

Classification and Prediction of Multi-Class DNA Sequences Utilizing Artificial Intelligence Technology

Minseong Hong, *L&N STEM Academy*, gomshong@gmail.com

Seongyong Hong, *Carson-Newman University*, shong@cn.edu

Abstract

Multi-class DNA sequence classification is of critical importance in bioinformatics, as it facilitates the accurate identification of species, gene functions, and disease-related variations. Conventional laboratories and alignment-based methodologies frequently require extended periods and are computationally onerous when applied to large-scale genomic datasets. This study proposes a set of propositions regarding the employment of artificial intelligence techniques for the classification and prediction of multi-class DNA sequences, with a view to achieving both efficiency and accuracy. This research includes an experimental component in which the techniques are applied to a dataset consisting of 10 classes of DNA. The DNA sequences obtained from synthetically random genomic files were subjected to preprocessing and transformed into numerical representations. This was achieved by employing feature engineering techniques, such as k-mer frequency extraction and one-hot encoding. Therefore, in this study, classification models, including the Random Forest algorithm, Support Vector Machine (SVM), and Convolutional Neural Network (CNN), were implemented and evaluated. Furthermore, this study proposed and experimented with CNN and SVM algorithms as hybrid models. The performance of each model was evaluated using standard evaluation metrics, including accuracy, precision, recall, and F1-score. The findings emphasize the efficacy of AI-driven methodologies in genomic data analysis, offering a scalable technique for the classification and prediction of multi-class DNA. This study contributes to advancements in bioinformatics, supporting applications in disease diagnosis, evolutionary studies, and personalized medicine.

Keywords: DNA Sequence Classification, Genomic Data Prediction, Artificial Intelligence in Bioinformatics, Support Vector Machine (SVM), Convolutional Neural Network (CNN)

Introduction

The fast improvement of next-generation sequencing technologies has led to a huge amount of new information about genomes. This has completely changed the way we do biological research and clinical medicine. DNA sequences are made up of four different bases: adenine (A), thymine (T), cytosine (C), and guanine (G). These bases encode the biological instructions that govern all living organisms. One of the most important challenges in modern bioinformatics is figuring out what these genetic sequences are. This is important because it helps scientists understand what species they are, what genes they are, or if they are related to diseases. Genomic databases are growing very quickly. This means that there is a great need for classification systems that can adapt to this growth. These systems should be able to scale up automatically and be intelligent. Conventional methodologies employed in DNA sequence analysis have historically relied on alignment-based computational methods, including pairwise sequence alignment, BLAST searches, and motif-detection algorithms (Satam et al., 2023). While these methods have contributed

substantially to foundational genomic discoveries, they are increasingly inadequate for large-scale classification tasks. As illustrated in Figure 1, four interconnected components of DNA biology are fundamental to sequence classification research. The DNA Structure and Composition offer a comprehensive illustration of the double helix architecture, in which nucleotides consisting of sugar-phosphate backbones and nitrogenous base pairs (A~T and C~G) form the foundational molecular structure. The organization of genes into functional sequences is a fundamental aspect of biology, elucidating the mechanisms that govern cellular development, growth, and reproduction. Replication and Sequence Conservation demonstrate the process by which the original DNA helix duplicates into identical daughter helices during cell division, thereby ensuring the preservation of genetic information across generations. Sequence Motifs and Their Identification discusses computational methods, including ChIP-seq, DNase Footprinting, and SELEX, which are used to detect recurring nucleotide patterns (Furey, 2012; Gusmao et al., 2016; Dao et al., 2016). These patterns are represented visually through Position Frequency Matrices and sequence logos. These patterns directly inform AI-based DNA classification models.

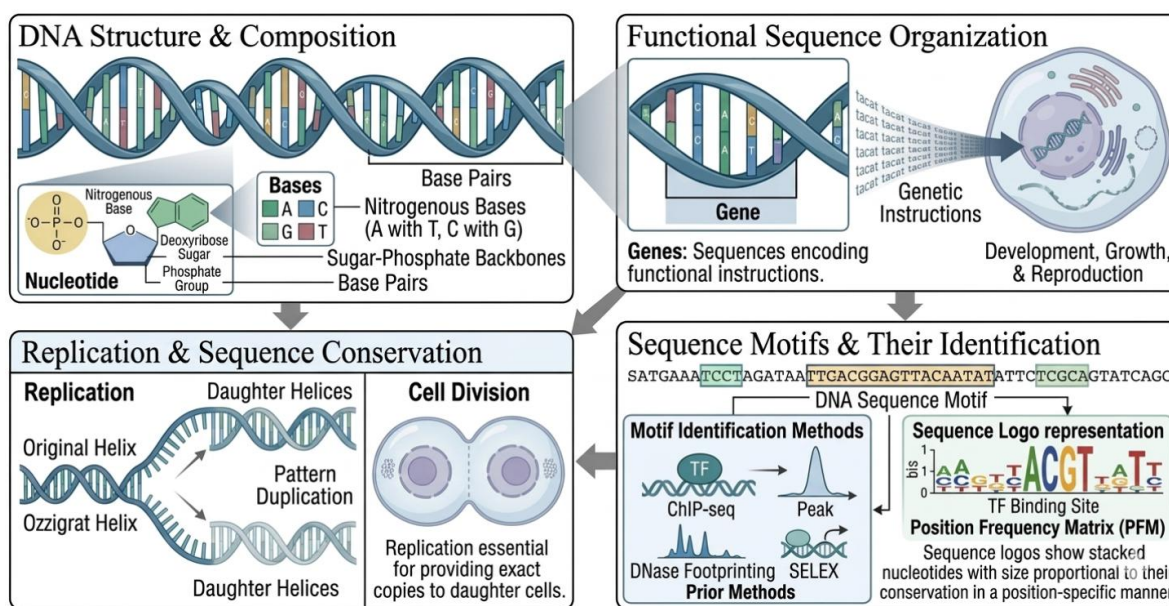


Figure 1. Fundamentals of DNA Architecture, Genetic Instruction Flow, and Sequence Motif Recognition

Stormo (2000) demonstrated that early sequence representation methods, although effective for small datasets, become computationally prohibitive as database sizes grow. Concurrently, Badger et al. (2012) evaluated conventional sequence classifiers and concluded that their performance deteriorated considerably with increasing sequence volume and diversity. This finding emphasizes the pressing need for more efficient, learning-based alternatives that can handle the complexity and scale of contemporary genomic repositories. The advent of machine learning and deep learning has engendered a transformative paradigm in DNA sequence analysis. Contrary to the conventional approach of utilizing manually engineered alignment rules, these algorithms leverage discriminative patterns directly from the labeled training data. Yang et al. (2020) conducted a comprehensive review of machine learning applications in DNA sequence mining and confirmed that ensemble methods such as Random Forest, kernel-based approaches such as Support Vector Machines, and neural network architectures consistently outperform classical alignment methods across a wide range of genomic classification tasks. Notably, their analysis underscored the efficacy of k-mer frequency encoding as a strategy for converting symbolic nucleotide sequences into structured numerical feature spaces conducive to machine learning algorithms (Campbell & Hong, 2024). This finding directly motivates the features of engineering methodology adopted in this study. Deep learning is a type of machine learning that can handle complicated relationships in biological data. LeCun, Bengio, and Hinton (2015) discovered that deep neural networks are very good at learning complex,

