A CAPSULE NETWORK-BASED HYBRID DEEP LEARNING MODEL FOR EFFICIENT PREDICTION OF CRISPR-CAS9 OFF-TARGET EFFECTS

Reference NO. IJME 2533, DOI: 10.5750/sijme.v167iA2(S).2533

Hamsika Chakilam*, Vaels International School, Chennai, India, **Kishore Kumar T**, National Institute of Technology, Warangal, India and **Tekyam Krishna Kumar Naidu**, Prathima Institute of Medical Sciences, Karimnagar, India

*Corresponding author. Hamsika Chakilam (Email): chakilamhamsika@gmail.com

KEY DATES: Submission date: 11.09.2024; Final acceptance date: 25.03.2025; Published date: 30.04.2025

SUMMARY

CRISPR-Cas9 genome editing has transformed biomedical research and the development of therapies, yet the challenge of unintended off-target effects remains a significant obstacle to its clinical use. In this study, we present a new deep learning model that combines Capsule Networks with Transformer blocks, bidirectional LSTM layers, and CNNs. The model is further strengthened by incorporating k-mer encoded sequence features and biological rule checks to predict CRISPR-Cas9 off-target activity with greater accuracy. It processes guide and off-target DNA sequences through a hybrid pipeline, which includes convolution, temporal modelling, attention-based representation learning, and spatial hierarchy encoding using capsule layers. At the same time, the model extracts and analyses numerical features, such as mismatch counts, GC content, and PAM motif patterns. To improve reliability, we introduce a biological constraint layer that filters predictions based on well-established domain knowledge. The final predictions result from integrating these various feature representations. Our results show that this biologically-informed architecture significantly enhances both sensitivity and specificity in off-target prediction, indicating its potential to improve the safety and design of CRISPR experiments.

KEY WORDS: CRISPR-Cas9, Capsule networks, Bidirectional LSTM, Transformer blocks, Off-target DNA sequences

1. INTRODUCTION

CRISPR-Cas9, short for Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR-associated protein 9, is a powerful gene-editing tool derived from a bacterial immune response. It works by guiding the Cas9 enzyme to a specific DNA sequence using a matching RNA guide, enabling precise double-strand breaks at targeted genomic locations. Once the DNA is cut, the cell's natural repair processes are activated. These repair mechanisms often introduce mutations that can disrupt the function of the targeted gene, effectively knocking it out. Alternatively, if a DNA template is provided during repair, the system can be used to insert or replace genetic material at the break site, allowing for accurate gene editing.

CRISPR holds immense potential with applications spanning various fields. It can be leveraged to investigate gene functions across species and transform agriculture by developing high-yield crops that are resistant to diseases and weeds. In the medical field, it offers promising avenues for treating genetic disorders. Despite its revolutionary capabilities, CRISPR faces significant challenges—most notably, Off-Target effects. These occur when the single guide RNA (sgRNA) binds to unintended DNA sites with minor mismatches, still leading to unintended double-strand breaks. Detecting these Off-Target sites using experimental approaches is often expensive and time-intensive. This challenge opens the door for computational models, which can provide efficient, accurate, and economical alternatives for predicting Off-Target events in CRISPR-based genome editing.

Accurately identifying potential Off-Target effects is vital for the effective use of the CRISPR Cas-9 system. In recent years, computational techniques—particularly those rooted in machine learning—have made significant progress in domains such as genomic analysis, drug design, and motif identification. In line with this trend, predictive models have also been developed to assess CRISPR Off-Target activity. Earlier approaches were largely score-based, assigning values according to the number and positions of mismatches between the guide RNA and target sequences. However, these scoring systems often overlook the intricate relationships within DNA sequences and are susceptible to inconsistencies arising from experimental noise or limitations.

2. LITERATURE REVIEW

The development of the CRISPR-Cas9 genome editing system has profoundly transformed modern biology, offering a precise, accessible and highly efficient approach for targeted DNA modification across a wide range of organisms. The system relies on a single guide RNA (sgRNA) that directs the Cas9 nuclease to the desired

genomic locus by complementary base-pairing with the target DNA sequence. Once the Cas9-sgRNA complex binds to its target, the nuclease introduces a double-strand break (DSB) at the site, which is subsequently repaired by the cell through either non-homologous end joining (NHEJ) or homology-directed repair (HDR) [1].

