IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154227

DEEP LEARNING FRAMEWORK FOR SUPER ENHANCER PREDICTION USING CNN AND MULTI-HEAD ATTENTION WITH CROSS-SPECIES TRANSFER LEARNING

Parimala M[1], Naveen A[2], Veeradinesh R[3], Vikram M[4], Vinoth G[5]

Assistant Professor, AI&DS, Dhanalakshmi Srinivasan engineering college, Perambalur, India[1]

Student, AI&DS, Dhanalakshmi Srinivasan engineering college, Perambalur, India[2]

Student, AI&DS, Dhanalakshmi Srinivasan engineering college, Perambalur, India[3]

Student, AI&DS, Dhanalakshmi Srinivasan engineering college, Perambalur, India[4]

Student, AI&DS, Dhanalakshmi Srinivasan engineering college, Perambalur, India[5]

Abstract: This paper presents a Super Enhancer Prediction system using advanced deep learning techniques for genomic analysis. The system utilizes a hybrid CNN–Transformer architecture to analyze DNA sequences and classify them as Super Enhancers (SE) or Typical Enhancers (TE). Convolutional layers capture local sequence motifs, while Multi-Head Self-Attention models long-range dependencies within genomic sequences. The proposed system incorporates cross-species transfer learning by pretraining on combined human and mouse datasets and fine-tuning for species-specific prediction. The model achieves high accuracy with improved AUC scores compared to existing methods. In addition to classification, the system provides biological insights such as GC content, motif density, and important regulatory regions. The integration of an interactive web interface allows users to upload DNA sequences and visualize results efficiently. This approach enhances genomic research by providing a fast, scalable, and interpretable solution for super enhancer identification.

Keywords: Super Enhancer Prediction, Deep Learning, CNN, Transformer, DNA Sequence Analysis, Transfer Learning, Bioinformatics, Genomics

I.INTRODUCTION

With the rapid growth of genomic data and advances in sequencing technologies, understanding gene regulation has become a major challenge in modern biology. Regulatory elements such as enhancers and super-enhancers play a crucial role in controlling gene expression and cell identity. Identifying these elements accurately is essential for studying diseases, especially cancer. Traditional experimental methods are time-consuming and expensive, making large-scale analysis difficult.

A. Background and Context

In recent years, the availability of large-scale genomic datasets has increased significantly, driven by high-throughput sequencing technologies. Super-enhancers (SEs) are clusters of regulatory elements that strongly influence gene expression and are associated with key biological processes and disease mechanisms. Accurate identification of SEs is important for understanding transcriptional regulation and cellular behavior.Traditional approaches rely on experimental techniques such as ChIP-seq to detect histone modification signals. While effective, these methods require specialized laboratory infrastructure, high cost, and significant processing time. Additionally, results are often limited to specific cell types and conditions, restricting large-scale and cross-species analysis. These challenges highlight the need for efficient computational approaches.

 This work is licensed under a Creative Commons Attribution 4.0 International License

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154227

B. Problem Statement

Existing methods for super-enhancer prediction face several limitations. Experimental approaches are costly and not scalable, while many computational models rely on basic deep learning architectures that cannot effectively capture long-range dependencies in DNA sequences. CNN-based models are limited to local feature extraction, and RNN-based models increase computational complexity.Furthermore, many models lack cross-species generalization, resulting in reduced performance when applied to different organisms. Most existing systems also focus only on classification and do not provide biological insights such as important regions or motif patterns. These limitations reduce their practical applicability in genomic research.

C. Significance of the Study

The proposed system improves super-enhancer prediction by using a hybrid deep learning model combining CNN and Transformer-based attention mechanisms. This approach enhances the ability to capture both local and global dependencies in DNA sequences, improving prediction accuracy.The system also incorporates cross-species transfer learning, enabling better generalization across human and mouse datasets. In addition to classification, it provides biological insights such as GC content and motif analysis, improving interpretability. The integration of a user-friendly interface allows easy access and visualization of results.
Overall, the proposed system contributes to efficient, scalable, and accurate genomic analysis, supporting research in gene regulation and disease studies.