nonlinear representations from high-dimensional data. This property is especially valuable in genomic contexts, where subtle nucleotide patterns have significant biological meaning. Zou et al. (2019) built on this research. They provided a thorough introduction to the use of deep learning in genomics. They showed that Convolutional Neural Networks are especially good at finding local sequence patterns and functional elements in raw DNA data. These advances suggest that combining classical machine learning and deep learning models in a unified framework is a rigorous and comprehensive approach for classifying DNA sequences. Despite these advances, there remains a significant research gap. This gap is in comparing traditional machine learning models and deep learning architectures. These models and architectures are used to classify DNA problems into more than one class. This comparison should be performed using standard datasets and evaluation protocols. Many existing studies have focused on either simple classification tasks or highly specialized biomedical datasets. This limits how often they can be reproduced and how widely they can be used. This study addresses the identified research gap by implementing and rigorously evaluating four artificial intelligence classification models: Random Forest (RF), Support Vector Machine (SVM), Convolutional Neural Network (CNN), and a Hybrid CNN+SVM architecture applied to a ten-class DNA sequence dataset. Each model was systematically assessed using a comprehensive suite of performance metrics, including accuracy, precision, recall, and F1-score. The primary objective was to determine which AI technique delivers the highest classification accuracy while maintaining computational efficiency and practical reproducibility. Beyond classification performance, this research aims to contribute a scalable and methodologically transparent AI technology that researchers and practitioners can readily replicate and extend to broader genomic applications, including genetic disease identification, pathogen detection, and personalized medicine.

Relative Work

The research conducted in the fields of DNA and bioinformatics has enabled the deciphering of the fundamental blueprints of life, thereby unveiling the genetic underpinnings of traits and diseases. This work is critical because it powers the development of personalized medicine and targeted therapies, moving healthcare from a one-size-fits-all approach to a precision treatment. The application of artificial intelligence to DNA sequence classification has evolved rapidly over the past decade, progressing from traditional alignment-based computational methods to sophisticated deep learning and transformer-based genomic architectures. This section reviews the most relevant literature organized thematically across machine learning foundations, deep learning advances, hybrid architecture, transformer-based models, and emerging challenges in AI-driven genomic analysis. Classical machine learning algorithms establish a foundational framework for computational DNA sequence classification. As illustrated in Figure 2, the adoption of Support Vector Machine models has been extensive because of their ability to construct optimal decision boundaries in high-dimensional genomic feature spaces.

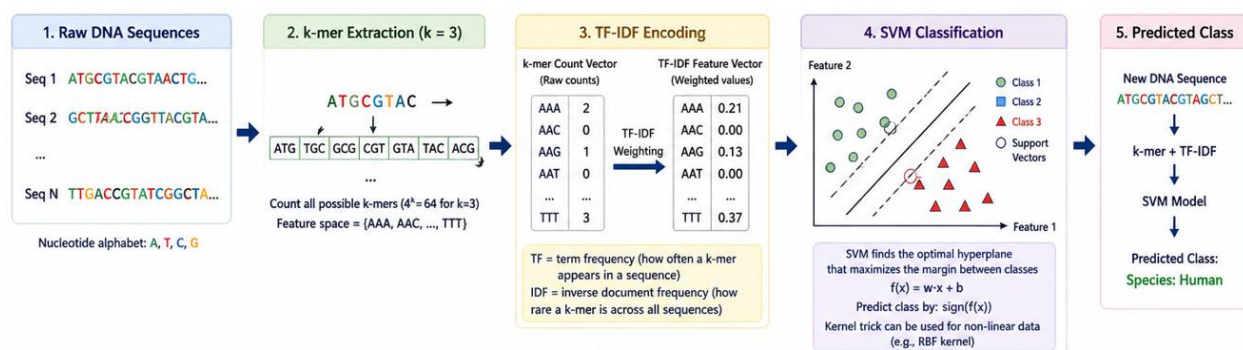


Figure 2. DNA Sequence classification technique based on SVM

Researchers commonly integrate SVM with k-mer extraction and TF-IDF encoding strategies to transform raw nucleotide sequences into structured numerical vectors, demonstrating strong classification performance across species identification, promoter prediction, and mutation classification tasks (Asgari and Mofrad, 2015; Kimothi et al., 2016; Yang et al., 2020). Breiman (2001) introduced Random Forest as an ensemble learning method that builds multiple decision trees and combines their predictions through majority voting, a design that confers robustness against overfitting and stable performance on high-dimensional genomic data. Yang et al. (2020) further confirmed that both SVM and Random Forest consistently outperform alignment-based approaches when applied to k-mer encoded DNA datasets, establishing these algorithms as reliable baseline models for genomic classification benchmarking. Pedregosa et al. (2011) developed Scikit-learn, a Python library that has since become the standard implementation environment for these classical machine learning approaches in bioinformatics research. Deep learning architecture has become dominant in genomic classification research because it eliminates the need for manual feature engineering by learning biologically meaningful representations directly from raw nucleotide sequences. LeCun, Bengio, and Hinton (2015) established a theoretical foundation for deep neural networks as universal function approximators capable of learning hierarchical nonlinear patterns from large datasets, a property particularly valuable in genomic contexts where subtle sequence motifs carry significant biological meaning. As illustrated in Figure 3, Convolutional Neural Networks (CNN) have proven especially effective for DNA classification because their local receptive fields detect nucleotide motifs while preserving sequence positional order. Kelley, Snoek, and Rinn (2016) demonstrated this capability through Basset, a CNN architecture that successfully learned regulatory sequence patterns from genomic accessibility data. Recently, a study published in 2024 demonstrated that image-based genomic representations, specifically Chaos Game Representation combined with CNN architectures, can effectively capture spatial and structural genomic relationships that conventional sequence-encoding methods may overlook (Alipour et al., 2024). Zou et al. (2019) provided a comprehensive review confirming that CNNs achieve state-of-the-art performance across multiple genomic tasks including sequence classification, variant calling, and functional genomic annotation.

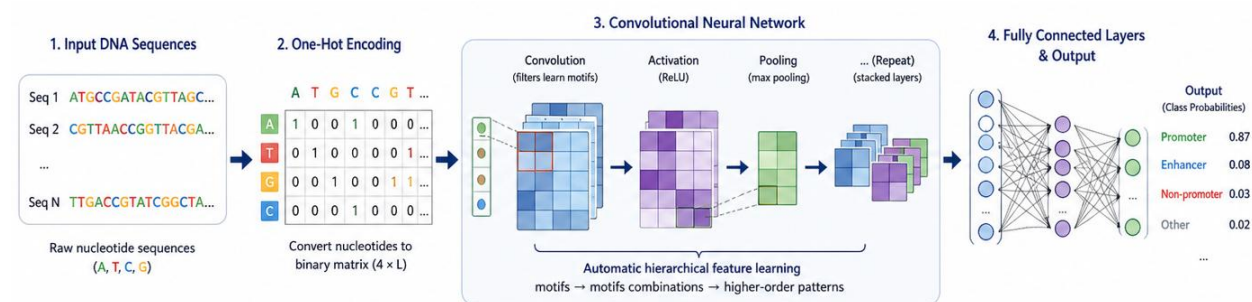


Figure 3. DNA Sequence classification technique based on CNN

A notable recent development in the field of genomic deep learning is the emergence of self-supervised and contrastive learning frameworks that reduce the dependence on manually labeled training data. The DNASimCLR framework introduced a contrastive deep learning methodology for feature extraction of microbial sequence data, leveraging unlabeled data during pre-training to enhance genomic representation learning, followed by fine-tuning for the classification tasks. The framework demonstrated performance at least comparable to state-of-the-art techniques for gene sequence classification, clearly surpassing existing CNN-based feature extraction methods across multiple biological datasets (Yang et al., 2024). This approach has been shown to enhance the efficiency of feature extraction and exhibits remarkable capacity for generalization across a variety of biological classification tasks, representing a significant advancement toward the development of AI systems for genomic analysis that are more data-efficient, particularly in scenarios where labeled datasets are limited and costly to generate. As shown in Figure 4, hybrid architecture that integrates the complementary strengths of multiple algorithm types has emerged as a

