Deep Vision For Genomic Data

LOKESH PUROHIT[1], **SAHAR HOOSHMAND**[1]

[1]Department of Computer Science, California State University-Dominguez Hills, Carson, CA, USA
E-MAIL: lpurohit1@toromail.csudh.edu, hooshmand@csudh.edu

Abstract:
Sequence classification is a core task in computational genomics with wide-ranging applications in gene regulation, functional annotation, and disease prediction. In this study, we propose a novel hybrid deep learning architecture that combines Vision Transformers (ViTs) with Convolutional Neural Networks (CNNs) for classifying human non-TATA promoter sequences. To harness the representational power of vision-based models, we introduce a multi-channel Hilbert Curve encoding technique that transforms linear DNA sequences into 2D image-like grids, spatially preserving k-mer relationships and regulatory motifs. This spatial restructuring enables CNNs to extract local features while allowing ViTs to capture global dependencies via self-attention. Evaluated on a balanced dataset of 36,131 promoter and non-promoter sequences (251 bp each), the proposed model achieves a validation accuracy of 90.28%, along with strong precision and recall across both classes. Our approach demonstrates that image-based sequence representation, when paired with hybrid architectures, offers a powerful alternative to traditional sequence modeling.

Keywords:
Vision Transformers; Convolution Neural Networks; Hilbert Curve; Deep Learning; Image Classification; DNA Classification; Hybrid Architecture; Multi-channel Hilbert Images

1. Introduction

Sequence classification is a widely explored problem with applications across several domains, including genomics, time series forecasting, intrusion detection, and natural language processing. In the domain of genomics, it has played a vital role in classifying DNA, RNA, or protein sequences, enabling downstream tasks such as gene regulation analysis, disease association studies, and functional annotation [1].

Traditionally, sequence classification methods are categorized into three types: (1) feature-based classification, where sequences are converted into feature vectors to apply conventional classifiers; (2) distance-based classification, where similarities among sequences are quantified and clustered accordingly; and (3) model-based classification using probabilistic models like Hidden Markov Models (HMMs) that learn statistical structures from sequences.

In recent years, deep learning methods— particularly CNNs— have proven effective in learning local motif patterns in biological sequences without the need for hand-engineered features [2]. However, CNNs are inherently limited in capturing long-range dependencies across nucleotides, which are often crucial in identifying regulatory elements like promoters. To address this limitation, researchers have explored Transformer-based architectures, originally introduced for natural language processing, for their ability to model global context using self-attention mechanisms [3].

This study introduces a hybrid deep learning framework combining CNNs with ViTs for classifying human non-TATA promoter sequences. CNN layers are employed to extract local motifs and spatial features from 2D image-like representations of DNA sequences generated using Hilbert Curve encoding, which spatially reorganizes 1D nucleotide sequences into 2D grids that preserve locality and structure. Stacked Transformer encoder blocks are then applied to capture long-distance dependencies among sequence elements. This integration enables a more expressive and comprehensive representation of promoter signals embedded within DNA. The goal of this work is to develop a high-performing deep learning model for promoter identification using an architecture capable of capturing both localized motif patterns and long-range dependencies. While the hybrid CNN–ViT model primarily focuses on maximizing classification performance, its structured representation of spatial and contextual features opens future avenues for enhancing interpretability. The model demonstrates strong potential in preliminary tests and may be applicable to other sequence classification tasks in computational biology.

2. Related Work

Deep learning models have increasingly become the foundation for sequence classification tasks across multiple disciplines. Models such as CNNs, Long Short-Term Memory networks (LSTMs), and bidirectional LSTMs (BiLSTMs) have been widely adopted in fields such as speech recognition, financial prediction, and bioinformatics [4, 5]. Specifically, CNNs and feature-based architectures have shown remarkable effectiveness in classifying genomic and proteomic sequences by identifying local motifs and patterns within raw input data [6, 7].

