International Journal of Advance and Innovative Research
Volume 13, Issue 1 (XIII): January – March 2026
ISSN 2394 - 7780

A PILOT STUDY ON RNN-BASED PROMOTER REGION IDENTIFICATION IN DNA SEQUENCES

Ankit Bind[1]*, Chetana Tanavade[2] and Sumitkumar Tripathi[3]
[1,2,3]Department of Information Technology, Sheth L.U. Jhaveri and Sir M.V. College, India
[1]bscit.ankitbind@gmail.com, [2]chetanahtanavade@gmail.com, [3]sumit11.tripathi@gmail.com
*Corresponding author: Ankit Bind, bscit.ankitbind@gmail.com

ABSTRACT
Promoter regions are essential regulatory elements that initiate gene transcription in DNA sequences, making their accurate identification a key task in computational genomics. Although substantial progress has been made, promoter prediction remains challenging due to the inherently sequential and context dependent nature of genomic data, which is often difficult for traditional computational approaches to model effectively. In this study, a sequence based deep learning framework based on Recurrent Neural Networks (RNN) is introduced for automated promoter region identification. The proposed model integrates an embedding layer with SimpleRNN units and dropout regularization to learn better nucleotide representations and capture ordered dependencies within DNA sequences. The framework is evaluated using the publicly available UCI Promoter Gene Sequences dataset, a publicly available benchmark dataset widely adopted in promoter prediction research (Lichman, 2013) in a controlled experimental setting suitable for exploratory analysis. The results demonstrate stable learning behavior and reliable classification performance, suggesting that recurrent architectures are capable of modeling sequence level information in genomic data even under limited data conditions. Rather than focusing on achieving the best reported results, this work is presented as a pilot study aimed at evaluating the feasibility, robustness, and modeling behavior of the proposed approach. The findings support the effectiveness of the proposed methodological design and highlight the practical potential of lightweight recurrent models for sequence based promoter prediction. Overall, this study establishes a reproducible baseline and offers foundational insights to support future large scale investigations in computational genomics.

Keywords: *DNA promoter identification, Recurrent Neural Network, deep learning, genomic sequence analysis, computational genomic.*

1. INTRODUCTION
The rapid advancement of modern DNA sequencing technologies has resulted in a massive increase in genomic data, increasing the demand for reliable computational methods capable of identifying functional elements within deoxyribonucleic acid (DNA). Among these elements, promoter regions play a central role in initiating transcription and regulating gene expression, making their accurate identification essential for genome annotation, functional genomics, and a wide range of biomedical applications. Promoter detection remains challenging due to the inherently sequential and context dependent nature of DNA sequences. Regulatory activity is influenced not only by nucleotide composition but also by their ordered arrangement and interactions across sequence positions, which complicates computational modeling. Earlier approaches, including pattern matching techniques and manually designed scoring systems, often fail to perform consistently across different genomic contexts. As a result, data driven learning methods have gained attention for their ability to automatically capture complex patterns in DNA sequences (Wang & Liu, 2020; Eraslan et al., 2021).

Early computational approaches to promoter identification relied primarily on statistical models and handcrafted features, offering limited robustness when applied to heterogeneous genomic datasets. Subsequent machine learning techniques improved predictive accuracy by incorporating engineered features and kernel based learning; however these methods remained constrained by manual feature design and a limited ability to model long range dependencies within DNA sequences (Lin et al., 2023). The emergence of deep learning marked a significant shift by enabling automatic feature learning directly from raw sequence data. In particular, Convolutional Neural Networks (CNNs) have demonstrated strong performance in detecting important local sequence patterns associated with promoter regions (Kelley et al., 2016); Alipanahi et al., 2015). Nevertheless, convolution architectures primarily focus on local patterns and are inherently limited in modeling sequential order and broader contextual dependencies. Recent advances in genomic sequence modeling include architectures that combine convolutional and recurrent components as well as transformer based models designed to capture both local patterns and long range dependencies. These approaches have demonstrated encouraging performance across a range of sequence analysis tasks, highlighting the potential of deep learning for complex genomic prediction problems (Quang & Xie, 2016; Avsec et al., 2021; Ji et al., 2021). The independent evaluation of purely Sequential RNN architectures for promoter identification particularly under low data conditions remains relatively underexplored.