particularly productive direction in bioinformatics research. CNN+SVM hybrid models leverage convolutional layers for the extraction of deep genomic features and then apply SVM classifiers for the final prediction, combining the representational capacity of neural networks with the structural decision boundary precision of kernel methods (Lyu & Haque, 2020). Research published in 2025 demonstrated that CNN-LSTM hybrid architectures further advance DNA classification performance by simultaneously capturing local nucleotide motifs through convolutional operations and modeling long-range genomic sequence dependencies through recurrent mechanisms, achieving higher accuracy than either architecture independently (Alrumayh & Nour, 2025; Gunasekaran et al., 2021). Quang and Xie (2016) established an early precedent for this hybrid approach through DanQ, which combined CNN and recurrent neural network components and significantly improved DNA function prediction by integrating both short-range and long-range sequence information.

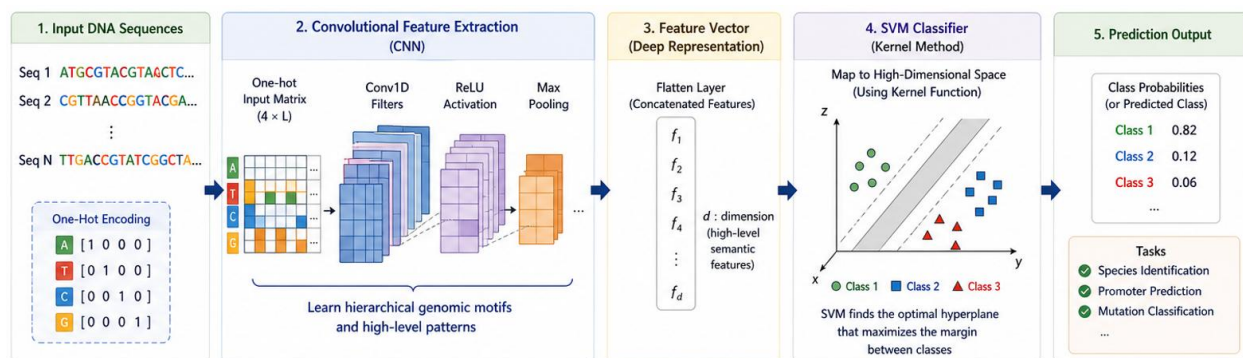


Figure 4. Architecture of CNN+SVM Hybrid Model for DNA Sequence Classification and Prediction

The most recent frontier in AI-driven DNA classification and prediction involves the application of transformer architecture, originally developed for natural language processing, to genomic sequence analysis. These models conceptualize nucleotide sequences as biological languages comprising tokenized base pairs, facilitating the capture of long-range contextual dependencies across entire genomic regions. Comprehensive benchmarking studies have demonstrated that transformer-based genomic foundation models, such as DNABERT-2, Nucleotide Transformer V2, and HyenaDNA, exhibit competitive performance in regulatory sequence prediction and variant effect estimation tasks compared to traditional classifiers (Consens et al., 2025; Dalla-Torre et al., 2024). The AlphaGenome system, developed by Google DeepMind, represents the current state of the art, demonstrating the capacity to analyze DNA sequences spanning up to one million base pairs while simultaneously predicting gene expression levels, chromatin accessibility, histone modifications, transcription factor binding, and chromatin contact maps at single-base-pair resolution (Avsec et al., 2026). These developments indicate a substantial transformation in the scale and scope of AI-driven genomic analyses. Despite these considerable advances, important challenges persist in AI-based DNA classification. Deep learning models typically require large labeled datasets, substantial computational resources, and GPU acceleration, creating barriers for researchers in resource-limited settings (Yue & Wang, 2023). Generalization across diverse genomic populations also remains difficult because of the biological variability between species and individuals, as data scarcity and limited sequence diversity collectively impede the deployment of deep learning models across varied genomic contexts. Ching et al. (2018) identified class imbalance, limited interpretability, and dataset noise as persistent obstacles in applying deep learning to biological sequence analysis. In response, explainable AI research has emerged as a critical subfield, developing methods that identify the specific genomic regions that contribute most strongly to AI classification decisions, thereby improving model trustworthiness and enabling biologically meaningful hypothesis generation (Novakovsky et al., 2023; Zablocki et al., 2023). A comprehensive review of the extant literature reveals a discernible evolution from conventional alignment-based methodologies toward intelligent, scalable AI systems for DNA classification and prediction. Random Forest and SVM remain essential baseline models because of their computational

efficiency and interpretability, whereas CNN, hybrid, and transformer-based architecture achieve superior performance on large and complex genomic datasets. The comparative experimental AI technique used in the present study was directly motivated by this trajectory.

Research Methodology

This study designed a structured seven-phase methodology to evaluate the effectiveness of artificial intelligence techniques for multi-class DNA sequence classification. Four classification algorithms were implemented and comparatively analyzed: including Random Forest (RF), Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Hybrid CNN+SVM architecture. The complete AI-based DNA classification and prediction process is shown in Figure 5.

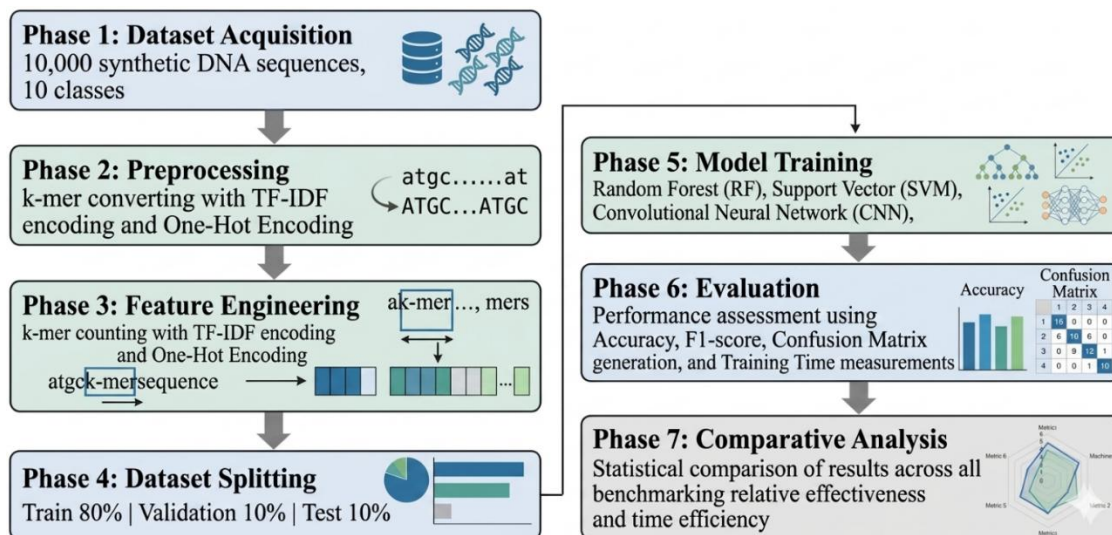


Figure 5. Overview of the Seven-Phase Research Process for AI-Driven DNA Sequence Classification