Despite its programmability, CRISPR-Cas9 is not infallible. The system can tolerate mismatches between the sgRNA and the DNA target, particularly in regions outside the seed sequence, which is located adjacent to the protospacer adjacent motif (PAM) [2]. This mismatch tolerance can lead to unintended binding and cleavage at off-target sites, a phenomenon that poses significant risks in both therapeutic and experimental applications. Off-target effects typically arise from imperfect sgRNA-DNA interactions, including single or multiple mismatches, small insertions or deletions (indels) that result in local structural distortions (known as bulges), or errors introduced during DSB repair by NHEJ, which may cause additional indels at the cleavage site [3,4,5]. In the context of gene therapy, such unintended alterations can disrupt tumour suppressor genes or activate oncogenes, raising substantial safety concerns [6]. Even in basic research, undetected off-target edits can compromise gene-function studies, introducing confounding variables and misleading interpretations [7].

Several factors influence how often off-target events occur and how severe they are, including the number and position of mismatches, the local chromatin state, the availability of PAM sites, and the concentration of the Cas9–sgRNA complex within the cell [8]. These complexities make it essential to develop accurate computational tools for predicting off-target effects, supported by experimental validation, to ensure genome editing remains both reliable and safe.

Early attempts to predict off-target activity mostly relied on alignment-based scoring and mismatch counting. Tools like CCTop [9] and CHOPCHOP [10] used straightforward heuristics to rank possible off-target sites based on sequence similarity and permitted mismatches, often placing particular emphasis on the presence of a protospacer adjacent motif (PAM). While these methods were useful for generating broad predictions, they fell short of capturing the more complex sequence dependencies and position-specific effects that are often seen in real CRISPR experiments.

The development of high-throughput experimental validation methods, such as GUIDE-seq [11] and CIRCLE-seq [12], allowed researchers to move beyond purely heuristic models by generating large-scale, experimentally verified off-target datasets. Alongside these, other benchmark datasets like SITE-Seq, as well as several mismatch-based datasets including those from Haeussler, Hek293t, K562, and Listgarten, have become essential resources. Together, these datasets form the foundation for training machine learning and deep learning models capable of identifying binding patterns directly from real biological data.

Deep learning has emerged as a particularly powerful approach for modelling genomic sequences. These architectures can automatically extract features from raw DNA data, allowing them to identify both low-level sequence motifs and more complex, higher-order dependencies that influence DNA–protein interactions. One of the earliest CRISPR-focused models, DeepCRISPR [13], used convolutional neural networks (CNNs) to detect local sequence patterns associated with Cas9 activity. Later models such as CnnCrispr [14] incorporated recurrent neural networks (RNNs) with CNNs, specifically long short-term memory (LSTM) layers, to capture broader sequence context and global dependencies across input sequences.

Further improvements have been achieved through the incorporation of attention mechanisms, originally introduced in the field of natural language processing [15]. Models that incorporate attention are able to assign varying levels of importance to different positions within a sequence, which often results in improved predictive accuracy and offers biologically interpretable insights into the regions of input that contribute most significantly to the model's decision-making process [16,17]. This approach has proven particularly valuable in sequence-to-function prediction tasks, such as protein–DNA binding and variant effect prediction, where only a specific subset of nucleotides may determine the biological outcome.

Another important aspect of CRISPR off-target prediction is the way in which DNA sequences are represented. Most existing models continue to rely on one-hot encoding, in which each nucleotide is expressed as a simple four-element vector. However, this method often restricts the model's ability to identify higher-order motifs and complex biological patterns. Alternative encoding strategies, such as k-mer tokenisation, have demonstrated superior performance across a range of genomic prediction tasks [18,19]. By representing sequences as overlapping fixed-length subunits, k-mers allow deep learning models to capture biologically meaningful motifs and contextual dependencies more effectively than one-hot encodings.

Building beyond these foundations, several specialised deep learning architectures have been developed to address the specific challenges associated with off-target prediction. Among these, the CRISPR-M model adopted a multi-branch design that combined CNN and RNN pathways, each independently processing local and global sequence features before merging their outputs for prediction [20]. This separation of features enabled the model to handle mismatches and indels more effectively, although the increased complexity of the architecture required larger datasets to achieve stable training.

Another significant contribution was made by R-CRISPR, which employed binary matrix encoding to preserve position-specific information throughout the learning process [21]. The combination of convolutional and recurrent layers enabled the model to generalise across a range of sequence types, although it introduced greater sensitivity to class imbalance, particularly in cases where true off-target events were relatively rare.