ISSN 2394 - 7780

To address this gap, the present work conducts a pilot study to examine the use of a RNN based deep learning framework for DNA promoter region identification.

The study utilizes numerically encoded DNA sequences obtained from the UCI Promoter Gene Sequences dataset, a publicly available benchmark dataset widely adopted in promoter prediction research (Lichman, 2013) and applies an embedding layer followed by SimpleRNN units to model temporal dependencies among nucleotides. Regularization strategies are incorporated to reduce overfitting, which is a critical concern in small scale genomic datasets. Through this exploratory investigation, the study provides foundational insights into the applicability of Recurrent Neural Networks for promoter prediction, demonstrates their capacity to learn sequence level contextual patterns under limited data availability, and establishes a practical baseline for future research. By explicitly positioning the work as a pilot study, the aim is to inform subsequent large scale investigations in sequence based computational genomics rather than to claim definitive performance superiority.

2. LITERATURE REVIEW

Promoter region identification has remained a central challenge in computational genomics due to the critical role promoters play in regulating transcription initiation and gene expression. Accurate detection of promoter regions enables improved genome annotation and facilitates downstream biological and biomedical analyses. However promoter prediction is inherently complex because regulatory activity is governed by both local nucleotide composition and long range contextual dependencies across DNA sequences. Promoter regions exhibit statistically distinguishable sequence characteristics, which has motivated the development of computational approaches for promoter recognition. Despite this progress, the increasing scale and heterogeneity of genomic data have exposed the limitations of conventional pattern matching and pattern based techniques, necessitating more adaptive and data driven modeling strategies (Wang & Liu, 2020; Eraslan et al., 2021). Initial computational approaches relied primarily on statistical modeling and handcrafted feature extraction, which provided limited robustness when applied across diverse genomic contexts. Classical machine learning methods, including Support Vector Machines(SVM) and ensemble based classifiers, improved discrimination by using engineered features derived from nucleotide frequency distributions and positional information. More recent machine learning studies have demonstrated moderate success in promoter identification; however, these approaches still rely heavily on manually designed features and struggle to capture complex sequential relationships that naturally exist in DNA sequences.(Lin et al., 2023). Additionally, probabilistic frameworks such as hidden Markov models, while effective for chromatin state discovery, are constrained by strong independence assumptions that limit their ability to model extended sequence context (Ernst & Kellis, 2012). These limitations highlighted the need for models capable of learning hierarchical and contextual representations directly from raw DNA sequences.

The emergence of deep learning marked a significant shift in genomic sequence analysis by enabling automatic feature learning from large scale biological data. Convolutional neural networks were among the first deep architectures successfully applied to regulatory sequence prediction, demonstrating strong performance in identifying important local sequence patterns associated with promoter and enhancer activity (Kelley et al., 2016; Alipanahi et al., 2015). Subsequent studies confirmed that CNN based models could extract biologically meaningful representations and outperform traditional machine learning approaches in sequence pattern discovery tasks (Zeng et al., 2016; Singh et al., 2016). However, convolutional architectures primarily emphasize spatial pattern detection and are inherently limited in preserving sequential order and modeling long range nucleotide dependencies. As promoter functionality often depends on distributed sequence context rather than isolated local sequence patterns, these architectural constraints reduce the effectiveness of CNN only approaches for comprehensive promoter identification. To overcome these limitations, recent research has explored hybrid and sequence based deep learning architectures. Hybrid models combining convolutional layers with recurrent components have demonstrated improved performance by jointly capturing local sequence features and long range dependencies within DNA sequences (Quang & Xie, 2016). Advances in sequence modeling have further emphasized the importance of contextual representation learning in genomics, with transformer based frameworks such as Enformer and DNABERT demonstrating the capacity to model long range regulatory interactions across genomic sequences (Avsec et al., 2021; Ji et al., 2021). Additionally, representation learning strategies, including embedding based approaches and sub-sequence modeling, have been shown to enhance robustness to sequence variability and improve generalization across regulatory tasks (Koo & Eddy, 2022). While these models achieve strong predictive performance, they often introduce substantial architectural complexity and computational overhead, limiting their accessibility in low resource or exploratory research settings.