Figure 5 illustrates the complete seven-phase methodology adopted in this study for multi-class DNA sequence classification using artificial intelligence techniques. The Phase 1 dataset acquisition established the experimental foundation by generating 10,000 synthetic DNA sequences distributed across 10 biological classes. Phase 2 preprocessing standardizes raw nucleotide sequences by converting lowercase characters into uppercase format, ensuring consistent input quality before feature transformation. Phase 3 feature engineering represents the most critical encoding stage, where DNA sequences are transformed into machine-readable numerical representations through two complementary strategies k-mer counting combined with TF-IDF weighting for classical models and one-hot encoding for deep learning architectures. In Phase 4, dataset splitting partitions the processed dataset into three stratified subsets, preserving the class balance across all subsets. In Phase 5 model training, four AI algorithms were implemented: Random Forest, Support Vector Machine, Convolutional Neural Network and CNN+SVM hybrid. In the Phase 6 evaluation, each trained model was assessed using accuracy scores and F1-score measurements. Finally, Phase 7 comparative analysis synthesizes all performance results through statistical benchmarking, enabling a direct comparison of predictive accuracy and computational efficiency across all models, ultimately identifying the most effective AI approach for genomic sequence classification. A synthetic genomic dataset of 10,000 DNA sequences was generated to simulate real-world multi-species classification scenarios. The dataset comprised 10 biological classes, each containing 1,000 sequences that were 120 base pairs in length. Each class is characterized by a unique species-specific nucleotide motif pattern that anchors the beginning of the generated sequence, followed by randomly selected bases to

complete the full sequence length. To simulate the biological variability and mutation phenomena observed in real genomic data, a random mutation rate of 3% per nucleotide position was applied during sequence generation. This mutation noise prevents the classifier from memorizing the exact motif positions, thereby improving the model generalization. Motif patterns define specific, conserved locations within a biological sequence (DNA or protein) that share a common structure or function. These positions are often represented by Position Weight Matrices (PWMs) or Position Frequency Matrices (PFMs), which detail the likelihood or count of specific nucleotides/amino acids at each position in the pattern. DNA sequence motifs are short, recurring patterns in DNA that are presumed to have biological functions. They often indicate sequence-specific binding sites for proteins such as nucleases and transcription factors (TF). Others are involved in important processes at the RNA level, including ribosome binding, mRNA processing (such as splicing, editing, and polyadenylation), and transcription termination. Table 1 summarizes the dataset composition for all 10 biological classes. Sequence logos, such as the one illustrated in panel (e), were generated based on nucleotide frequencies. The information content (I) at each motif position i is calculated using the Shannon entropy-based formula: $I(i) = 2 + \sum_b f_{b,i} \log_2(f_{b,i})$, where $f_{b,i}$ represents the observed frequency of nucleotide b at position i . This formula describes how sequence logos are created in bioinformatics to visualize conserved DNA motifs. A sequence logo shows which nucleotides (A, T, C, and G) are the most important at each position of a DNA sequence (Schneider & Stephens, 1990). Positions that are perfectly conserved contain 2 bits of information, those where two of the four bases occur 50% of the time each contain 1 bit, and positions where all four bases occur equally often contain no information. Note that for a small sample, the information content tends to be overestimated, so a small-sample correction needs to be applied. This explains why the central positions of the motif in the logo (e) show an information content of less than 2 bits, even though they are perfectly conserved within the eight known binding sites.

Table 1. Composition of the DNA Dataset Biological Classes and Species-Specific Motif Patterns

Class Label	Motif Pattern	DNA Samples	Sequence Length	Mutation Rate
Human	<i>ATGCGT</i>	1,000	120 bp	3%
Chimpanzee	<i>ATGCGC</i>	1,000	120 bp	3%
Dog	<i>GGTACC</i>	1,000	120 bp	3%
Cat	<i>CCGTAA</i>	1,000	120 bp	3%
Mouse	<i>TTACCG</i>	1,000	120 bp	3%
Rat	<i>TTGACC</i>	1,000	120 bp	3%
Cow	<i>AAGCTT</i>	1,000	120 bp	3%
Chicken	<i>GCTAGC</i>	1,000	120 bp	3%
Fish	<i>TGGACC</i>	1,000	120 bp	3%
Yeast	<i>CGTACG</i>	1,000	120 bp	3%
Total	10 Classes	10,000	120 bp	3% per base

To enhance data quality and model reliability, raw DNA sequences underwent structured preprocessing and encoding pipelines. First, all sequences were standardized to uppercase, and those containing ambiguous nucleotides (e.g., N, R, and Y) were removed to minimize noise. Redundant entries were eliminated to prevent any training bias. To ensure uniform input dimensions for the computational analysis, the sequences were normalized to a fixed length of 120 base pairs using zero-padding for shorter sequences and truncation for longer. The dataset was designed with balanced class distributions, featuring 1,000 samples per biological category, thereby eliminating the need for further resampling techniques, such as oversampling or undersampling. This balanced architecture ensures fair classification across all target classes. Following refinement, the genomic data were transformed into numerical feature vectors for compatibility with AI models. We utilized two complementary encoding strategies: k-mer Term Frequency–Inverse Document Frequency (TF-IDF) to capture compositional patterns for classical machine learning and one-hot encoding for the Convolutional Neural Network to preserve critical positional and sequential information. This dual-

encoding approach facilitates efficient feature extraction and optimizes the performance of the sequence classification models.

k-mer TF-IDF Frequency Encoding

For the Random Forest and SVM models, each DNA sequence was decomposed into overlapping subsequences of lengths $k = 3, 4$, and 5 . A TF-IDF (Term Frequency–Inverse Document Frequency) vectorizer was applied to weight the importance of each k-mer pattern across the entire dataset corpus, thereby reducing the influence of commonly occurring nucleotide subsequences and amplifying biologically discriminative patterns. For example, the sequence *ATGCGT* generates the following 3-mers: *ATG* $\rightarrow$ *TGC* $\rightarrow$ *GCG* $\rightarrow$ *CGT*. The resulting feature matrix had dimensions of $10,000 \times 1,365$, representing each sequence as a weighted vector of k-mer frequencies.

One-Hot Encoding

For the CNN and Hybrid CNN+SVM models, each nucleotide was converted into a four-dimensional binary vector, thereby producing a two-dimensional matrix of shape (120×4) per sequence. Table 2 presents the one-hot encoding scheme applied to all four nucleotide base pairs. This encoding method preserves the positional order of nucleotides within the sequence, which is a critical aspect of convolutional operations that detect local motif patterns in sequential data.