The CRISPR-IP model further extended the hybrid approach by integrating CNN, BiLSTM, and attention mechanisms within a single pipeline [22]. This design achieved strong predictive accuracy and improved biological interpretability, as the attention mechanism was able to highlight influential nucleotides within candidate off-target sequences. However, attention-based models tend to be computationally intensive and often require longer training cycles to converge effectively.

The CRISPR-DIPOFF model shifted the focus of the field towards interpretability rather than maximising predictive performance. By using integrated gradients, this framework identified the sequence positions that exerted the greatest influence on model predictions [23]. Although this approach provided valuable biological insights, the predictive performance of DIPOFF was occasionally limited by its relatively shallow architecture when compared with deeper, multi-layer hybrid models [24].

In this study, novel deep learning architecture that integrates Capsule Networks with transformer blocks, bidirectional LSTM layers, and CNNs, all enriched by k-mer encoded sequence features and biological rule checks, to predict CRISPR-Cas9 off-target activity with high accuracy. The model is trained on a collection of experimentally validated datasets, including both indel-based sources (GUIDE-seq, CIRCLE-seq) and mismatch-based sources (Haeussler, Hek293t, K562, Listgarten, and SITE-Seq). To improve the model's ability to detect biologically meaningful sequence patterns, input sequences are represented using k-mer-based tokenisation. This design aims to enhance both the accuracy and interpretability of off-target prediction, contributing to the continued development of computational tools for safer and more reliable genome editing.

3. DATA COLLECTION AND PREPROCESSING

3.1 DATASETS

In this study, we utilized multiple publicly available datasets sourced from the Kaggle repository titled "Off-Target Datasets for CRISPR-Cas9." The datasets, provided in CSV format, consist of both Indel and Mismatch types, each representing distinct aspects of CRISPR-Cas9 off-target prediction. Specifically, the Indel datasets include GUIDE-Seq and CIRCLE-Seq, while the Mismatch datasets consist of Haeussler, Hek293t, K562, Listgarten, and SITE-Seq. These datasets contain DNA sequences and their corresponding off-target predictions, which are crucial for evaluating the efficiency and accuracy of CRISPR-Cas9 targeting.

The datasets were processed through a standardized pipeline to maintain consistency. Columns were renamed according to a common convention, aligning guide sequences, off-target sequences, and labels across datasets [25]. This allowed the datasets to be merged into one unified dataset. Metadata, such as the source of the dataset and the type of sequence (Indel or Mismatch), was included for better clarity in subsequent analyses.

3.2 DATA ANALYSIS

Once the datasets were standardized and combined, the next step involved exploring their structure and class distribution. We examined the shape of each dataset and the overall combined dataset to ensure correct merging. A breakdown of the class distribution revealed an imbalanced dataset, with a higher proportion of samples belonging to one class over the other, which is common in biological datasets.

To further analyze the characteristics of the data, we focused on the mismatch count between the guide and off-target sequences. The number of mismatches is an important feature in evaluating the specificity of CRISPR-Cas9 targeting. We calculated the mismatch count for each guide-off sequence pair, iterating through both sequences and counting the positions where the nucleotides differ. This provides insights into the potential off-target effects of the CRISPR system.A detailed mismatch distribution was generated by grouping the data based on the mismatch count and the associated label [26]. This analysis offers an overview of how mismatch counts are distributed across off-target and on-target sequences, highlighting the prevalence of mismatches in different scenarios.

3.3 CREATION OF SYNTHETIC DATA

In order to augment the dataset and improve the model's ability to generalise, we generated synthetic data by introducing controlled mismatches to the sequences. This allowed us to simulate various levels of off-target activity, which is crucial for training a robust model.

To augment the dataset, synthetic off-target sequences were created in three distinct categories based on their resemblance to the guide sequences. The first group comprised perfect matches, which were identical to the guide and presumed to have no off-target activity [27]. The second group, near perfect involved sequences with 1–2 mismatches, considered likely to remain on-target. The third group intermediate matches included sequences with 3–4 mismatches, reflecting uncertain off-target behavior.

Labels were assigned using heuristic criteria, such as the presence of PAM-like motifs. These synthetic examples were then combined with the original data to support improved generalization and robustness in subsequent model training.

4. FEATURE ENGINEERING AND K-MER PROCESSING

Once the synthetic data was generated and merged with the original dataset, we proceeded with feature extraction. We focused on extracting meaningful features that could capture the complexities of the CRISPR-Cas9 off-target prediction task.