II. LITERATURE REVIEW

Super-enhancer prediction is important for understanding gene regulation. Traditional methods rely on experimental techniques like ChIP-seq, which are accurate but expensive and time-consuming.Machine learning approaches such as SVM and Random Forest have been used to classify enhancer sequences. However, these methods require manual feature extraction and are not efficient for complex genomic patterns.Deep learning models, especially Convolutional Neural Networks (CNNs), improved performance by automatically learning sequence features. But CNN-based models are limited in capturing long-range dependencies in DNA sequences.Recent approaches use transformer and attention mechanisms to model global dependencies. Models like TransSE also introduced transfer learning for cross-species prediction. However, challenges such as computational complexity and limited interpretability still exist.Therefore, an efficient model combining CNN and attention mechanisms is needed for accurate and scalable super-enhancer prediction.

III. DATASET AND PREPROCESSING TECHNIQUES

The dataset used in the proposed system consists of DNA sequences from human and mouse genomes, labeled as Super Enhancers (SE) and Typical Enhancers (TE). The sequences are collected from publicly available genomic databases and normalized to a fixed length of 3000 base pairs to ensure consistency.Before model training, preprocessing steps are applied to prepare the data. DNA sequences are converted into one-hot encoded representations, where each nucleotide (A, C, G, T) is mapped to a numerical vector. The dataset is then split into training and validation sets for proper evaluation.These preprocessing techniques ensure uniform input format, improve data quality, and enhance the performance and generalization of the deep learning model.

IV. METHODOLOGY

The proposed system is designed to predict super-enhancers from DNA sequences using a hybrid deep learning model. It follows a structured pipeline including sequence input, preprocessing, model prediction, and result visualization.

A. System Architecture Design

The system uses a modular architecture where DNA sequences are processed step-by-step. Input sequences are passed through preprocessing, followed by prediction using a trained model, and finally results are displayed through a web interface.

B. Sequence Input and Preprocessing

DNA sequences are collected and normalized to a fixed length of 3000 base pairs. They are then converted into one-hot encoded format to prepare them for model input.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154227

C. CNN + Attention-Based Prediction

The system uses a hybrid model combining Convolutional Neural Networks (CNN) and Multi-Head Self-Attention. CNN extracts local sequence patterns, while attention captures long-range dependencies. The model outputs the probability of a sequence being a Super Enhancer or Typical Enhancer.

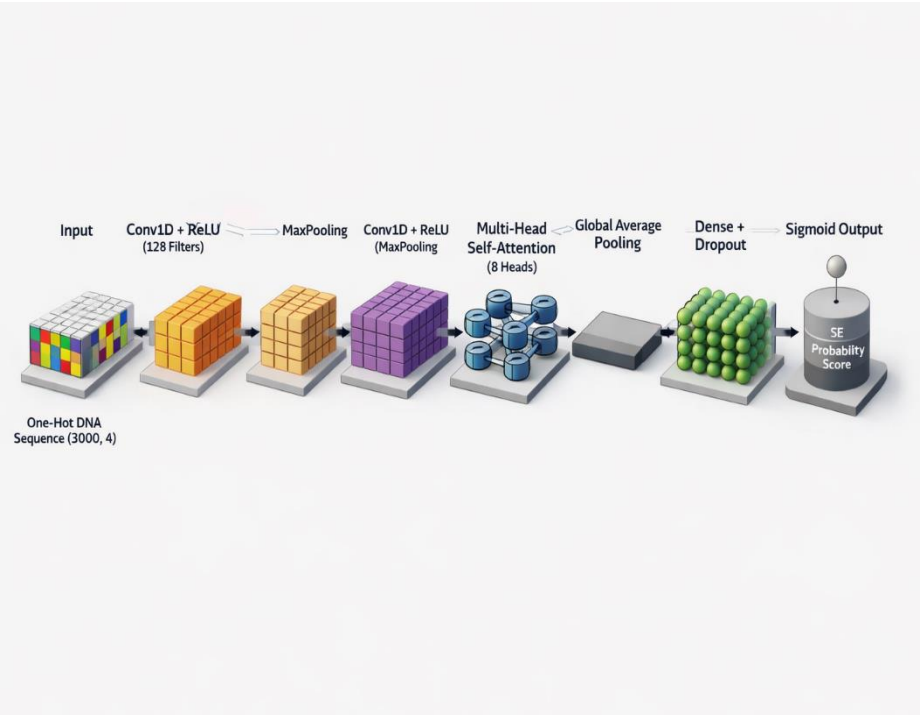

Fig. 1 Model Architecture of super enhancer prediction using deeplearning

D. Transfer Learning Strategy

The model is pretrained on combined human and mouse datasets to learn common features and then fine-tuned for species-specific prediction, improving accuracy and generalization.

