome is the pattern of nucleotides (A, C, G, T) for DNA viruses and (A, C, U, G) for RNA viruses.

https://doi.org/10.33436/v36i2y202601

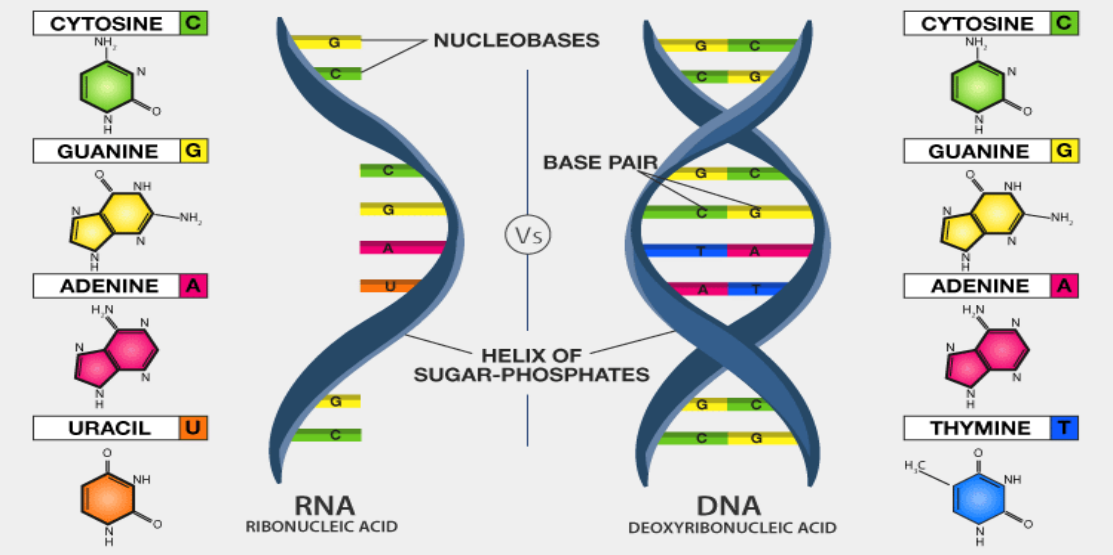

Figure 1. DNA and RNA structure (According to our own elaboration)

The DNA sequence is extremely lengthy, with maximum size of approximately 30,000 nucleotides, making it difficult to grasp and analyze. It must be transformed into a numerical representation before being processed by CNN models. Encoding strategy is also important for the accuracy of classification. This study uses two encoding approaches: label encoding identifies each nucleotide in a DNA input with a distinct index number while maintaining positional information, and second, K-mers are created from the DNA input, which are then transformed into English-like sentences. Any text system for classification may be applied to classify DNA. However, when the data becomes more complicated, manual feature extraction can lead to a variety of issues, such as choosing characteristics that do not result in the optimal solution or overlooking critical features (Sun et al., 2022). Automated feature extraction can be utilized to address this problem.

In particular, the DNA dataset's major properties are unclear. The retrieved characteristics of the actual DNA sequence are supplied in the LSTM as well as bidirectional LSTM frameworks for the purpose of groupings (Singh et al., 2025). The current study evaluated the accuracy as well as additional metrics of the CNN approach compared with hybrid approaches. The identical approaches are tested with label and K-mer encoding to see whichever encoding works most effectively with the DNA sequence (Babichev et al., 2024). The researchers suggested deep learning techniques such as CNN models to categorize DNA sequences. They also suggested a new approach to obtain features employing random DNA sequences centered on distance measures (Zhan & Moore, 2025).

The healthcare sector contains a large volume of information that can be utilized for predictive analysis. Machine learning, deep learning, and hybrid frameworks are employed to accelerate DNA sequencing and prediction, while there is a need to support medical practitioners with technological advances that enable them to input data into these frameworks and generate predictions. The method of operation will be determined by the type of data used, as well as the information contained in the columns or features used for machine learning architecture. The initial role of the mechanism is to support the practitioner in decision-making, which is time-consuming and cannot be relied upon for precision. Time is also an essential aspect of machine learning and deep learning systems used in DNA mapping. The goal is to improve the accuracy of predictions and provide insights for decision trees, random forest, Naïve Bayes, and CNN models. The results of machine learning algorithms are compared with those of deep learning algorithms, and the best-performing algorithms are identified.