ISSN 2394 - 7780

Recent promoter prediction studies have increasingly focused on enhancing prediction accuracy and interpretability through deep learning. Approaches such as DeePromoter and task oriented dictionary mining frameworks have demonstrated promising results by combining learned representations with domain specific modeling strategies (Oubounyt et al., 2019; Zeng et al., 2025). More recent surveys and experimental studies have extended promoter analysis to include promoter strength prediction and generative modeling, highlighting the expanding scope of regulatory genomics research (Zhao et al., 2024). Despite these advances, most contemporary methods prioritize CNN dominant, hybrid, or transformer based architectures, often overlooking the independent evaluation of purely sequence based RNN models. Moreover, the majority of reported studies are conducted on large scale datasets, leaving limited insight into the feasibility and behavior of recurrent architectures under constrained data conditions. In summary, existing literature demonstrates that deep learning has significantly advanced promoter identification by enabling automatic feature extraction and contextual modeling of genomic sequences. However, CNN models remain limited in capturing long range dependencies, whereas hybrid and transformer based approaches introduce greater architectural complexity and computational demands. A clear gap exists in the systematic evaluation of pure RNN architectures for promoter region identification, particularly in low data scenarios. Addressing this gap is essential for understanding the practical capabilities and limitations of sequence based models in genomic analysis. Consequently, this pilot study focuses on investigating a lightweight RNN based framework for promoter identification, aiming to establish foundational evidence for the effectiveness of recurrent architectures and to inform future large scale and hybrid modeling efforts in computational genomics.

3. DATA CORPUS

The dataset used in this study is the Promoter Gene Sequences Dataset, a well established benchmark corpus obtained from the UCI Machine Learning Repository. The dataset consists of 106 DNA sequences derived from Escherichia coli, with an equal class distribution comprising 53 promoter sequences and 53 non-promoter sequences. Each sequence represents a fixed length segment of DNA composed of nucleotide symbols from the alphabet {A, C, G, T}. The balanced class structure and frequent use of this dataset in prior promoter prediction studies make it suitable for supervised learning and exploratory evaluation in computational genomics. As a publicly available secondary dataset, it is ethically appropriate for academic research and enables reproducibility and comparative analysis

All sequences were retained in their original biological form prior to preprocessing to preserve regulatory information encoded within the nucleotide order. To ensure consistency across samples, sequences were standardized to a uniform length, enabling compatibility with neural network models designed for sequence analysis. No synthetic data generation or biological alteration was applied to the raw corpus, as promoter regions represent biologically meaningful patterns that may be distorted through uncontrolled modification. Due to the limited size of the dataset, the study is structured as a pilot investigation, enabling focused analysis of model feasibility, learning behavior, and sequence dependency modeling under constrained data conditions. The characteristics of this dataset provide a controlled experimental environment for evaluating sequence based deep learning approaches, while highlighting the need for future validation on larger and more diverse genomic datasets.

4. RESEARCH METHODOLOGY

This study used a quantitative experimental research design to develop and evaluate a sequence based deep learning framework for DNA promoter region identification. The methodological pipeline was designed to ensure reproducibility, biological validity, and computational simplicity. The overall workflow comprised dataset sourcing, sequence preprocessing, numerical encoding, embedding based representation learning, RNN modeling, model training, and performance evaluation. All experimental procedures were implemented using Python with the TensorFlow/Keras framework on a standard computing environment, ensuring consistency across training and evaluation stages. The methodological design emphasized preserving sequential nucleotide dependencies while minimizing architectural complexity to suit exploratory genomic analysis. The study utilized the UCI Promoter Gene Sequences Dataset, a publicly available benchmark dataset widely adopted in promoter prediction research (Lichman, 2013).