In the realm of genomic analysis, CNNs have been successfully employed for promoter prediction, enhancer detection, transcription factor binding site classification, and chromatin accessibility modeling. For example, DeepBind and DeepSEA demonstrated how CNNs could automatically extract high-resolution features from DNA sequences, reducing the need for manual motif engineering [8]. Similarly, hybrid methods that combine CNNs with LSTMs have been proposed to model both local and contextual information in genomic data. These methods, however, often struggle with capturing long-range dependencies that span beyond the receptive field of convolutional filters or recurrent memory limitations [6].

To overcome this, attention-based models—most notably Transformer architectures—have been introduced. Originally developed for natural language processing, Transformers have demonstrated strong performance on sequence data due to their ability to learn relationships across entire input sequences using self-attention. Applications such as DNABERT, which adapts BERT-style pretraining for DNA k-mer sequences, and Enformer, which models regulatory activity across thousands of bases, underscore the growing impact of attention mechanisms in genomics [9, 10].

Recent efforts have explored integrating CNNs with Transformers into a unified framework. This hybrid architecture leverages the local feature extraction capability of CNNs alongside the global context modeling of Transformers [3]. Vision Transformers, in particular, have become a promising solution by treating genomic representations as images or patch-based embeddings, making it possible to learn from both spatial and sequential features simultaneously [11]. While ViTs are data-hungry, they have been shown to outperform CNNs in some domains when trained properly or combined with convolutional preprocessing layers.

In the specific context of promoter classification, several studies have tackled the challenge of distinguishing canonical and non-canonical promoters. Non-TATA promoters, which do not contain the traditional TATA box motif, are especially difficult to classify due to their sequence variability and subtle regulatory signals [12]. Traditional machine learning approaches such as Support Vector Machines (SVMs) have been employed with hand-crafted features, but scalability and generalizability have remained concerns [13].

Our work contributes to this evolving landscape by proposing a CNN + ViT hybrid model tailored for the classification of non-TATA promoter sequences in the human genome. The model is designed to effectively capture both short-range motifs and long-range dependencies, providing a comprehensive representation of promoter characteristics. By using Hilbert Curve encoded DNA sequences transformed into 2D images, and applying convolutional layers followed by Transformer encoders, the model takes advantage of both inductive biases [14]. Furthermore, this study addresses not just the accuracy of promoter classification but also its scalability. With over 36,000 sequences in the dataset, efficient learning and generalization become essential. Preliminary results from our experiments suggest that hybrid architectures of vision models may offer a superior balance of performance and interpretability for regulatory element prediction, and represent a promising direction for future genomic research.

3. Hybrid CNN ViT Model for DNA Classification

3.1. Background

Recent advancements in artificial intelligence, particularly in deep learning, have significantly enhanced sequence-based classification tasks, including applications in genomics [15]. CNNs have been widely used for analyzing spatial patterns in image-like data, showing strong capabilities in learning localized features. In genomic analysis, CNNs can effectively extract important motifs and short-range dependencies from sequence representations, enabling robust feature extraction without manual engineering. On the other hand, ViTs leverage the self-attention mechanism to model global dependencies across input data sequences. The self-attention mechanism calculates the relationships between all parts of the input simultaneously, enabling the model to capture complex, long-range patterns that CNNs may miss. Mathematically, the scaled dot-product attention can be defined as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

where q_i, k_i, and v_i, represent the ith rows of matrices Q, K, and V respectively. Moreover, d_k represents the dimensionality

of the input queries and keys, acting as a mechanism to mitigate the potential computational instability caused by large inputs in the attention function, thus ensuring computational stability.

It was discovered to be advantageous to linearly transform the queries, keys, and values multiple times (denoted as h) using different learnable matrices, each with dimensions d_k, d_k, and d_v. Specifically, each set of these learnable matrices is referred to as a "head," and the transformer model incorporates a Multi-head Self-Attention (MSA) layer. These output heads are concatenated, as shown in equation (2), to create a unified output that is then fed into the feed-forward layer. The feed-forward layer, a multi-layer perception (MLP) network, aids in identifying various features within the input sequence. This structure significantly enhances performance and enables parallel processing of sequence data, particularly in Natural Language Processing applications.

$$MSA(Q, K, V) = Concat(head_1, ..., head_h)W^O \quad (2)$$

Recently, the transformer model has garnered significant interest due to its exceptional performance in image processing, particularly for classification purposes. Vision transformers, operate by dividing an input image into patches and treating these patches as a sequence of linear embeddings. The use of CNNs and transformers in this capacity represents a cutting-edge approach in medical technology. In the context of DNA sequence analysis, combining the local feature extraction power of CNNs with the global context modeling ability of transformers provides a powerful hybrid architecture. The CNN captures sequence motifs and short-range dependencies, while the transformer captures relationships across distant regions of the DNA sequence, leading to a more holistic understanding crucial for complex classification tasks.