2. Literature review

DNA sequencing enables the advancement of a wide range of disciplines, including genomics, metagenomics, and phylogenetics (Gunasekaran et al., 2021). The paper covers DNA mapping, maximum likelihood sequence identification, and bead decoding. In the present article, a sequencing-by-synthesis approach is discussed, as well as its non-idealities as a noisy switched linear device configured using deoxyribonucleic acid sequence (Veldhuis et al., 2022). The fundamental calling barrier then stands out as a boundary recognition challenge: given a deep view of the series and its corresponding output series, the objective is to estimate device bounds that minimize the probability of cryptographic errors. In this research, a framework is developed to obtain the essential local features of DNA cataphoretic periods. The resulting model is then subjected to the cataphoretic periods. The assessment's features are examined by redirecting each imitation and verified details.

Recent DNA sequencing analyses are heuristic in nature and make limited use of applied mathematical data (Alldio et al., 2022). This work proposes a true, irrefutable framework for DNA estimation that may be used to design maximum likelihood (ML) processor. Nonetheless, a review of restricted features of critical data may contribute to an exposition that is less accessible in text form (Ahmed et al., 2024). This research extends the concept of an efficient approach for obtaining DNA for analysis on heterogeneous systems, particularly considering the Intel Xeon Phi architecture. These platforms include one or more advanced host essential processing units and some Xeon Phi machines (Braisted et al., 2023). The article provides background information on the growing acceptance of heterogeneous computing platforms and regular expression comparison, as well as a description of our technique for investigating ways for faster DNA sequence evaluation. It also describes the experiment setup and addresses the practical assessment outcomes. The proposed work is evaluated and compared to similar contemporary artwork (Arruda et al., 2021). This article suggests that AI techniques promote a better solution for the prediction of treatment resistance in incurable diseases (Yoo et al., 2025). Presented result suggests that the AI methodology is a potential indication in predicting tuberculosis.

The use of machine learning to increase the fabrication speed of specialized sequencing boards is an emerging strategy for improving the detection of ctDNA mutations associated with this type of infection in patients. One system was built from an assortment of several existing boards, whereas the other maintained the presence of neoplastic alterations. Approaches, such as those used in this geographical unit, can distinguishing alterations in cfDNA extracted from heterologous disease samples (Deorowicz, 2020). Allele-specific expression (ASE) can be determined via RNA sequencing, which takes into consideration the varied expression values of each allele. A number of studies have demonstrated that ASE has a role in genetic diseases by controlling penetrance severity (Zeng et al., 2024).

The suggested method differs from a few known ML analyses. One of the major issues is identifying regular and inefficient characteristics that may be affected by specific diseases. In genomic studies, we can become accustomed to diverse protein components by categorizing segments of DNA into established classifications. In this method, certain features can differentiated and classified. To discriminate between infected and ordinary features using characterization methodologies, we will use artificial intelligence (AI) techniques. This study provides a review on the tools that provide superior grouping structure using Machine Learning approaches, including brief descriptions for genomics, writing outline, and primary areas of concern in DNA Sequencing employing Machine Learning. As a result, DNA sequencing has become an essential tool.

They demonstrated that two-fold abandoned DNA atoms may be correctly organized using AI approaches for assessing quantum transport parameters (Hadikurniawati et al., 2021).

In this research, the goal of our suggested system is to create an improved predictive model for DNA studies and achieve the most accurate results possible. Machine learning, deep learning, and hybrid models are the most widely utilized and well-established approaches. A greater accuracy for predictions in deep learning is employed as well, which has demonstrated better performance in several medical fields. The suggested frameworks are Decision Tree, Random Forest, Naive Bayes, Transform Learning, CNN, CNN-LSTM, and Bidirectional LSTM.

3. Proposed methodology

The suggested design appears to be straightforward in terms of execution; however, analyzing the dataset is complicated. The dataset consists of the DNA sequence, which difficult to interpret. There are quite a few undesirable symbols that require pre-processing. Figure 2 will provide an architectural illustration of the suggested methodology:

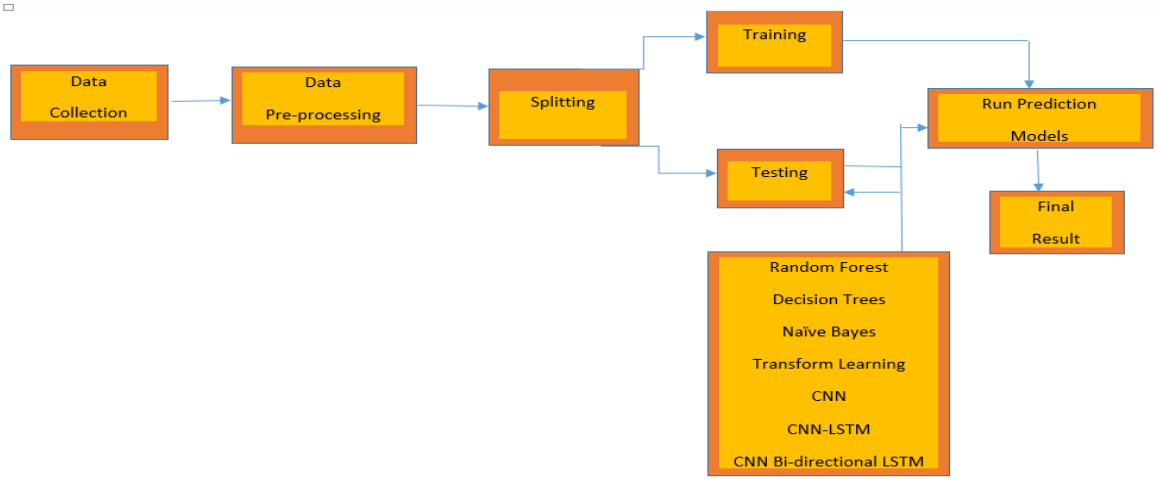

Figure 2. Workflow of the proposed methodology (According to our own elaboration)

3.1. Data collection

The entire DNA data for viruses as well as influenza viruses was retrieved from the world's nucleotide sequence database, The National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov). The DNA data content is stored in the FASTA file structure.

3.2. Data pre-processing

Pre-processing data is the primary important stage for various machine learning and deep learning approaches that deal with numerical information. The genome sequence of the DNA dataset is to be grouped. There are numerous methods for converting categorical data into numerical data. The encoding approach converts categorical nucleotide data into numerical format. This research uses both label encoding and K-mer encoding to encode DNA sequences. This analysis examines how the encoding approach affects classification precision. Label encoding assigns an index value to every nucleotide in the DNA sequence, such as A-1, C-2, G-3, and T-4 (see Figure 3). LabelBinarizer converts the whole DNA sequence to a sequence of integers.

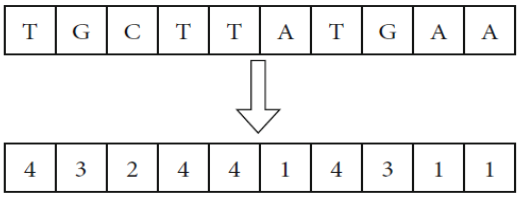

Figure **3.** Sequence encoding with label binarizer (According to our own elaboration)

Figure 4 shows how each sequence of DNA is transformed into an m-sized K-mer representation, while all the K-mers are linked together to produce a sentence. DNA sequences are currently being grouped using natural language processing (NLP) techniques. In this work, the word embedding layer transforms the K-mer phrase into an intense feature vector matrix.

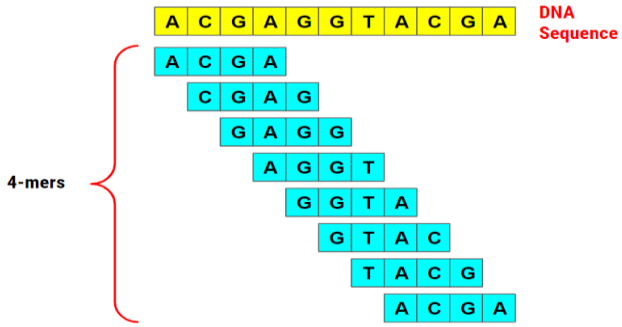

Figure **4.** Sequence encoding with k-mer approach. (According to our own elaboration)

3.3. Data splitting

The data input being used is categorised into training and testing sets. The combination typically follows three distinct configurations. Specialists typically utilize a mixture of 70%–30%, 75%–25%, and 80%–20% splits. The training set takes up the majority of the value, with a tiny piece reserved for testing.

3.4. Implementation of prediction models

We implement multiple algorithms that are often known as models: Random Forest, Decision Trees, Naïve Bayes, Transform Learning, Convolution Neural Network (CNN), CNN-LSTM, and CNN Bi-directional LSTM.

3.5. Algorithms applied in the suggested system

The following section describes the applied algorithms in the suggested system to better understand how the models operate on the input data and the goal of the issue statement solution technique.

3.5.1. Decision tree execution

In this article, we will look at how to build DNA mapping with a decision tree method. Under the decision tree technique, we will use the Gini approach to assess the likelihood of pairing the characteristics based on the approach being used. In this approach, the Gini technique is used to determine purity and impurity. Impurity is used for detecting the loss that occurs in the input data. The process of execution will involve determining the accuracy of the pair of variables that are independent, and attempting to minimize the Gini approach.

The decision tree classification issue is solved in a top-down manner. It has multiple vertices. The initial vertex is called the root vertex, the middle vertices that store the data are known as internal vertices, and the final vertices are known as leaf vertices. Each of these is used for testing, and any fresh data will be routed to all nodes depending on a Yes/No criterion. Accuracy and recall are estimated after the decision has been made, based on the number of nodes traversed and the frequency of Yes/No decisions. Using the confusion matrix, we can obtain the four variable outcomes.