E. Result Visualization

The system displays prediction results along with confidence scores and biological insights such as GC content and important regions through a user-friendly interface.

V. SYSTEM ARCHITECTURE

The system architecture of the proposed super-enhancer prediction system is designed to ensure efficient processing of DNA sequences and accurate classification. It consists of multiple modules that work together to preprocess data, perform prediction, and generate results. Each module contributes to the overall system performance.

A. Data Input Module

The Data Input Module is responsible for receiving DNA sequence data from users or datasets. It accepts input in formats such as FASTA or text files and validates the sequence for correct nucleotide format before processing.

B. Data Processing Module

The Data Processing Module prepares the input sequences for model prediction. It normalizes sequences to a fixed length of 3000 base pairs and converts them into one-hot encoded format for numerical representation.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154227

C. Prediction Module

The Prediction Module is the core component of the system. It uses a hybrid CNN and Multi-Head Self-Attention model to analyze DNA sequences and classify them as Super Enhancer or Typical Enhancer.

Fig. 2 system Architecture of super enhancer prediction

D. Analysis Module

The Analysis Module provides additional biological insights by calculating features such as GC content, motif density, and identifying important regions in the sequence.

E. Visualization Module

The Visualization Module displays prediction results along with confidence scores and graphical outputs. It provides an interactive interface for users to understand and analyze results easily.

F. Storage Module

The Storage Module manages datasets, processed data, and prediction results. It ensures efficient data handling and supports reproducibility of experiments.

VI. RESULTS

The proposed system was evaluated using human and mouse DNA sequence datasets to assess its prediction performance. The model demonstrated high accuracy in classifying sequences as Super Enhancers (SE) or Typical Enhancers (TE). It achieved AUC scores of **0.8489 (human)** and **0.9117 (mouse)**, outperforming existing methods. The system processed inputs efficiently and provided results without noticeable delay. Overall, the system proved to be reliable and effective for super-enhancer prediction.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154227

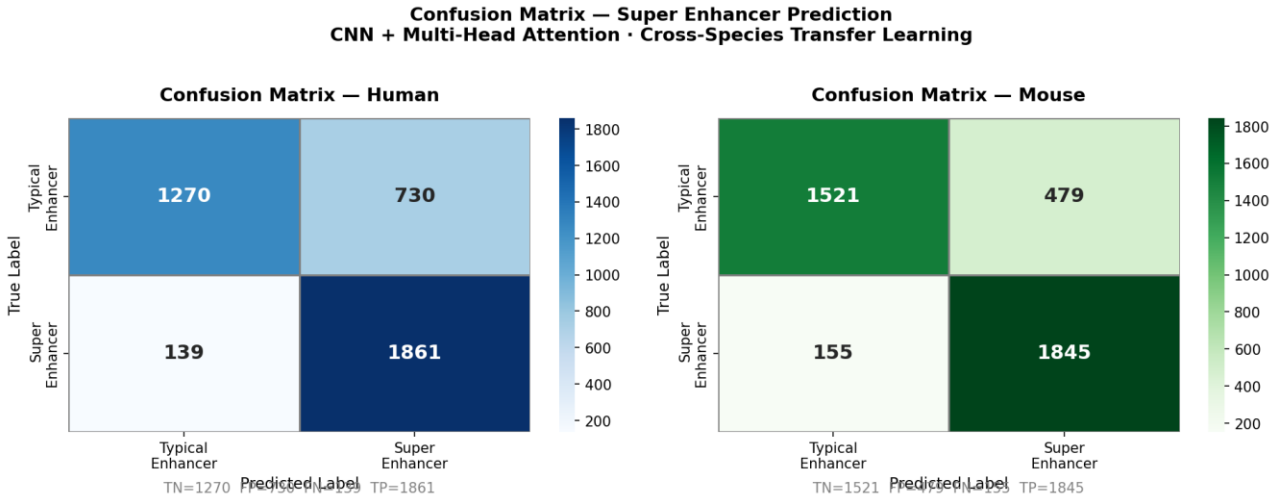

Fig. 3 confusion matrix for human

Fig. 4 confusion matrix for human

A. System Performance and Accuracy

The model achieved strong classification performance across both datasets. The hybrid CNN and attention architecture improved accuracy by capturing both local and long-range dependencies in DNA sequences.