3.2 Proposed Pipeline

The proposed hybrid framework integrates a custom Convolutional Neural Network (CNN) and a Vision Transformer (ViT) to classify DNA sequences efficiently. The DNA sequences, specifically human non-TATA promoter regions, are first converted into multi-channel Hilbert Curve image representations. Each nucleotide (A, T, G, C) is one-hot encoded into separate channels, and a fifth channel encodes the relative position to the Transcription Start Site (TSS). This encoding preserves both local sequence information and spatial organization critical for downstream analysis.

The custom CNN serves as the initial feature extractor, applying multiple convolutional and pooling layers to learn local sequence patterns embedded in the Hilbert images. The CNN architecture is tailored to effectively process the five-channel input, extracting low-level spatial features across the encoded sequence.

Once feature maps are generated by the CNN, they are partitioned into fixed-size patches. Each patch is linearly embedded into a high-dimensional vector space, and positional encodings are added to maintain spatial relationships. These embeddings are then fed into a transformer encoder composed of multi-head self-attention layers and feed-forward networks. The transformer layers capture long-range dependencies across the entire sequence, modeling complex genomic relationships.

The output of the transformer encoder is flattened and passed through a multi-layer perceptron (MLP) head with GELU activation. A final classification layer with a softmax activation function predicts whether the DNA sequence belongs to the promoter class or not. Dropout layers are incorporated within the MLP and transformer blocks to prevent overfitting.

Gaussian noise is added during training as a form of data augmentation to improve model robustness. Furthermore, early stopping based on validation loss is employed to halt training when overfitting is detected, ensuring a well-generalized model.

This comprehensive pipeline captures both localized and globalized features within DNA sequences, providing an effective framework for genomic sequence classification.

4. Experimental Results

In this section, the dataset specifics, preprocessing steps, implementation details of the proposed model are provided and discussed. Additionally, the experimental results are thoroughly examined, focusing on performance metrics and the model's effectiveness.

4.1. Dataset

Our study utilized the human non-TATA promoters dataset from the genomic-benchmarks repository [16], which contains 36,131 DNA sequences of precisely 251 base pairs each. These sequences span from -200 to +50 base pairs relative to the transcription start site (TSS). The dataset comprises 14,742 positive samples (actual non-TATA promoters) and 12,355 negative samples in the training set, along with 4,915 positive and 4,119 negative samples in the testing set, representing a slight class imbalance ratio of 1.2:1. The negative sequences were carefully constructed from random fragments of human genes located after first exons to ensure biological relevance

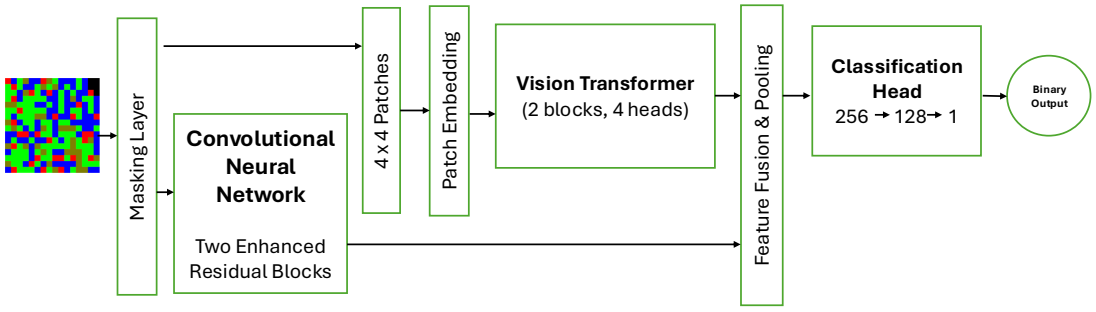

FIGURE 1. Proposed Pipeline

in the model's discrimination task. Originally adapted from Umarov and Solovyev's 2017 research, this dataset is particularly valuable for studying promoter regions that lack the canonical TATA-box motif, which constitute the majority of human gene promoters and present unique challenges for computational identification due to their diverse sequence characteristics [17].