4.1 K-MER EXTRACTION

Once the synthetic data was generated and merged with the original dataset, we proceeded with feature extraction. We focused on extracting meaningful features that could capture the complexities of the CRISPR-Cas9 off-target prediction task.A critical feature in understanding sequence data is the use of k-mers, which are subsequences of length k that capture local patterns in DNA sequences. In this study, we used a k-mer length of 3 (trimer analysis), as it strikes a balance between capturing sequence motifs and maintaining computational efficiency.We created a mapping of all possible 3-mers from the nucleotide alphabet (A, C, G, T) and assigned each k-mer a unique index. This allowed us to convert the DNA sequences into numerical representations that could be fed into machine learning models.

4.2 SEQUENCE REPRESENTATION

For each sequence pair (guide and off-target), we computed several features to capture relevant sequence characteristics. First, both the guide and off-target sequences were converted into k-mer representations using a previously created mapping. This conversion allowed the model to capture sequence patterns that might indicate off-target potential. We also calculated the total number of mismatches between the guide and off-target sequences by comparing each nucleotide position in the two sequences. Additionally, we computed position-weighted mismatches, where mismatches closer to the PAM region were given higher importance due to their greater relevance in CRISPR-Cas9 targeting efficiency. The GC content of the off-target sequence was also calculated, as it is known to influence the efficiency of CRISPR-Cas9 targeting; higher GC content can affect the stability and binding efficiency of the guide sequence. Lastly, since the PAM region plays a crucial role in CRISPR-Cas9 targeting, we extracted the last three bases of the off-target sequence to check for the presence of common PAM motifs such as NGG, NAG, and NGA. Binary features were created to indicate the presence of each PAM sequence.

Each sequence in the augmented dataset, which included both original and synthetic data, was processed using the above feature extraction methods. For each guide-off-target pair, we calculated the k-mer features for both the guide and off-target sequences, the total mismatch count, and the weighted mismatch count. Additionally, we computed the GC content of the off-target sequence and created binary flags indicating the presence of different PAM motifs. After processing, the extracted features were stored in a DataFrame for further analysis.

After feature extraction, we calculated summary statistics for key features, including mismatch count, weighted mismatches, and GC content. We also examined the distribution of PAM motifs across the dataset. This analysis helped ensure that the feature engineering process had captured relevant characteristics of the sequences and prepared the data for model training. Once the feature extraction process was complete, we performed several steps to ensure the data was properly prepared for model training, focusing on sequence padding, numerical feature integration, and data balancing.

4.3 CLASS IMBALANCE HANDLING

Upon analyzing the class distribution of the dataset, it became evident that the dataset was imbalanced, with one class (either on-target or off-target) being more prevalent than the other. Addressing this class imbalance is critical for improving model performance and ensuring that the model is capable of predicting both classes accurately.

To mitigate the impact of class imbalance, we employed a two-step process involving undersampling of the majority class and oversampling of the minority class.

4.3.1 Majority Class Reduction (Undersampling)

To begin addressing the class imbalance, we reduced the number of samples from the majority class in the training set. We identified the minority class (either 0 or 1, depending on the dataset) and calculated the number of samples in the minority class. The majority class was then undersampled to achieve a 2:1 ratio, where the majority class samples were reduced to twice the number of the minority class samples.

4.3.2 Minority Class Oversampling (SMOTE)

Following the undersampling step, we used the Synthetic Minority Over-sampling Technique (SMOTE) to oversample the minority class and increase its representation in the training set. SMOTE creates synthetic samples of the minority class by interpolating between existing minority class samples, allowing the model to better learn the characteristics of the minority class.

SMOTE was applied to both the sequence data and numerical features. The sequence data was reshaped before applying SMOTE, and the output was reshaped back to

its original structure. This allowed us to balance the class distribution in the training set, ensuring that both classes had sufficient representation for model training. The final balanced dataset, with both undersampled majority class samples and oversampled minority class samples, was used for model training.

5. METHODOLOGY

This section describes the construction of an advanced hybrid model for CRISPR-Cas9 off-target prediction, integrating Capsule Networks, Transformer layers, and a detailed PAM (Protospacer Adjacent Motif) detection mechanism. The model effectively combines both sequence-based and numerical features to enhance the prediction accuracy. Figure 1 shows the breakdown of the methodology for each layer and component in the architecture.