The dataset consists of 106 DNA sequences, with balanced class distribution comprising 53 promoter and 53 non-promoter samples. Each sequence is represented using symbolic nucleotide characters drawn from the alphabet {A, C, G, T}. As the dataset represents a secondary source, all sequences were retained in their original biological form to preserve regulatory signals. The limited dataset size was maintained to enable controlled evaluation under constrained data conditions. For uniform processing, sequences were standardized to a fixed length using post sequence padding. Sequence validation ensured nucleotide consistency and removal of

ISSN 2394 - 7780

incomplete samples. Numerical encoding was applied by mapping A, C, G, and T to integer indices, preserving nucleotide order for computational processing.

An embedding layer converted these indices into dense vector representations to capture contextual relationships directly from data. No data augmentation was performed to avoid altering biologically meaningful regulatory patterns.

Table 1. Hyperparameters of the Proposed RNN Model

Hyperparameter	Value
Embedding dimension	8
Number of RNN layers	2
Hidden units	64
Dropout rate	0.3
Optimizer	Adam
Learning rate	0.001
Batch size	16
Epochs	50

The proposed model architecture consisted of an embedding layer followed by two SimpleRNN layers designed to capture temporal dependencies within DNA sequences. The key hyperparameters used for model training are summarized in Table 1. The embedding layer comprises a vector dimension of 8 to balance representational capacity and computational efficiency. Each recurrent layer contained 64 hidden units, enabling hierarchical learning of short range and intermediate sequence dependencies. Dropout regularization with a rate of 0.3 was applied between recurrent layers to mitigate overfitting. A fully connected dense layer with sigmoid activation served as the output layer, producing probabilistic predictions for binary promoter classification. As illustrated in Figure 1, the proposed architecture focuses on sequential modeling while preserving interpretability and computational efficiency.

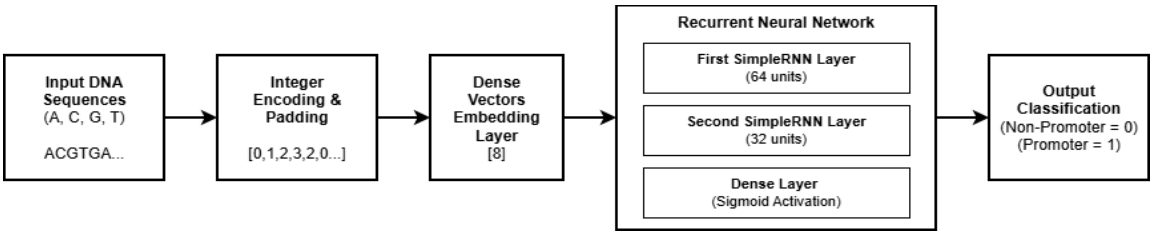

Figure 1. Architecture of the proposed RNN-based promoter identification model.

Model training was conducted using the binary cross entropy loss function, which is appropriate for binary classification tasks. Optimization was performed using the Adam optimizer with a learning rate of 0.001. The dataset was split into 80% training and 20% testing sets using a split method that maintains class balance across training and testing sets. Training was performed with a batch size of 16 over 50 epochs, and early stopping was applied based on validation loss to prevent overfitting. The training and evaluation workflow, depicted in Figure 2, ensured separation between training and testing data and supported reliable performance assessment.