4.2. Implementation Detail

The proposed hybrid CNN–Vision Transformer model is implemented using TensorFlow 2.17 with Keras [18]. The model accepts multichannel $16 \times 16 \times 5$ Hilbert-encoded DNA images as input. All experiments were conducted on a Linux workstation using a single GPU. [1]

To ensure that the model learns only from biologically relevant inputs, a custom `MaskingLayer` is applied at the input level. This layer masks out the 5 empty cells resulting from encoding 251-bp sequences onto a 16×16 Hilbert grid. The CNN pathway consists of convolutional blocks with Batch Normalization, ReLU activations, and dropout regularization. It includes enhanced residual blocks augmented with Squeeze-and-Excitation (SE) mechanisms and a spatial attention module. This pathway is responsible for extracting local motif-level features from the spatial DNA representation.

The ViT pathway begins with a `PatchExtract` layer that divides the image into 4×4 non-overlapping patches, followed by a `PatchEmbedding` layer which projects patches into a 64-dimensional embedding space. Two stacked `TransformerBlock` layers are used, each consisting of multi-head self-attention (MHSA), layer normalization, dropout, and GELU-activated MLP blocks. A positional embedding mechanism is integrated to maintain spatial structure awareness across the patches.

Input & Masking	16×16×5 Hilbert-encoded DNA "images," with a custom MaskingLayer zeroing out fixed regions
CNN Backbone (Residual + SE + Spatial Attention)	Two enhanced residual blocks (Conv→BN→ReLU→Conv→BN + Squeeze-Excitation + spatial-attention + skip connections + Dropout).
Patch Extraction & Embedding	Split masked input into 4×4 patches, project each to a 64-dim vector + positional encoding.
Transformer Encoder	Two stacked TransformerBlocks (LayerNorm→Multi-Head Self-Attention→MLP→residuals + Dropout + LayerNorm).
Feature Fusion & Pooling	Global average-pool the CNN map (128-dim) and transformer sequence (64-dim), then concatenate into a 192-dim feature.
Classification Head & Output	MLP (Dense→ReLU→Dropout→Dense→ReLU→Dropou) → final Dense(1, sigmoid) producing the binary-class probability.

FIGURE 2. Detailed Implementation Architecture

The outputs from both pathways are processed by global pooling layers (1D for ViT and 2D for CNN), concatenated, and passed through a classification head comprising dense layers with dropout and L2 regularization. A final `sigmoid` activation is used for binary promoter classification.

The model has early stopping of 30 epochs, with a batch size of 64 using the Adam optimizer and a fixed learning rate of 1×10^{-5}. The binary cross-entropy loss is used along with label smoothing to improve generalization. Class imbalance is addressed using dynamically computed class weights based on training label frequencies. To prevent overfitting and adaptively manage the learning rate, the following callbacks are employed:

- **Early Stopping:** Monitors validation AUC with a patience of 15 epochs.

[1] A GitHub repository at Vision for Genomic Data has been developed, and access can be granted upon request.