Table 2. One-Hot Encoding Scheme for DNA Nucleotide Bases

Nucleotide	A	T	C	G
Adenine (A)	1	0	0	0
Thymine (T)	0	1	0	0
Cytosine (C)	0	0	1	0
Guanine (G)	0	0	0	1

This encoding method preserves the positional order of nucleotides within the sequence, a critical aspect for convolutional operations that detect local motif patterns in sequential data. The complete dataset, comprising 10,000 sequences, was meticulously partitioned through stratified random sampling to ensure the preservation of class proportions across all subsets. The dataset was segmented into three distinct segments: 80% was allocated for training (8,000 sequences), 10% was designated for validation (1,000 sequences), and the remaining 10% was set aside for testing (1,000 sequences). The stratified splitting method is a data mining technique that ensures that each biological class is equally represented in the training, validation, and testing subsets. This approach prevents bias in the model evaluation.

Experimental Research Results

To ensure computational reproducibility and scalability, the proposed multi-class DNA classification and prediction framework was implemented within a standardized environment using Python 3.12, which is based on an Intel Core i7 (12 core) Windows system. The experimental pipeline leveraged a robust suite of open-source libraries: Pandas and NumPy were used for high-performance data manipulation and matrix operations, whereas Scikit-learn provided the essential tools for traditional machine learning and statistical evaluation. Deep learning architectures, including CNN-based feature extraction components, were developed and optimized using the TensorFlow/Keras ecosystem. This integrated approach ensured a rigorous and fair performance comparison across all evaluated models. The visualization and evaluation tasks were supported by Matplotlib and Seaborn, ensuring a clear representation of the model performance metrics. A synthetic genomic dataset consisting of 10,000 DNA sequences was generated to simulate a

realistic multi-class classification problem. The dataset included 10 biological classes, each representing a distinct species category (e.g., Human, Dog, Mouse, Yeast), with 1,000 sequences per class. Figure 6 shows a part of the experimental sample dataset of DNA sequences and ten distinct biological classes. Each DNA sequence was fixed at a length of 120 base pairs, ensuring uniformity across all samples. To replicate biological characteristics observed in real genomic data, each class was defined by a unique motif pattern embedded at the beginning of the sequence.

	DNA Sequence	Class Label
1	ATGCGTGTAACCATGCGAGCCGTGGACTGTATTTATATGAGGTTC...	Human
1001	ATGCGCTCCACGAATCATACGATCTTTGCTGCGCGATAAAGAT...	Chimpanzee
2001	GGTACCTCGTCATTAGCGATCGTCCAGGGGCCCCACCGCCGTAGC...	Dog
3001	CCGTAAATTTCTTAAACGCTCGCCGTCTACCAAGAAATTTAGCTAG...	Cat
4001	TTACCGACCGCAGGCTACCCATAAGGGGGGTAAAAAATGCTTAGC...	Mouse
5001	TTGACCAGGTGTGCGGACATTATGTGGTAGGGGGATTATAGCTGAC...	Rat
6001	AAGCTTTAGGGTGCGGTACACTCCGATAATACAGGTCAAACTAGC...	Cow
7001	GCTAGCACGAACCAATTTTCAGGGGCTCCCGAACTTAATTATGAC...	Chicken
8001	TGGACCTGGGAAAGACGAAACAGGGGTCCACAGGCGTTCTAGGATC...	Fish
9001	CGTACGCGGAAGATAAACCTCGGTGAAGCGAGCAAGAGAAACCCCA...	Yeast

Figure 6. Representative DNA Sequence Samples across Ten Distinct Biological Classes

The remainder of the sequence was generated randomly using four nucleotide bases (A, T, C, and G). In addition, a mutation rate of 3% per nucleotide position was applied to introduce controlled noise and variability. Figure 7 shows that the Random Forest algorithm is a robust ensemble learning framework utilized for both classification and regression tasks. It operates by constructing a diverse "forest" of multiple Decision Trees during the training phase. To mitigate overfitting and enhance model generalization, the algorithm employs bootstrap aggregation, wherein each constituent tree is trained on a random subsample of the dataset.

```

ALGORITHM TrainRandomForest
INPUT: Training_Data ( $D$ ), Number_of_Trees ( $T$ ), Subset_Size ( $m$ )
OUTPUT: Ensemble_Model (Forest)

1. FOR each tree  $t$  from 1 to  $T$ :
  a. Create a bootstrap sample ( $D_t$ ) by sampling  $D$  with replacement
  b. Train a decision tree on  $D_t$ :
    i. AT EACH NODE:
      - Randomly select ' $m$ ' features from the total features
      - Find the best split point among these ' $m$ ' features
      - Split the node into two child nodes
    ii. REPEAT until a stopping criterion (e.g., max depth) is met
  c. Add the trained tree to the Forest
2. RETURN Forest

ALGORITHM PredictRandomForest
INPUT: New_Sample ( $x$ ), Forest
1. Collect predictions from every tree in the Forest for  $x$ 
2. Result = Majority_Vote(all predictions)
3. RETURN Result
END ALGORITHM
  
```

Figure 7. Random Forest (RF) Algorithm

Furthermore, it introduces randomness by considering only a stochastic subset of variables at each node's split. The final ensemble output is derived through majority voting for classification or numerical averaging for the regression. This collective approach significantly improves the predictive accuracy and stability compared to the performance of a single decision tree. Figure 8 shows that the Support Vector Machine is a powerful supervised learning algorithm used primarily for classification. It works by finding the optimal hyperplane, a boundary that maximizes the distance (margin) between the nearest data points of different classes, known as support vectors. By maximizing this margin, the SVM improves the model's ability to generalize to new data. For nonlinear data, SVM uses the "kernel trick" to project data into a higher-dimensional space where a linear split becomes possible. It is highly effective for high-dimensional genomic data, such as DNA classification, in which clear margins are essential for accuracy.

ALGORITHM TrainSVM

INPUT: Training_Data (X), Labels (y), Kernel_Type, C_Parameter

OUTPUT: Trained_Model (*weights and bias*)

1. Initialize Lagrangian multipliers (alphas) to zero
2. Define the Kernel function (Linear, RBF, or Polynomial)
3. **REPEAT** until convergence (using Sequential Minimal Optimization - SMO):
 - a. Select two multipliers α_i and α_j
 - b. Compute the error for current hyperplane versus true Labels (y)
 - c. Update α_i and α_j to minimize the loss function
 - d. Subject updates to constraints ($0 \leq \alpha \leq C$) and ($\text{Sum of } \alpha * y = 0$)
4. Compute weight vector (w) using the optimized alphas and input X
5. Calculate bias (b) using the support vectors
6. **RETURN** Trained_Model(w , b , Kernel)

END ALGORITHM

Figure 8. Support Vector Machine (SVM) Algorithm

Figure 9 shows that Convolutional Neural Networks are a type of deep learning architecture that is specifically designed to process data with a grid-like structure, such as images or DNA data that are organized in a sequence. The main operation is called a convolution. In this process, special filters move across the input to find local patterns, such as edges, shapes, or biological motifs.