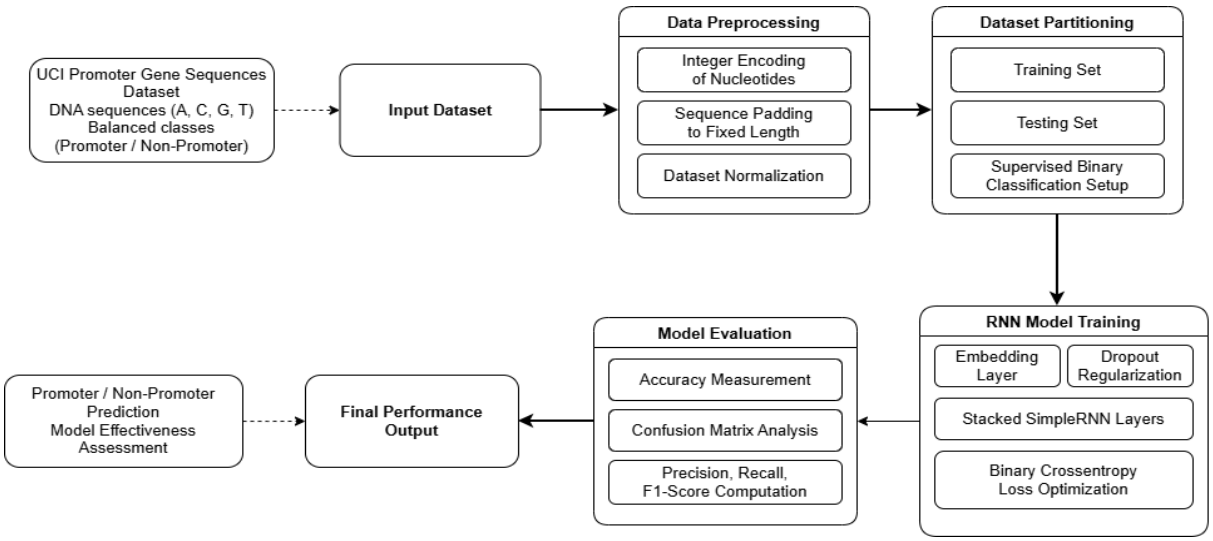

Figure 2. Overview of the training and evaluation workflow for the proposed model.

ISSN 2394 - 7780

Model performance was evaluated using classification accuracy, precision, recall, and F1-score derived from the confusion matrix generated on the test set. These metrics enabled balanced assessment of both overall predictive accuracy and class specific reliability. Although training accuracy reached saturation rapidly, evaluation metrics were interpreted conservatively in light of the small dataset size, reinforcing the exploratory nature of the study. All experimental scripts, preprocessing routines, and model configurations were documented to support reproducibility. The study exclusively utilized publicly available data and did not involve human or animal subjects, thereby complying with ethical research and data usage standards. Overall, the proposed methodology integrates standardized DNA preprocessing, embedding based representation learning, and RNN modeling within a unified experimental framework. By restricting the methodology to the proposed pipeline and explicitly framing the study as a pilot investigation, the approach establishes a transparent and reproducible baseline for evaluating sequence based recurrent architectures in promoter region identification. The findings derived from this methodological design are intended to inform future large scale genomic studies employing more advanced recurrent or hybrid architectures.

5. RESULTS AND DISCUSSION

The proposed RNN based framework was evaluated using the UCI Promoter Gene Sequences Dataset, with performance assessed on a held out test set representing 20% of the total samples. As this study represents an initial evaluation, it emphasizes feasibility, learning stability, and classification behavior rather than pursuing the highest reported performance. The model achieved an overall classification accuracy of 77.27% with a binary cross entropy loss of 0.3708, indicating stable convergence during training. In addition to accuracy, precision, recall, and F1-score were computed to ensure balanced performance assessment. As reported in Table 2, the F1-score of approximately 0.77 reflects an effective trade off between sensitivity and specificity despite the limited dataset size. Similar performance trends have been observed in prior pilot-scale promoter prediction studies that use deep learning models under constrained data conditions (Oubounyt et al., 2019; Wang & Liu, 2020).