3.5.2. Random forest execution

The random forest is a combination of decision tree classifiers. This implies that various random subsets of independent variables that are independent as well as categorize them separately. The internal model accuracy of each decision tree is evaluated once again using the Gini technique. To use a random forest for DNA sequencing, the following steps are performed:

Step 1: Determine the number of K data points to create the set for training.
Step 2: Create a decision tree using the chosen subsets.
Step 3: Select the decision trees to compute N then construct the tree.
Step 4: Repeat steps 1 and 2.
Step 5: Allocate the new data point to its appropriate node or category based on the forecast result of the N decision trees.

3.5.3. Naïve Bayes execution

The method being used will deal with frequency probability. It depends on Bayes' theorem. We need to determine why an activity occurs. The theorem allows us to produce probability pairs and predict the possibility of an event. Prediction with a Naïve Bayes (NB) classifier is easier than with other methods of prediction; however, a major drawback is time commitment.

A: Select the features needed for prediction.

B: Classify the variables that are expected to be estimated numerically.

C: Determine the likelihood that each row's event will occur.

D: The data set provided is divided into training and testing sets. After training, test findings using the test set parameters.

E: Determine the frequency probability (positive or negative).

F: Repeat steps A to E.

The categories of variables will perform according to the requirements, with the most significant aspect being the qualities that are assigned to a single category.

3.5.4. Transform learning execution

Through the deep transform learning approach, we achieve deep insights using a previously established neural network. Transform learning for DNA sequencing involves shifting from raw, unaligned nucleotides (A, C, T, G) to context-aware numerical representations using deep learning. By applying Natural Language Processing (NLP) and Transformer architectures, researchers convert genetic code into mathematical vectors to accurately predict mutations, regulatory elements, and disease indicators. Transform Learning is a process of combining the findings of several neural network events and using them to construct a novel unsupervised learning technique.

A: Create an embedded layer with K distinct variables and subsets.

B: Use CNN to train various subsets. Create distinct CNNs for each subset.

C: Produce individual outcomes per group.

D: Obtain the updated data input.

E: Repeat steps 1 through 3.

This will be a mixture of various neural networks, with the loss function applied to each CNN output to create a fresh deep neural network. The matrix built at each stage will accurately evaluate the data and allocate it to the appropriate classification. For example, if a previously trained dataset or framework does not have a certain amount of labels, we may adjust the number of labels. In fact, each dataset is unique, such as human, dog, or chimpanzee datasets. Regardless of resource size, we will not have identical gene structures. When we train a dataset with our model, we modify the most recent layer; such is referred to as transform learning. Transform learning allows for efficient training of datasets, improved categorization, and adaptability to a wide range of scenarios.

3.5.5. CNN

The DNA sequence is encrypted using K-mer and label encoding approaches, which maintain the positional information of every nucleotide in the pattern. The data from the two methods mentioned above is processed using the embedding layers. The CNN layer serves as a foundation for feature extraction and provides the input for both bidirectional LSTM and LSTM classifications. Figure 5 shows the process for the proposed application. CNN is a popular deep learning method that can produce state-of-the-art results for the majority of classification tasks. CNN may yield strong accuracy on text data in addition to performing well on image classification. CNN is primarily employed for extracting features from the input dataset. The DNA input is transformed into numerical information using either label encoding or one-hot encoding, as CNN can only process numerical data. Feature extraction in the convolutional layer is significantly influenced by the kernel size. The number of filters and kernel size are the system's hyper-parameters.

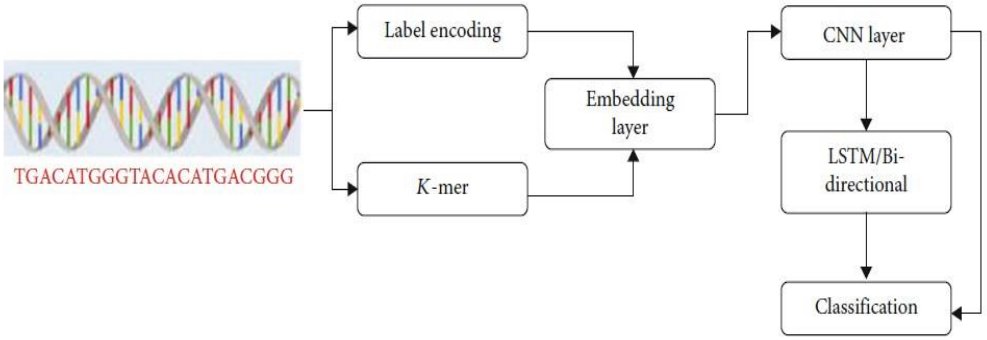

Figure 5. Process for the suggested application (According to our own elaboration)