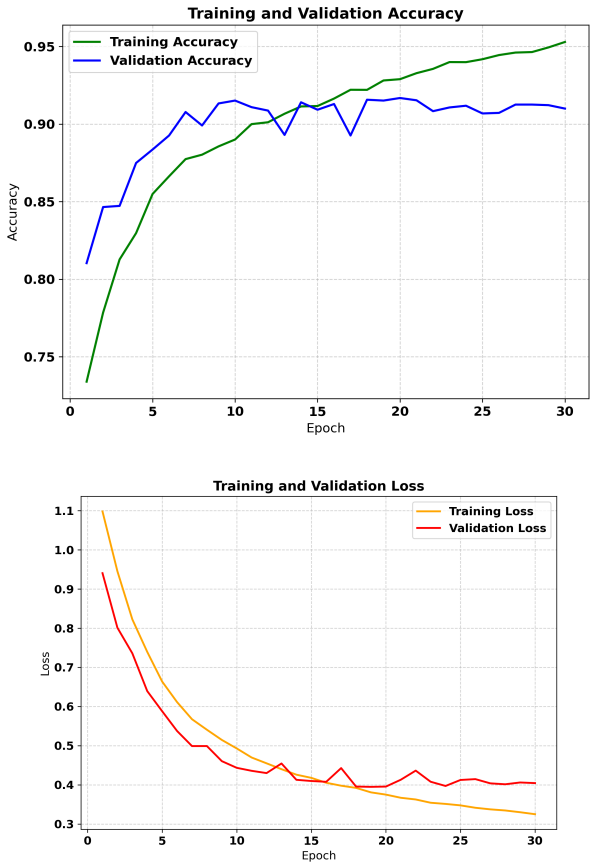

FIGURE 3. Training and validation accuracy and loss for the proposed model

- **ReduceLROnPlateau:** Reduces the learning rate by a factor of 0.5 when validation loss plateaus, with a minimum threshold of 1×10^{-7}.
- **Model Checkpointing:** Saves the best model based on validation AUC.
- **TensorBoard Logging:** Logs training metrics and weight histograms for visualization.

4.3. Results and Discussions

The performance of the proposed hybrid ViT + CNN model was evaluated on a dataset of 9,034 human DNA sequences (4,915 promoters and 4,119 non-promoters). The model trained for 30 epochs before the early stopping was triggered. Adam optimizer was used with a learning rate scheduler, and convergence was monitored via validation loss.

Figure 3 demonstrates that the model steadily improved its performance, with training accuracy reaching approximately 95.5% and validation accuracy stabilizing around 91.5%. The loss curves also show effective convergence, with both training and validation loss decreasing significantly over time, indicating successful learning and generalization.

The confusion matrix in Figure 4 shows that the model correctly classified 3,584 non-promoter sequences and 4,572 promoter sequences, with relatively few misclassifications (535 and 343, respectively). The overall test classification accuracy was 90.28%. As detailed in Table 1, and according to the classification metrics, the model achieved strong performance across both classes, as well.

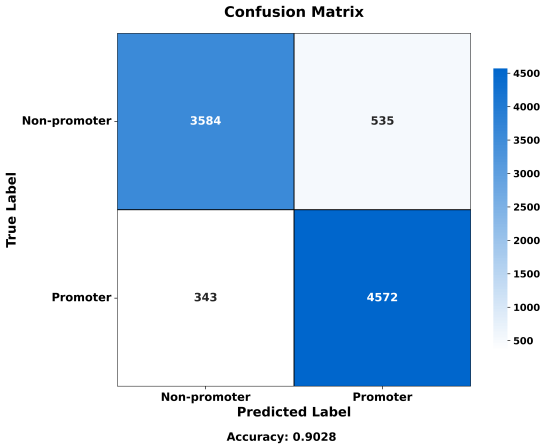

FIGURE 4. Confusion matrix of the proposed model

Class	Precision	Recall	F1-score	Support
Non-promoter	0.9127	0.8701	0.8909	4119
Promoter	0.8952	0.9302	0.9124	4915
Accuracy		0.9028		
Macro avg	0.9039	0.9002	0.9016	9034
Weighted avg	0.9032	0.9028	0.9026	9034

TABLE 1. Classification Metrics

These results suggest that the hybrid ViT-CNN architecture is effective in learning both local and global patterns within genomic sequences. Its ability to distinguish promoter from non-promoter sequences highlights its potential utility in broader sequence classification tasks in computational biology.

5. Conclusions and Future Work

This study presents a hybrid deep learning model combining CNNs and Vision Transformers (ViTs) to classify human non-TATA promoter sequences. Using multi-channel Hilbert Curve encoding, DNA sequences are transformed into 2D image-like representations that preserve both local motifs and long-range dependencies. The model achieved 95.5% training, 91.5% validation accuracy, and 90.28% validation accuracy, demonstrating strong generalization. Future work will explore higher-order Hilbert encodings and biologically enriched channels to improve detection of subtle regulatory motifs, advancing vision-based approaches in genomics.