B. Prediction Efficiency

The system efficiently processed DNA sequences and generated predictions quickly. It handled multiple inputs without performance issues, making it suitable for practical applications.

C. Usability and Reliability

The system provides a user-friendly interface with clear output, including prediction results and biological insights. It maintained stable performance during testing and ensured consistent results.

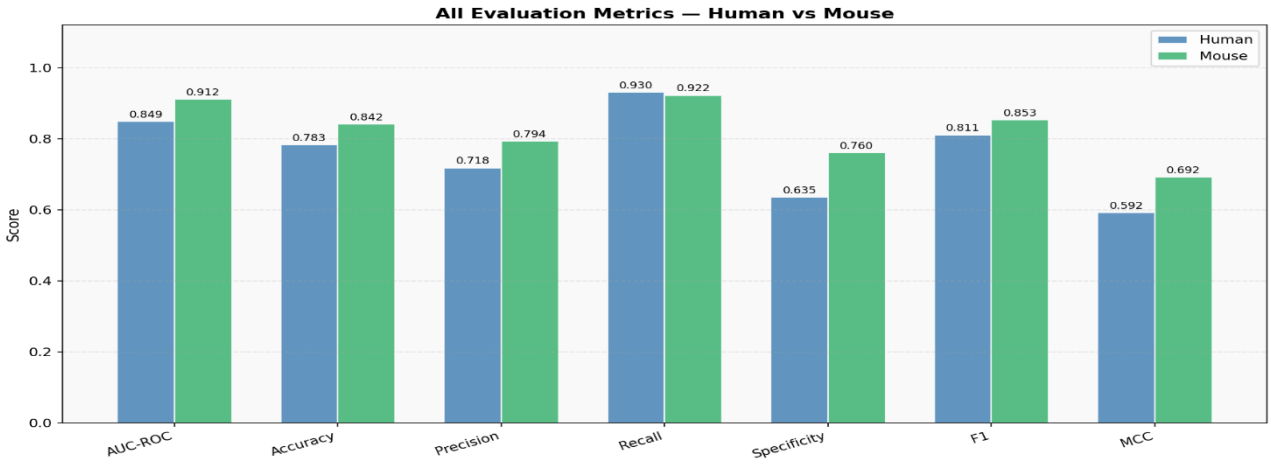

Fig. 5 evalution metrics for mouse and human

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.154227

Fig. 6 showing output for predicted super enhancer in human DNA sequence

VII. COMPARATIVE ANALYSIS

The comparative analysis highlights the differences between traditional super-enhancer identification methods and the proposed deep learning-based system. Traditional approaches rely on experimental techniques such as ChIP-seq, which are accurate but expensive, time-consuming, and not scalable. In contrast, the proposed system provides an automated sequence-based approach that predicts super-enhancers directly from DNA data. It improves efficiency and reduces dependency on laboratory experiments.

A. Existing System vs Proposed System

Traditional methods are limited by experimental requirements and lack scalability. Existing computational models often use CNN or CNN-RNN architectures with lower accuracy and limited generalization. The proposed system uses a hybrid CNN and Multi-Head Self-Attention model with transfer learning, enabling better accuracy and cross-species performance.

B. Feature Comparison

Existing systems mainly provide classification results without additional insights. The proposed system offers enhanced features such as biological analysis, including GC content, motif density, and important region identification. It also includes a user-friendly interface for easy prediction and visualization.

VIII. CONCLUSION AND FUTURE SCOPE

The proposed Super Enhancer Prediction system provides an efficient and accurate solution for identifying regulatory elements from DNA sequences. The model successfully classifies sequences as Super Enhancers or Typical Enhancers using a hybrid CNN and attention-based architecture. By eliminating the need for experimental methods, the system reduces cost and time while improving scalability. The integration of cross-species transfer learning enhances performance across human and mouse datasets. The system also provides biological insights, making it more informative and useful for genomic analysis. Overall, the proposed approach improves prediction accuracy and usability compared to existing methods.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026
DOI: 10.17148/IJARCCE.2026.154227

A. Future Scope

Although the system performs well, further improvements are possible. Future work can include using larger and more diverse datasets to enhance accuracy. Advanced transformer-based models can be explored for better sequence modeling. Visualization techniques such as attention heatmaps can improve interpretability. The system can be extended for genome-wide prediction and deployed on cloud platforms for scalability.