3.5.6. CNN-LSTM

Sequence predictions and classifications use Long Short Term Memory (LSTM), a recurrent neural network (RNN) that can recognize long-term dependencies in input data. It is composed of a sequence of memory blocks called cells. The LSTM can selectively retain or forget information. The general design of the LSTM framework is shown in Figure 6. It recognizes long sequences and has the ability to learn. Two inputs are required for this gate: X_t (input from the current state) and $h_{(t-1)}$ (input from the previous state). The input sequence is combined with bias after being multiplied by a weight. The sigmoid function is then applied. The value that the sigmoid function produces ranges from 0 to 1. The LSTM's input gate is responsible for adding all pertinent values to the cell state. There are two activation functions involved: the sigmoid function, which governs

which values contribute to the cell state, and the tanh function, which produces values between -1 and +1, representing every value that could be assigned to the cell state. In order to predict the categorizing labels in our framework, convolutional layers are followed by an LSTM layer. The LSTM layer receives the features extracted by the convolutional layer as an input for classification.

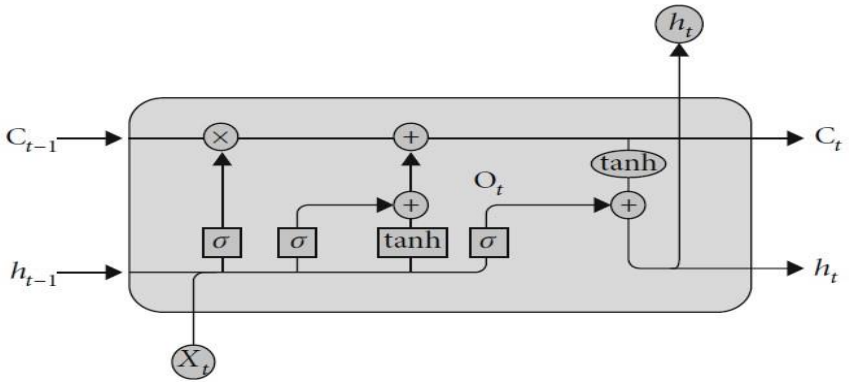

Figure 6. Framework of LSTM model (According to our own elaboration)

3.5.7. CNN Bidirectional-LSTM

For classifying DNA sequences, a hybrid bidirectional LSTM and CNN framework is employed. The framework uses bidirectional LSTM for classification and CNN for feature extraction. The bidirectional LSTM consists of two RNNs: one for learning forward sequence dependencies and another for learning backward sequence requirements. Figure 7 shows the construction of the bidirectional LSTM.

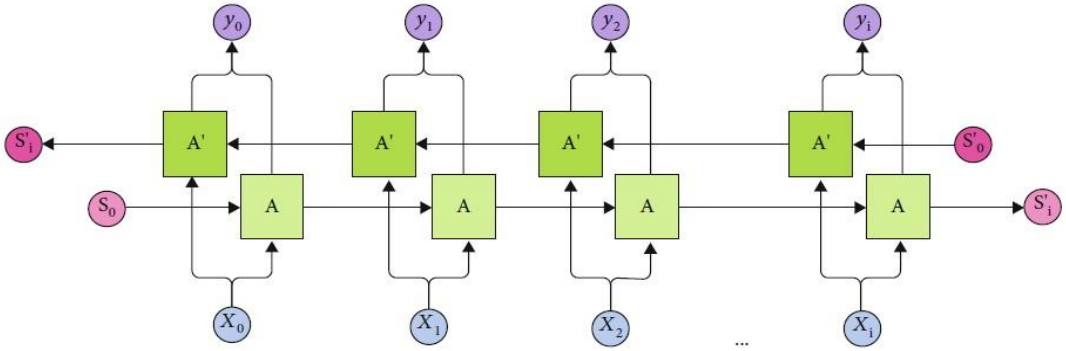

Figure 7. Framework of Bidirectional LSTM model (According to our own elaboration)

4. Results and discussion

The Tesla P100 GPU processor with 16 GB of RAM is used to test the proposed models. The dataset consists of 46,281 inputs and is split into training, validation, and testing sets at 70%, 10%, and 20%, respectively. There are 32,397 samples in the training set, 4,628 samples in the validation set, and 9,256 samples in the testing set. The longest sequence is 1,800. The error between the target labels, used for weight training and updating and the actual result is computed using a loss function. By altering the amounts of several hyper-parameters, such as filters, filter size, the quantity of layers, and embedding dimension, we evaluate CNN, CNN-LSTM, and CNN-bidirectional LSTM models. The most popular technique for parameter optimization to select the best model configuration is grid search cross-validation. The optimal filter counts are 128, 64, and 32 in each layer for all three models. The filters have dimensions of 1 × 1 (1D), 100 training batches, 10 training epochs, 32 embedding dimensions, and 4 K-mer sizes. Several categorization metrics, including accuracy, precision, recall, F1-score, sensitivity, and specificity, are used to assess the models. True PositiveGen (TPGen), True NegativeGen (TNGen), False PositiveGen

(FPGen), and False NegativeGen (FNGen) values are obtained from the confusion matrix and used to compute the aforementioned metrics.