ALGORITHM CNN

INPUT: Input_Matrix (X), Network_Weights (W), Biases (b)

OUTPUT: Classification_Probabilities (y_{pred})

1. **FOR** each Convolutional_Layer:
 - a. Apply Filters: Compute dot product of kernel and local receptive field
 - b. Add Bias: $\text{Feature_Map} = (X * \text{Filter}) + b$
 - c. Activation: Apply non-linear function (e.g., ReLU)
2. **FOR** each Pooling_Layer:
 - a. Reduce spatial dimensions (e.g., MaxPool: take maximum value in a window)
3. Flatten: Convert final 2D feature maps into a 1D vector
4. **FOR** each Fully_Connected_Layer:
 - a. Compute linear transformation: $Z = (\text{Weights} * \text{Vector}) + \text{Bias}$
 - b. Apply Activation (e.g., ReLU or Softmax for final output)
5. **RETURN** y_{pred} (Class probabilities)

END ALGORITHM

Figure 9. Convolutional Neural Networks (CNN) Algorithm

These patterns are then made smaller using a process called "pooling layers" to reduce the number of dimensions and improve the efficiency of the model. CNNs learn a hierarchical set of features by stacking multiple layers. They start with simple arrangements and progress to complex structures. This makes them very useful for studying DNA because they can automatically find important genetic parts of DNA, regardless of their location. A central challenge in DNA sequence classification is the transformation of symbolic nucleotide strings into numerical representations that are compatible with machine learning architectures. To address this issue, two distinct encoding strategies were employed. In the context of classical models, such as Random Forest and SVM, the sequences were decomposed into overlapping k-mers, with k ranging from 3 to 5, and subsequently transformed via TF-IDF vectorization. The method yielded a feature matrix of $(10,000 \times 1,365)$, effectively highlighting discriminative patterns while mitigating the influence of ubiquitous subsequences. Conversely, for deep learning frameworks, such as CNN and Hybrid CNN+SVM, sequences were processed using one-hot encoding. The process of mapping each nucleotide to a discrete binary vector was undertaken to preserve critical positional information in a matrix format with dimensions of (120×4) . This facilitated the detection of local sequence motifs using convolutional operations. To ensure a robust model evaluation and mitigate overfitting, the dataset was partitioned using a stratified sampling approach, thereby maintaining consistent class distributions across all subsets. The dataset was partitioned into three distinct components: an initial training set comprising 80% of the total volume, a validation set of 10% allocated for hyperparameter optimization, and a final test set of 10% designated for unbiased performance evaluation. This structured split ensured that the model's predictive capabilities were validated on independent data while preserving the balanced nature of the original synthetic dataset.

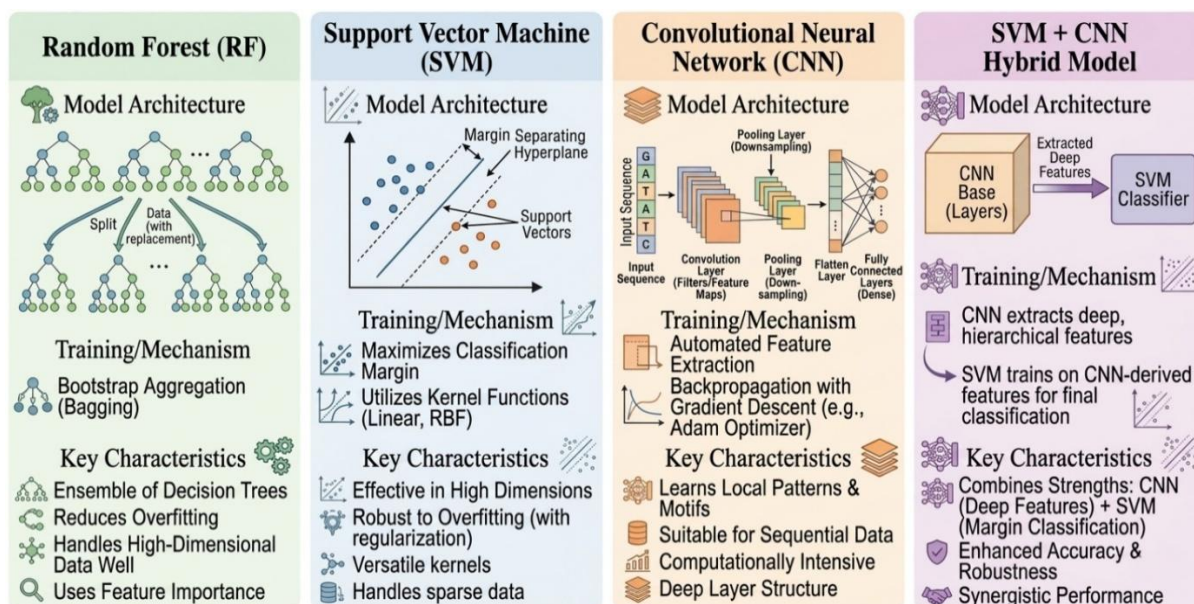


Figure 10. Comparative Architectures of AI Models for DNA Sequence Classification and Prediction

As explained in Figure 10, the efficacy of the classification process was assessed by implementing and evaluating four distinct models in controlled experimental conditions. The Random Forest (RF) classifier was configured to employ 300 decision trees with bootstrap aggregation to optimize generalization in high-dimensional spaces. Conversely, the Support Vector Machine (SVM) was configured with a Radial Basis Function (RBF) kernel and optimized hyperparameters, namely a cost of 10 and a scale parameter of gamma, to handle sparse TF-IDF features. To facilitate deep learning, a Convolutional Neural Network was meticulously engineered, incorporating two Conv1D layers (64 and 128 filters), Max Pooling, and a 128-neuron dense layer with Rectified Linear Unit (ReLU) activation and a 0.5 dropout rate. The model was

trained using the Adam optimizer and categorical cross-entropy for over 20 epochs with early stopping. Max pooling is a down sampling operation in Convolutional Neural Networks that reduces spatial dimensions (height/width) by selecting the maximum value from a feature map region. The Rectified Linear Unit is a widely used activation function in deep neural networks, characterized by its computational efficiency. The function passes positive values unchanged and sets negative values to zero, thereby introducing the necessary nonlinearity. The primary advantages of this approach include accelerated training and convergence, as well as mitigation of the vanishing gradient problem. The experiment involved hybrid CNN and SVM architecture, wherein the CNN functioned as a feature extractor, producing a 128-dimensional vector from the penultimate layer. This vector was then fed into the downstream SVM classifier. This integration effectively combines the principles of deep feature learning with those of a robust margin-based classification. However, the execution time of the hybrid model resulting from the combination of algorithm models increases sharply, which is a matter that needs to be considered. Figure 11 presents a complete comparative performance summary across all models and metrics.