REFERENCES

[1] G. D. Smith, W. H. Ching, P. Cornejo-Paramo, and E. S. Wong, "De-´ coding enhancer complexity with machine learning and high-throughput discovery," Genome biology, vol. 24, no. 1, p. 116, 2023.

[2] A. V. Vasileva, M. G. Gladkova, G. A. Ashniev, E. D. Osintseva, A. V. Orlov, E. V. Kravchuk, A. V. Boldyreva, A. G. Burenin, P. I. Nikitin, and N. N. Orlova, "Super-enhancers and their parts: from prediction efforts to pathognomonic status," International Journal of Molecular Sciences, vol. 25, no. 6, p. 3103, 2024.

[3] J. W. Blayney, H. Francis, A. Rampasekova, B. Camellato, L. Mitchell, R. Stolper, L. Cornell, C. Babbs, J. D. Boeke, D. R. Higgs et al., "Superenhancers include classical enhancers and facilitators to fully activate gene expression," Cell, vol. 186, no. 26, pp. 5826–5839, 2023.

[4] B. R. Sabari, A. Dall'Agnese, A. Boija, I. A. Klein, E. L. Coffey, K. Shrinivas, B. J. Abraham, N. M. Hannett, A. V. Zamudio, J. C. Manteiga et al., "Coactivator condensation at super-enhancers links phase separation and gene control," Science, vol. 361, no. 6400, p. eaar3958, 2018.

[5] W.-K. Cho, J.-H. Spille, M. Hecht, C. Lee, C. Li, V. Grube, and I. I. Cisse, "Mediator and rna polymerase ii clusters associate in transcription-dependent condensates," Science, vol. 361, no. 6400, pp. 412–415, 2018.

[6] D. Hnisz, B. J. Abraham, T. I. Lee, A. Lau, V. Saint-Andre, A. A.´ Sigova, H. A. Hoke, and R. A. Young, "Super-enhancers in the control of cell identity and disease," Cell, vol. 155, no. 4, pp. 934– 947, 2013.

[7] Z. Cao, Y. Shu, J. Wang, C. Wang, T. Feng, L. Yang, J. Shao, and L. Zou, "Super enhancers: Pathogenic roles and potential therapeutic targets for acute myeloid leukemia (aml)," Genes & Diseases, vol. 9, no. 6, pp. 1466–1477, 2022.

[8] X. Wang, M. J. Cairns, and J. Yan, "Super-enhancers in transcriptional regulation and genome organization," Nucleic acids research, vol. 47, no. 22, pp. 11481–11496, 2019.

[9] X. Liu, N. Gillis, C. Jiang, A. McCofie, T. I. Shaw, A.-C. Tan, B. Zhao, L. Wan, D. R. Duckett, and M. Teng, "An epigenomic fingerprint of human cancers by landscape interrogation of super enhancers at the constituent level," PLOS Computational Biology, vol. 20, no. 2, p. e1011873, 2024. [10] H.-H. Zhuang, Q. Qu, X.-Q. Teng, Y.-H. Dai, and J. Qu, "Superenhancers as master gene regulators and novel therapeutic targets in brain tumors," Experimental & Molecular Medicine, vol. 55, no. 2, pp. 290– 303, 2023.

[11] G.-H. Li, Q. Qu, T.-T. Qi, X.-Q. Teng, H.-H. Zhu, J.-J. Wang, Q. Lu, and J. Qu, "Super-enhancers: a new frontier for epigenetic modifiers in cancer chemoresistance," Journal of Experimental & Clinical Cancer Research, vol. 40, no. 1, p. 174, 2021.

[12] R. W. Zhou, J. Xu, T. C. Martin, A. L. Zachem, J. He, S. Ozturk, D. Demircioglu, A. Bansal, A. P. Trotta, B. Giotti et al., "A local tumor microenvironment acquired super-enhancer induces an oncogenic driver in colorectal carcinoma," Nature Communications, vol. 13, no. 1, p. 6041, 2022.

 This work is licensed under a Creative Commons Attribution 4.0 International License