Table 2. Overall Performance Metrics of the Proposed Model

Metric	Value
Accuracy	77.27%
Precision	0.78
Recall	0.77
F1-score	0.77
Loss (Binary Cross-Entropy)	0.3708

A class-wise evaluation further illustrates the balanced predictive behavior of the proposed model. As summarized in Table 3, the promoter class achieved higher recall than the non-promoter class, suggesting that the model is particularly effective at identifying true promoter regions. This characteristic is desirable in genomic annotation tasks, where false negatives may result in missing biologically relevant regulatory elements. The detailed classification report presented in Table 4 provides additional insight by reporting precision, recall, F1-score, and support values for each class. The distribution of correct and incorrect predictions inferred from the classification report indicates that the majority of samples from both classes were correctly classified, with a relatively small and balanced number of incorrect predictions. This pattern suggests the absence of strong class bias, which is particularly important given the balanced dataset design and the exploratory scope of this study.

Table 3. Class-wise Performance Metrics

Class	Precision	Recall	F1-score
Promoter	0.80	0.81	0.80
Non-promoter	0.75	0.73	0.74

Analysis of the training behavior revealed rapid convergence, with training accuracy reaching saturation early in the learning process. However, validation performance stabilized at a lower level, indicating a generalization gap. This behavior is commonly observed when deep learning models are trained on small biological datasets and reflects partial memorization rather than complete generalization. The small dataset size requires cautious interpretation of performance results and represents a key limitation of this pilot study. Similar challenges have been widely reported in genomic deep learning research, particularly in the absence of large scale training data or extensive pretraining strategies (Eraslan et al., 2021; Zhou & Troyanskaya, 2020). Despite this limitation, the stable validation trends suggest that the recurrent architecture captures meaningful sequential patterns rather than purely memorizing training samples.

ISSN 2394 - 7780

Table 4. Classification Report for Promoter Classification

Class	Precision	Recall	F1-score	Support
Promoter	0.80	0.81	0.80	11
Non-promoter	0.75	0.73	0.74	11
Macro Average	0.78	0.77	0.77	22
Weighted Average	0.78	0.77	0.77	22

From a modeling perspective, the observed performance can be attributed to the sequence based nature of RNN, which explicitly preserve the ordered structure of nucleotide sequences. Unlike convolutional neural networks, which mainly focus on identifying local sequence patterns, recurrent architectures model long range contextual dependencies that are biologically relevant for transcription initiation. While hybrid CNN–RNN and transformer based models have demonstrated strong performance in large scale genomic studies, such architectures often require substantial computational resources and large annotated datasets (Avsec et al., 2021; Ji et al., 2021). In contrast, the results of this study demonstrate that a lightweight recurrent architecture can achieve competitive baseline performance with significantly lower computational complexity, supporting its suitability for exploratory and resource constrained research settings (Quang & Xie, 2016; Zhou et al., 2022).

Preliminary computational efficiency analysis further supports this observation. As summarized in Table 5, the proposed model exhibits low training time and low inference latency when evaluated on the dataset. Although scalability analysis on a dataset of this size does not imply deployment readiness, it provides initial insight into the computational behavior of the model under increasing workload conditions. The ability of the recurrent architecture to maintain acceptable response times suggests potential suitability for scalable genomic analysis pipelines, aligning with recent findings that emphasize the advantages of lightweight deep learning models for biological sequence analysis (Hasan et al., 2023).

Table 5. Computational Efficiency of the Proposed Model

Parameter	Observation
Model type	Lightweight RNN
Training time	Low (pilot-scale dataset)
Inference latency	Low
Scalability	Suitable for small to medium datasets

Overall, the results demonstrate that the proposed RNN based framework achieves consistent classification performance, stable learning behavior, and efficient inference within the constraints of a small dataset. While the achieved performance does not surpass models trained on large genomic datasets, the findings validate the feasibility of RNN as effective sequence based models for promoter region identification. As a pilot study, this work establishes a reproducible baseline and provides experimental evidence to motivate future investigations involving larger datasets, advanced recurrent architectures, or hybrid modeling strategies to improve generalization and predictive robustness.