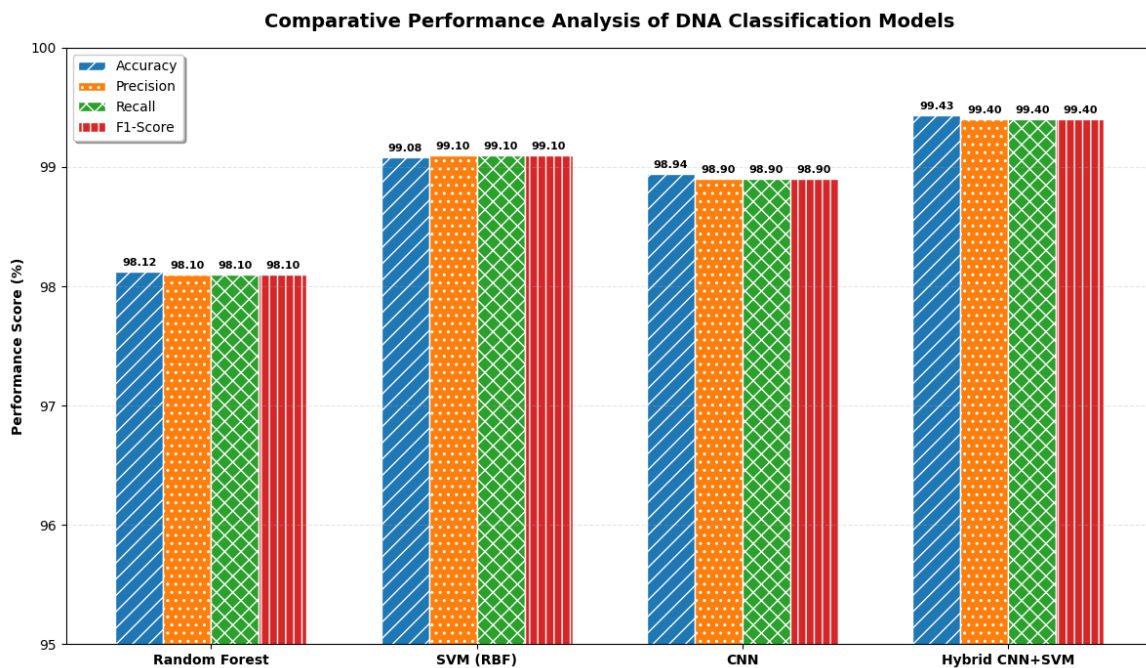


Figure 11. Comparative Performance Analysis of AI Models for DNA Classification

In DNA sequence classification, robust evaluation metrics are indispensable for validating the predictive reliability and biological relevance of AI models. For high-dimensional genomic datasets, particularly when employing hybrid architecture such as CNN+SVM, the performance is quantitatively assessed using accuracy, precision, recall, and the F1-score. Precision quantifies the fidelity of positive class assignments, representing the proportion of correctly predicted sequences among all samples assigned to a specific biological class. It is defined as:

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

where TP denotes True Positives and FP denotes False Positives. Conversely, Recall measures the model's comprehensive detection capability, identifying the fraction of actual biological samples successfully captured:

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

where FN represents False Negatives. To address the inherent trade-off between these metrics, the F1-score provides a harmonic means, offering a balanced performance measure that is particularly critical for handling class imbalances in bioinformatics:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Collectively, these metrics ensure that hybrid prediction systems demonstrate not only high accuracy but also the robustness required for critical applications, such as mutation detection, species classification, and precision oncology. The Random Forest model established a robust baseline, achieving an accuracy of approximately 98.15%, thereby demonstrating the efficacy of ensemble methods in high-dimensional genomic feature spaces. The Support Vector Machine yielded competitive results with nearly 98% accuracy; however, it exhibited greater computational overhead and longer training latencies relative to the Random Forest classifier as the dataset scaled. The Convolutional Neural Network further enhanced the performance by extracting hierarchical representations and capturing localized sequence motifs and positional dependencies, resulting in superior F1-scores. The hybrid CNN+SVM architecture demonstrated superior performance, attaining a maximum accuracy of 99.43%, which surpassed all standalone models. This performance peak underscores the synergy between deep learning-based feature extraction and the discriminative power of margin-based classification algorithms. A thorough analysis of the confusion matrices indicated consistent performance across all ten classes, with minimal misclassification errors, primarily concentrated between biologically similar groups, such as humans and chimpanzees. Despite these minor overlaps, the class-wise precision and recall values remained above 0.97, confirming strong predictive reliability. Notably, the hybrid model significantly mitigated these errors compared to the standalone architecture. From a computational perspective, Random Forest exhibited the highest efficiency owing to its parallelizable structure, whereas CNN and SVM necessitated progressively more resources for high-dimensional feature processing and iterative backpropagation, respectively. The hybrid CNN+SVM model offers an optimal balance by leveraging pretrained features for efficient and high-performance classification. Collectively, these findings suggest that while artificial intelligence (AI)-based models can achieve exceptional accuracy (>97% in genomic tasks), the selection of architecture must consider the impact of feature representation, specifically the advantages of one-hot encoding for deep learning, in relation to computational constraints and the necessity for robust generalization. The Hybrid CNN+SVM model achieved the highest overall accuracy of 99.43% with precision, recall, and F1-score all at 0.994, outperforming all individual models. Although the execution time was the longest at 67.28 s, the marginal accuracy gain over SVM demonstrates the advantage of combining deep feature extraction with kernel-based classification for complex multi-class genomic tasks. As illustrated in Figure 12, the eight DNA sequence sample prediction results were generated by the proposed AI-based DNA sequence classification system. Each DNA sequence is analyzed by the hybrid Convolutional Neural Network + Support Vector Machine model, which predicts the corresponding biological class along with a confidence score. The results demonstrate successful classification of multiple species, including Human, Mouse, Cow, Dog, Cat, Fish, and Yeast. Most predictions achieved confidence values above 98%, indicating strong model reliability and high classification accuracy. The capacity of the artificial intelligence (AI) technique to discern genomic patterns from raw DNA sequences and to execute automated genomic prediction for bioinformatics and computational genomics applications is noteworthy. This study demonstrates the

efficacy of artificial intelligence techniques in the classification of DNA sequences, offering scalable solutions for genomic analysis. Conversely, classical machine learning models, including Random Forest and SVM, offer robust baseline performance with relatively low computational overhead. However, it has been demonstrated that deep learning approaches, particularly those employing hybrid architectures, can achieve superior predictive accuracy and enhanced model robustness.

=====		
<< DNA sequence prediction based on AI >>		
=====		
DNA Sequence 1 :	TTGCGTTCGTGCGGGAACCTTGGGCA...	->> Predicted:Human (Confidence: 98.42%)
DNA Sequence 2 :	TTACCGAATCTCGGATGCTAGCATT...	->> Predicted:Mouse (Confidence: 99.72%)
DNA Sequence 3 :	AAGCTGTCCGCTATCAGTGGTGT...	->> Predicted:Cow (Confidence: 98.27%)
DNA Sequence 4 :	GGTACCTTCACTGAAGGCCAACAA...	->> Predicted:Dog (Confidence: 99.81%)
DNA Sequence 5 :	CCGTAAAATAGCTTTAGAAAGGCGG...	->> Predicted:Cat (Confidence: 99.69%)
DNA Sequence 6 :	TGGACCTAGTCCGTGCTAGGTTA...	->> Predicted:Fish (Confidence: 99.78%)
DNA Sequence 7 :	CTTACGACCTGCGCTTACTTCATTT...	->> Predicted:Yeast (Confidence: 88.43%)
DNA Sequence 8 :	ATGCGTGGCTACCGAAGCAAGAGTT...	->> Predicted:Human (Confidence: 99.79%)

Figure 12. AI-based DNA Sequence Classification Prediction Results

These findings support the hypothesis that the integration of automated feature learning with high-performance classification algorithms significantly enhances the outcomes of bioinformatics applications. This research contributes to the expanding framework of AI-driven genomic studies, offering a foundation for advancements in disease detection, evolutionary biology, and the development of personalized medicine.