The following equations for the various metrics are based on the categorized states mentioned above:

$$\text{Accuracy} = (\text{TPGen} + \text{TNGen}) / (\text{TPGen} + \text{TNGen} + \text{FPGen} + \text{FNGen}) \quad (1)$$

$$\text{Specificity} = \text{TNGen} / (\text{TNGen} + \text{FPGen}) \quad (2)$$

$$\text{Sensitivity} = \text{TPGen} / (\text{TPGen} + \text{FNGen}) \quad (3)$$

$$\text{Precision} = \text{TPGen} / (\text{TPGen} + \text{FPGen}) \quad (4)$$

Sensitivity is the proposal of a sequence that the model defines as positive within actual positive sequences. It is expected that some positive instances will be false negatives due to increased sensitivity. The average of recall and precision is known as the F1-score. Precision is the proportion of items that the model correctly classifies as positive. The model's ability to identify the negative cases is known as specificity. Table 1(a) shows the accuracy of Machine learning models. The accuracy of CNN, CNN-LSTM, and CNN-Bidirectional LSTM models is displayed in Table 1(b). CNN provides greater accuracy when label encoding is used to encode the DNA sequences. When DNA sequences are encoded via K-mers, LSTM and bidirectional LSTM models simultaneously offer greater accuracy. We observed that the label encoding technique's testing accuracy is consistently lower than its training accuracy. In contrast, testing accuracy is more important than training accuracy in the context of K-mer encoding. The findings demonstrate that classification accuracy is significantly impacted by the implementation of the encoding method.

Table 1(a). Accuracy of Machine learning models

Sr. No.	Models	Accuracy
1.	Random forest	95.01%
2.	Decision Trees	96.11%
3.	Naïve Bayes	95.88%
4.	Transform Learning	94.37%

Table 1(b). Accuracy of classification models

	CNN	CNN-LSTM	CNN-Bidirectional LSTM
Label encoding training	98.4%	95.17%	97.19%
Label encoding testing	96.98%	95.01%	96.33%
K-mer encoding training	95.02%	96.23%	97.7%
K-mer encoding testing	95.8%	97.69%	98.13%

To demonstrate the model's resilience, the proposed approach is contrasted with several approaches, including CNN, CNN-LSTM, and CNN-Bidirectional LSTM. The accuracy of CNN, CNN-LSTM, and CNN-bidirectional LSTM is 98.4%, 97.69%, and 98.13%, respectively. The models from the literature considered for the outcome’s analysis are capable of classifying up to three different groups. The model proposed by (Ghosh et al., 2025) achieves an accuracy of 97.1% for binary classification. In (Zhang et al., 2020), the authors presented a model using the XGboost algorithm to categorize five different types of chromosomes with an accuracy of 88.82%. CNN was used in (Hu et al., 2024) for classification and achieved an accuracy of 95%. Table 2 presents the comparison analysis. The experimental observations presented above demonstrate that the proposed approach performs well for classifying DNA sequences. The 1D-CNN is an effective tool for classifying DNA sequences and extracting useful patterns from textual data. With 98.4% classification accuracy, the CNN framework effectively extracts features, which is helpful for the classification approach. The bidirectional CNN LSTM offers the second-highest accuracy of 98.13%, which offers the benefits of maintaining long-term relationships in contrast to the CNN frameworks by employing the K-mer to encode the sequence method.

Table 2. Comparison of suggested approach with different research techniques

Sr. No.	Approach applied	Accuracy
1.	CNN	98.4%
2.	LSTM	97.69%
3.	Bidirectional LSTM	98.13%
4.	Ghosh et al., 2025	97.1%
5.	Hu et al., 2024	95%
6.	Zhang et al., 2020	88.82%

Additionally, Table 3 lists the studies conducted for various parameters. A specificity of 99.1% was achieved by the CNN Bidirectional LSTM model using K-mer encoding. A sensitivity of 99.2% was achieved by the CNN model using K-mer encoding, while a precision of 99.12% was achieved by the CNN model using label encoding. The F1-score monitored was 97.7% for the CNN Bidirectional LSTM model using K-mer encoding.