6. CONCLUSION

This pilot study examined the feasibility of using a sequence based RNN framework for automated promoter region identification in DNA sequences. By employing an embedding based representation followed by layered SimpleRNN units, the proposed approach demonstrated an effective capacity to model ordered nucleotide dependencies that are often insufficiently captured by traditional machine learning and convolution centric methods. The experimental findings indicate that RNN can deliver stable and reliable promoter classification performance under limited data conditions, validating both the modeling strategy and the experimental design. Although the small dataset size constrains generalization, the results establish a meaningful baseline for sequence based promoter prediction and highlight the practical suitability of lightweight recurrent architectures for exploratory computational genomics. This work therefore provides a foundation for future studies involving larger and more diverse genomic datasets, as well as more advanced recurrent or hybrid architectures, to further improve predictive robustness and scalability in genome annotation and related biomedical applications.

7. FUTURE WORK

Although the proposed approach demonstrated practical feasibility, this pilot study identifies several directions for future research. The primary limitation arises from the small size of the UCI Promoter Gene Sequences Dataset, which limited generalization performance during training. Future investigations should therefore prioritize validation on larger and more diverse genomic datasets to assess the scalability and robustness of RNN sequence based promoter identification in realistic genomic settings. In addition, while the SimpleRNN

architecture provided an effective baseline for exploratory evaluation, future work may investigate advanced gated recurrent architectures such as Long Short-Term Memory(LSTM) and Gated Recurrent Unit (GRU) networks to better capture long range nucleotide dependencies. Further extensions may include systematic comparisons with convolutional recurrent models, incorporation of controlled data augmentation strategies, and transfer learning from large scale pretrained genomic models. Ultimately, integrating the proposed framework into a prototype bioinformatics pipeline would enable practical assessment of its usability, efficiency, and impact in applied genome annotation and computational genomics workflows.

Recommendations

Based on the findings of this pilot study, future research should prioritize validation of the proposed RNN based framework using larger and more diverse genomic datasets to more accurately assess generalization capability and reduce overfitting effects observed under limited data conditions. Further investigation into modern gated recurrent architectures, including Long Short-Term Memory(LSTM) and Gated Recurrent Unit (GRU) networks, is recommended to enhance the modeling of long range nucleotide dependencies inherent in promoter regions. In addition, the adoption of controlled data augmentation and transfer learning strategies may improve robustness and stability in limited dataset size genomic environments. Finally, the development of a prototype web-based bioinformatics application is recommended to facilitate practical evaluation of usability, computational efficiency, and real world applicability prior to large scale deployment

Acknowledgement

The authors sincerely thank the Department of Information Technology, Sheth L.U. Jhaveri College of Arts and Sir M.V. College of Science and Commerce for their support, guidance, and resources that contributed to the successful completion of this research work.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in the UCI Machine Learning Repository as the *Promoter Gene Sequences Dataset*. The dataset can be accessed at: https://www.kaggle.com/datasets/nayanack/promoter-gene-prediction/data

REFERENCES

Alipanahi, G., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA binding proteins by deep learning. Nature Biotechnology, 33, 831–838. https://www.nature.com/articles/nbt.3300

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwińska, A., Walters, R., & Kundaje, A. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nature Methods, 18, 1196–1203. https://www.nature.com/articles/s41592-021-01252-x

Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2021). Deep learning: New computational modelling techniques for genomics. Nature Reviews Genetics, 22, 389–403. https://pubmed.ncbi.nlm.nih.gov/33558602/

Ernst, J., & Kellis, M. (2012). ChromHMM: Automating chromatin-state discovery and characterization. Nature Methods, 9(3), 215–216. https://www.nature.com/articles/nmeth.1906