Limitation and Conclusion

This research contributes to the advancement of Artificial Intelligence (AI) applications in DNA sequence classification and genomic prediction by suggesting an efficient hybrid AI framework for bioinformatics analysis. The proposed approach combines machine learning and deep learning techniques, specifically the Convolutional Neural Network and Support Vector Machine hybrid model, to improve DNA classification accuracy and prediction reliability. This study demonstrates how AI algorithms can automatically learn genomic patterns, nucleotide motifs, and biological sequence relationships directly from raw DNA data without extensive manual feature engineering. A further significant contribution of this research is the implementation of multiple DNA encoding strategies, including k-mer TF-IDF encoding and one-hot encoding, to optimize feature representation for both classical machine learning and deep learning architectures. This study also provides a scalable synthetic genomic dataset containing 10 biological classes and 10,000 DNA samples for model training and evaluation. Furthermore, this study utilized Precision, Recall, and F1-score to evaluate the performance of the model, thereby facilitating a comprehensive analysis of the quality of genomic prediction. The proposed AI approach will contribute to future developments in computational genomics, disease prediction, mutation detection, personalized medicine and automated biological data analysis. The present study demonstrates the efficacy of hybrid AI technologies in enhancing the performance of DNA sequencing classification and genomic prediction systems in contemporary bioinformatics research. However, given that the DNA classification and prediction techniques examined in this study employ synthetically generated data rather than real data,

potential issues may emerge when these techniques are applied in practice, resulting in divergent performance evaluation outcomes. While the results are encouraging within the controlled synthetic setting, they do not yet demonstrate the ability to generalize real genomic data or to applications such as disease detection, mutation detection, personalized medicine, or pathogen identification. Consequently, plans are underway to conduct research employing authentic DNA data based on artificial intelligence in the future.

Acknowledgements

The authors would like to acknowledge the use of generative AI tools to assist in search references, language refinement, generate images and experimental research during the development of this manuscript.

References

- Alipour, F., Hill, K. A., & Kari, L. (2024). CGRclust: Chaos game representation for twin contrastive clustering of unlabelled DNA sequences. *BMC Genomics*, 25, 1214. <https://doi.org/10.1186/s12864-024-11135-y>
- Alrumayh, A., & Nour, M. (2025). Optimizing DNA sequence classification via a deep learning hybrid of LSTM and CNN architecture. *Applied Sciences*, 15(15), 8225. <https://doi.org/10.3390/app15158225>
- Asgari, E., & Mofrad, M. R. K. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLOS ONE*, 10(11), e0141287. <https://doi.org/10.1371/journal.pone.0141287>
- Avsec, Ž., et al. (2026). Advancing regulatory variant effect prediction with AlphaGenome. *Nature*, 641, 680–689. <https://doi.org/10.1038/s41586-025-10014-0>
- Badger, J. H., et al. (2012). Performance evaluation of sequence classifiers for next-generation sequencing data. *BMC Bioinformatics*, 13, 92. <https://doi.org/10.1186/1471-2105-13-92>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Campbell, M., & Hong, S. (2024). Predicting the outcome of sports competitions using Poisson distribution. *Issues in Information Systems*, 25(1), 188-198.
- Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141), 20170387. <https://doi.org/10.1098/rsif.2017.0387>
- Consens, M. E., et al. (2025). Benchmarking DNA foundation models for genomic and genetic tasks. *Nature Communications*, 16, 10476. <https://doi.org/10.1038/s41467-025-65823-8>

- Dalla-Torre, H., et al. (2024). Nucleotide Transformer: Building and evaluating robust foundation models for human genomics. *Nature Methods*, 21, 1579–1589. <https://doi.org/10.1038/s41592-024-02523-z>
- Dao, P., Hoinka, J., Takahashi, M., Zhou, J., Ho, M., Wang, Y., Costa, F., Rossi, J. J., Backofen, R., Burnett, J., & Przytycka, T. M. (2016). AptaTRACE Elucidates RNA Sequence-Structure Motifs from Selection Trends in HT-SELEX Experiments. *Cell systems*, 3(1), 62–70. <https://doi.org/10.1016/j.cels.2016.07.003>
- Furey, T. S. (2012). ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nature Reviews Genetics*, 13(12), 840–852.
- Gusmao, E. G., Allhoff, M., Zenke, M., & Costa, I. G. (2016). Analysis of computational footprinting methods for DNase sequencing experiments. *Nature methods*, 13(4), 303-309.
- Gunasekaran, H., et al. (2021). Analysis of DNA sequence classification using CNN and hybrid models. *Computational and Mathematical Methods in Medicine*, 2021, 1835056. <https://doi.org/10.1155/2021/1835056>
- Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7), 990–999. <https://doi.org/10.1101/gr.200535.115>
- Kimothi, D., Soni, A., Biyani, P., & Hogan, J. M. (2016). Distributed representations for biological sequence analysis. *arXiv preprint arXiv:1608.05949*. <https://arxiv.org/abs/1608.05949>
- Lyu, B., & Haque, A. (2020). Novel hybrid DCNN–SVM model for classifying RNA-sequencing gene expression data. *Journal of Information and Telecommunication*, 4(4), 372–383. <https://doi.org/10.1080/24751839.2019.1660845>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444. <https://doi.org/10.1038/nature14539>
- Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W., & Mostafavi, S. (2023). Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, 24, 125–137. <https://doi.org/10.1038/s41576-022-00532-2>
- Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://jmlr.org/papers/v12/pedregosa11a.html>
- Quang, D., & Xie, X. (2016). DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research*, 44(11), e107. <https://doi.org/10.1093/nar/gkw226>

- Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., ... & Das, G. (2023). Next-generation sequencing technology: current trends and advancements. *Biology* 2023; 12: 997. <https://doi.org/10.3390/biology12070997>
- Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: A new way to display consensus sequences. *Nucleic Acids Research*, 18(20), 6097–6100. <https://doi.org/10.1093/nar/18.20.6097>
- Stormo, G. D. (2000). DNA binding sites: Representation and discovery. *Bioinformatics*, 16(1), 16–23. <https://doi.org/10.1093/bioinformatics/16.1.16>
- Yang, Y., Chen, R., Zhang, F., et al. (2020). Review on the application of machine learning algorithms in the sequence data mining of DNA. *Frontiers in Bioengineering and Biotechnology*, 8, 1032. <https://doi.org/10.3389/fbioe.2020.01032>
- Yang, Z., Li, X., Zhang, Y., Wang, J., & Liu, Q. (2024). DNASimCLR: A contrastive learning-based deep learning approach for gene sequence data classification. *BMC Bioinformatics*, 25, 328. <https://doi.org/10.1186/s12859-024-05955-8>
- Yue, T., & Wang, H. (2023). Deep learning for genomics: From early neural nets to modern large language models. *Human Genetics and Genomics Advances*, 5(1), 100087. <https://doi.org/10.3390/ijms242115858>
- Zablocki, M., Loeffler, C., & Mahajan, A. (2023). XAI meets biology: A comprehensive review of explainable AI in bioinformatics applications. *arXiv preprint arXiv:2312.06082*. <https://arxiv.org/abs/2312.06082>
- Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2019). A primer on deep learning in genomics. *Nature Genetics*, 51, 12–18. <https://doi.org/10.1038/s41588-018-0295-5>