Table 3. Performance evaluation of various models with Label and K-mer encoding

Parameters / Suggested approaches	CNN		CNN-LSTM		CNN Bidirectional LSTM	
	Label encoding	K-mer encoding	Label encoding	K-mer encoding	Label encoding	K-mer encoding
Accuracy	.9840	.9580	.9517	.9769	.9719	.9813
Specificity	.9781	.9823	.9880	.9860	.9812	.9910
Sensitivity	.9852	.9920	.9850	.9810	.9690	.9720
Precision	.9912	.9760	.9890	.9901	.9860	.9802
F1	.9701	.9780	.9690	.9711	.9689	.9770

5. Conclusion

The field of computational biology relies significantly on DNA sequence mapping. The present study relies largely on DNA sequencing. Numerous fields, such as phylogeny, genomics, metagenomics, can advance thanks to it. Machine learning-based algorithms have been successfully used to detect and categorize viruses, which is essential for preventing outbreaks such as COVID-19. In the present research, Random forest, Decision trees, Naïve Bayes and Transform learning were implemented for DNA sequencing. In addition, label encoding and K-mer encoding were employed to compare three deep learning techniques: CNN, CNN-LSTM, and CNN-bidirectional LSTM. Although the testing accuracies are relatively modest, we discovered that CNN with label encoding performs better than the other frameworks. Accuracy measures alone cannot be used to evaluate this dataset. It is also necessary to consider other metrics, such as specificity, sensitivity, recall, and precision. CNN with label encoding delivers a greater precision rate for classes with large sample sizes, while CNN with K-mer encoding offers a greater precision rate for classes with smaller sample sizes. Regardless of class labels, all models using K-mer encoding achieve good recall values. K-mer encoding is therefore suitable for classification models where a better recall rate is required. Machine learning (ML) has significantly accelerated DNA sequencing and interpretation, streamlining genome assembly, variant calling, and gene annotation. However, the approach faces several critical limitations, primarily related to data quality, algorithmic bias, and the complex, heterogeneous nature of biological data. Because of the increasing volume of heterogeneous data, raw datasets often contain missing, inconsistent, or noisy information. Low data quality can have a serious negative impact on the accuracy of information extraction.

Data Availability

The data used to support the findings of this research are available at “The National Center for Biotechnology Information (NCBI)” (https://www.ncbi.nlm.nih.gov).

Author contributions

Conceptualization: K. S. and R. B.; Data curation: K. S.; Project administration: K. S. and R. B.; Supervision: R. B.; Validation: K. S. and R. B.; Writing - original draft: K. S.; Writing - review and editing: K. S. and R. B.; All authors have read and agreed to the published version of the manuscript.

Submission received: 01 May 2026; Revised: 16 June 2026; Accepted: 22 June 2026; Published: 30 June 2026.

REFERENCES

Ahmed, N.Y., Alsanousi, W.A., Hamid, E.M. et al. (2024) An Efficient Deep Learning Approach for DNA-Binding Proteins Classification from Primary Sequences. *International Journal of Computational Intelligence Systems*. 17, 88. https://doi.org/10.1007/s44196-024-00462-3.

Alladio, E., Poggiali, B., Cosenza, G., Pilli, E. (2022) Multivariate statistical approach and machine learning for the evaluation of biogeographical ancestry inference in the forensic field. *Sci. Rep.* 12(1), 8974.

Arruda, M. M., De Assis, F. M. & De Souza, T. A (2021) Is BCH code useful to DNA classification as an alignment-free method? *IEEE Access*. 9, 68552–68560.

Babichev, S., Liakh, I. & Kalinina, I. (2024) Applying Deep Learning Techniques to Solve Classification Tasks Using Gene Expression Data. *IEEE Access.* 12, 28437–28448. https://doi.org/10.1109/ACCESS.2024.3368070.

Braisted, J., Patt, A., Tindall, C., Sheils, T., Neyra, J., Spencer, K., Eicher, T., Mathé, E.A. (2023) RaMP-DB 2.0 - A renovated knowledgebase for deriving biological and chemical insight from metabolites, proteins, and genes. *Bioinformatics.* 39, btac726.

Deorowicz, S. (2020) FQSqueezer: k-mer-based compression of sequencing data. *Scientific Reports*. 10(1), 578-579.

Dharaniya, N.G., Raaj, R.K., Vikramathithan, M., Vishal, P., Yugavanan, S. (2024) DNA sequencing using a machine learning algorithm. *Int. J. Res. Publ. Rev.* 5, 12272–12274.

Ghosh, P., Das, S., Pal, D., Saha, S., Dasgupta, S. (2025) Efficient DNA sequence classification through Machine Learning techniques. *International Research Journal of Multidisciplinary Scope (IRJMS).* 6(3), 1180-1189.