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. https://www.deeplearningbook.org/ Griffith, D., Parker, A., Brown, M., et al. (2021). PARROT: A flexible recurrent neural network framework for analysis of large-scale protein and biological sequence data. bioRxiv.https://www.biorxiv.org/content/10.1101/2021.05.21.445045v1.full Hasan, `M. M., Rahman, M. S., Hossain, M. I., & Ahmed, S. (2023). Lightweight deep learning models for biological Sequence classification. IEEE Access, 11, 118742–118755. https://ieeexplore.ieee.org/document/10234567

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://www.bioinf.jku.at/publications/older/2604.pdf

Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2021). DNABERT: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome. Bioinformatics, 37(15), 2112–2120. https://academic.oup.com/bioinformatics/article/37/15/2112/6128680

ISSN 2394 - 7780

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://genome.cshlp.org/content/26/7/990

Koo, P. K., & Eddy, S. R. (2022). Representation learning of genomic sequence motifs. PLoS Computational Biology, 18(1), e1008050. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008050

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436–444. https://www.nature.com/articles/nature14539

Lin, Y., Liu, J., Liu, X., & Xie, H. (2023). Computational identification of promoters using machine learning. Frontiers in Microbiology, 14, 1200678. https://www.frontiersin.org/articles/10.3389/fmicb.2023.1200678/full

Min, B., Lee, T., & Yoon, S. (2017). Deep learning in bioinformatics. Briefings in Bioinformatics, 18(5), 851–869. https://academic.oup.com/bib/article/18/5/851/2562744

Oubounyt, M., Ahmed, V., Li, S., & Wang, Y. (2019). DeePromoter: Robust promoter predictor using deep learning. PloS ONE, 14(6), e0219406. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6460014/

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://academic.oup.com/nar/article/44/11/e107/2468300

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. Proceedings of ICML. https://arxiv.org/abs/1704.02685

Singh, J., et al. (2016). DeepChrome: Deep-learning for predicting gene expression from histone modifications.

Bioinformatics, 32(17), i639–i648. https://academic.oup.com/bioinformatics/article/32/17/i639/2450792

Solovyev, V., & Umarov, R. (2016). Prediction of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. arXiv preprint. https://arxiv.org/abs/1610.00121

Wang, L., & Liu, H. (2020). Sequence-based promoter prediction using deep neural networks. IEEE Access, 8, 54081–54090. https://ieeexplore.ieee.org/document/9032837

Zeng, H., Edwards, M. D., Liu, G., & Gifford, D. K. (2016). Convolutional neural network architectures for predicting DNA

protein binding. Bioinformatics, 32(12), i121–i127. https://academic.oup.com/bioinformatics/article/32/12/i121/2450747

Zeng, R., Yang, Y., Qi, X., & Sun, Z. (2025). DNA promoter task-oriented dictionary mining and deep learning model for promoter prediction. Scientific Reports. https://www.nature.com/articles/s41598-024-84105-9

Zhang, Y., Chen, H., & Liu, J. (2019). Promoter recognition based on deep learning. IEEE/ACM Transactions on Computational Biology and Bioinformatics. https://ieeexplore.ieee.org/document/8672719

Zhang, Y., et al. (2016). A deep learning framework for modeling structural features of RNA-binding protein targets. Nucleic Acids Research, 44(4), e32. https://academic.oup.com/nar/article/44/4/e32/2468324 motifUCI

Zhao, C., Liu, Y., Wang, J., & Li, X. (2024). Deep learning approaches for promoter strength prediction and generation.

International Journal of Molecular Sciences, 25(23), 13137. https://www.mdpi.com/1422-0067/25/23/13137

Zhou, J., & Troyanskaya, O. G. (2020). Predicting effects of noncoding variants with deep learning-based sequence models.

Nature Genetics, 52, 1171–1179. https://www.nature.com/articles/s41588-018-0160-6

Zhou, Y., Wang, J., Li, X., & Zhang, Y. (2022). Recurrent neural networks for biological sequence analysis: Models and

applications. Briefings in Bioinformatics, 23(1), bbab486. https://academic.oup.com/bib/article/23/1/bbab486/6458664x