References

[1] M. Deshpande and G. Karypis, "Evaluation of techniques for classifying biological sequences," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2002, pp. 417–431.

[2] Y. He, Z. Shen, Q. Zhang, S. Wang, and D.-S. Huang, "A survey on deep learning in dna/rna motif mining," *Briefings in Bioinformatics*, vol. 22, no. 4, p. bbaa229, 2021.

[3] S. R. Choi and M. Lee, "Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review," *Biology*, vol. 12, no. 7, p. 1033, 2023.

[4] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in bioinformatics*, vol. 18, no. 5, pp. 851–869, 2017.

[5] R. R. Nejad and S. Hooshmand, "Hvit4lung: Hybrid vision transformers augmented by transfer learning to enhance lung cancer diagnosis," in *2023 5th International Conference on Bio-engineering for Smart Technologies (BioSMART)*. IEEE, 2023, pp. 1–7.

[6] H. Gunasekaran, K. Ramalakshmi, A. Rex Macedo Arokiaraj, S. Deepa Kanmani, C. Venkatesan, and C. Suresh Gnana Dhas, "Analysis of dna sequence classification using cnn and hybrid models," *Computational and Mathematical Methods in Medicine*, vol. 2021, no. 1, p. 1835056, 2021.

[7] A. Kaur, A. P. S. Chauhan, and A. K. Aggarwal, "Prediction of enhancers in dna sequence data using a hybrid cnn-dlstm model," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 20, no. 2, pp. 1327–1336, 2022.

[8] Y. Zhang, W. Bao, Y. Cao, H. Cong, B. Chen, and Y. Chen, "A survey on protein–dna-binding sites in computational biology," *Briefings in functional genomics*, vol. 21, no. 5, pp. 357–375, 2022.

[9] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, "Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome," *Bioinformatics*, vol. 37, no. 15, pp. 2112–2120, 2021.

[10] Ž. Avsec, V. Agarwal, D. Visentin, J. R. Ledsam, A. Grabska-Barwinska, K. R. Taylor, Y. Assael, J. Jumper, P. Kohli, and D. R. Kelley, "Effective gene expression prediction from sequence by integrating long-range interactions," *Nature methods*, vol. 18, no. 10, pp. 1196–1203, 2021.

[11] H. Xu, Q. Xu, F. Cong, J. Kang, C. Han, Z. Liu, A. Madabhushi, and C. Lu, "Vision transformers for computational histopathology," *IEEE Reviews in Biomedical Engineering*, vol. 17, pp. 63–79, 2023.

[12] S. Menon, S. Piramanayakam, and G. Agarwal, "Computational identification of promoter regions in prokaryotes and eukaryotes," *EPRA International Journal of Agriculture and Rural Economic Research (ARER)*, vol. 9, no. 7, pp. 21–28, 2021.

[13] D. Chicco, "Support vector machines in bioinformatics: a survey," *Politecnico di Milano, Dipartimento di Elettronica e Informazione*, 2012.

[14] S. Anders, "Visualization of genomic data with the hilbert curve," *Bioinformatics*, vol. 25, no. 10, pp. 1231–1235, 2009.

[15] C. Ao, S. Jiao, Y. Wang, L. Yu, and Q. Zou, "Biological sequence classification: A review on data and general methods," *Research*, vol. 2022, p. 0011, 2022.

[16] K. Grešová, V. Martinek, D. Čechák, P. Šimeček, and P. Alexiou, "Genomic benchmarks: a collection of datasets for genomic sequence classification," *BMC Genomic Data*, vol. 24, no. 1, p. 25, 2023.

[17] K. Zaytsev, A. Fedorov, and E. Korotkov, "Classification of promoter sequences from human genome," *International Journal of Molecular Sciences*, vol. 24, no. 16, p. 12561, 2023.

[18] T. Authors, "Tensorflow: Large-scale machine learning on heterogeneous systems," https://www.tensorflow.org/, 2024, version 2.17 with Keras.