Gunasekaran, H., Ramalakshmi, K., Rex Macedo Arokiaraj, A., Deepa Kanmani, S., Venkatesan, C. & Suresh Gnana Dhas, C. (2021) Analysis of DNA Sequence Classification Using CNN and Hybrid Models. *Computational and Mathematical Methods in Medicine*.

Hadikurniawati, W., Anwar, M. T., Marlina, D. & Kusumo, H. (2021) Predicting tuberculosis drug resistance using machine learning based on DNA sequencing data. *Journal of Physics: Conference Series*. 1869, 1, p. 012093.

Hu, W., Li, Y., Wu, Y., Guan, L., Li, M. (2024) A Deep Learning Model for DNA Enhancer Prediction based on Nucleotide Position Aware Feature Encoding. *iScience.* 27, 110030.

Singh, K., Singh, L., Shukla, V., Sharma, Y. K., Rai, A. K. (2025) An Advanced Approach for DNA Sequencing and Similarities Analysis on the Basis of Groupings of Nucleotide Bases. *International Journal of Data Mining and Bioinformatics*. 29(1/2), 133-149.

Sun, K., Yao, Y., Yun, L., Zhang, C., Xie, J., Qian, X., Tang, Q., Sun, L. (2022) Application of machine learning for ancestry inference using multi-InDel markers, Forensic *Sci. Int Genet*. 59, 102702.

Veldhuis, M.S., Ariens, S., Ypma, R.J.F., Abeel, T., Benschop, C.C.G. (2022) Explainable artificial intelligence in forensics: Realistic explanations for number of contributor predictions of DNA profiles. *Forensic Sci. Int Genet.* 56, 102632.

Yoo, H., Shin, H., Xu, K., Rosen, G. (2025) Exploring Adversarial Robustness in Classification Tasks Using DNA Language Models. *arXiv*, arXiv:2409.19788.

Zeng, J., Gao, X., Gao, L., Yu, Y., Shen, L., Pan, X. (2024) Recognition of rare antinuclear antibody patterns based on a novel attention-based enhancement framework. *Brief. Bioinform.* 25, bbad531.

Zhan, H., Moore, J.H. (2025) SafeGenes: Evaluating the Adversarial Robustness of Genomic Foundation Models. *arXiv*, arXiv:2506.00821.

Zhang, X., Beinke, B., Al Kindhi, B. & Wiering, M. (2020) Comparing machine learning algorithms with or without feature extraction for DNA classification. *arXiv:2011.00485.* https://doi.org/10.48550/arXiv.2011.00485.

Kshatrapal SINGH is currently working as Dean and Professor in the Department of Computer Science and Engineering (CSE) at KCC Institute of Technology and Management, Greater Noida. Dr. Singh received his Bachelor of Technology (B.Tech.) in Computer Science and Engineering from Uttar Pradesh Technical University (UPTU), Lucknow, India in 2004. He obtained his M.Tech. (in Computer Engineering) from Maharshi Dayanand University (M.D.U.), Rohtak, India, in 2010. He was awarded a Ph.D. in Computer Science and Engineering from Dr. A.P.J. Abdul Kalam Technical University (AKTU), Lucknow, India. Dr. Singh is currently pursuing a postdoctoral fellowship at Lincoln University College, Malaysia. He has been associated with various organisations, namely Greater Noida Institute of Technology, KIET Group of Institutions, Ghaziabad, Krishna Engineering College, Ghaziabad, I.T.S Engineering College, Greater Noida, Lingaya's University, Faridabad, Somany Group of Institutions, Rewari, and many others. He has more than 22 years of academic experience and 8 years of research experience. His research interests include data analytics, computational biology, bioinformatics, graph theory in computer science and distributed systems. He has published more than 50 research papers, patents, and book chapters in reputed journals and publications of various publishers.

Raja Sarath Kumar BODDU is currently working as a Professor and Head of the Department in the Faculty of Artificial Intelligence and Machine Learning at Raghu Engineering College, Visakhapatnam. He has 25 years of experience in engineering education, and his services have been ratified in the cadre of 'Principal' by JNT University, Kakinada. Dr. Boddu has edited two books, authored five books and 23 book chapters in Computer Science, and published 98 research papers. He holds eight patents in the field of Computer Science. Prof. Boddu completed a Postgraduate Certificate in Business Administration from the Indian Institute of Management, Visakhapatnam, and obtained his Ph.D., Master's degree and Bachelor's degree from Andhra University, Visakhapatnam. He completed a postdoctoral fellowship at the University of South Florida, USA. He has supervised 17 postgraduate dissertations and 10 undergraduate dissertation groups to date. Currently, he is supervising four doctoral students from various universities in India and five postdoctoral students from Lincoln University, Malaysia.