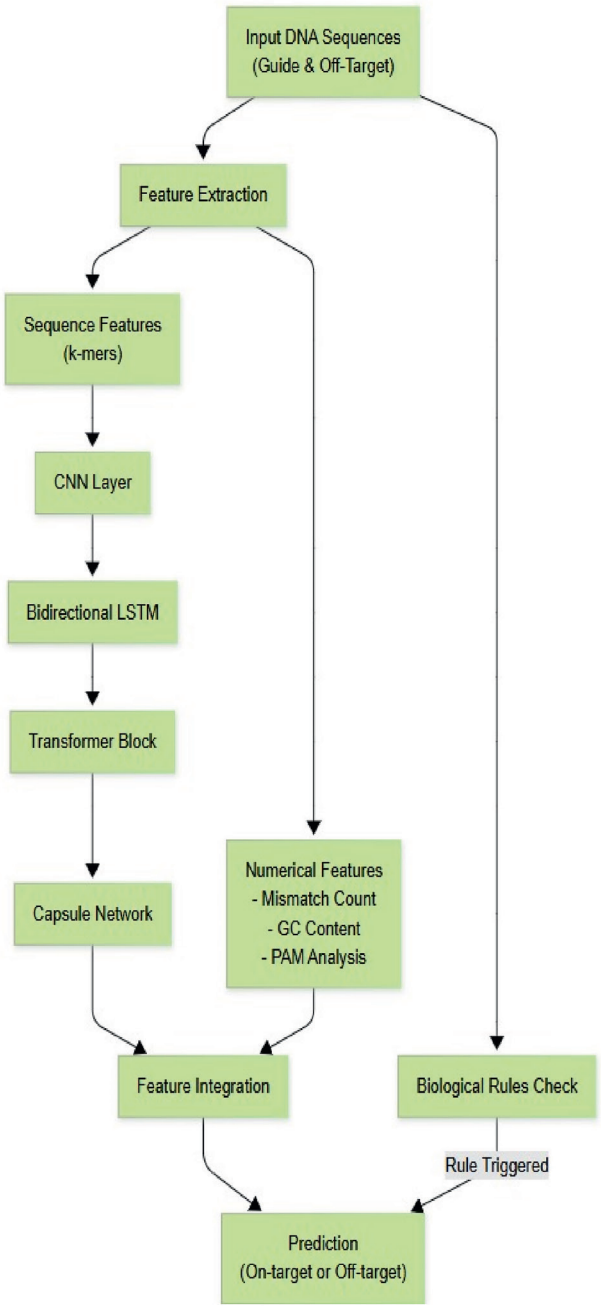

Figure 1. Architecture of the proposed model

5.1 CAPSULE LAYER

A Capsule Layer is used to capture spatial hierarchies in the data. The architecture of the Capsule Layer includes several key components:

The input sequence is transformed by a weight matrix, which is trainable, enabling the model to learn the optimal transformation for each input. The output of each capsule is squashed to ensure the magnitude of the vector is between 0 and 1, thereby maintaining the notion of certainty in predictions.Capsules interact with each other through routing by agreement. This is achieved through the iterative refinement of the agreement between the capsules' predictions, improving the ability of the model to understand relationships between various features of the sequence. The CapsuleLayer accepts sequence input, transforms it via capsules, and produces output for further processing in the model. The capsule layer provides a robust mechanism for the model to capture complex patterns and hierarchical relationships within the sequence data.

5.2 TRANSFORMER BLOCK

The Transformer block is implemented to capture long-range dependencies and global patterns within the sequence data, which are essential for understanding sequence-specific motifs like PAM. The Transformer block includes Multi-Head Self-Attention which allows the model to focus on different parts of the input sequence simultaneously. It helps identify important features and relationships at various scales.A fully connected network follows the self-attention mechanism, consisting of a ReLU activation function and a second dense layer, which further processes the output of the attention layer.Layer Normalization is applied to stabilize the training process and speed up convergence by maintaining the distribution of activations.The transformer block refines sequence representations by leveraging attention mechanisms to learn contextual dependencies, which are crucial for tasks such as off-target prediction in CRISPR-Cas9 applications.

5.3 PAM DETECTION

To capture the critical Protospacer Adjacent Motif (PAM), a dedicated step is introduced within the model. The PAM sequence is extracted from the target sequence to aid in off-target prediction:The target sequence is sliced to isolate the last three positions. After extraction, the PAM sequence is passed through a fully connected dense layer to further process and transform the extracted PAM features into a usable representation for downstream tasks. This

explicit detection of the PAM sequence is a significant enhancement, allowing the model to better understand the specificity of CRISPR-Cas9 targeting.

5.4 FEATURE FUSION

The model incorporates multiple types of data: sequence-based features (guide and target sequences), numerical features (such as GC content, off-target scores, etc.), and the PAM-specific features.

5.4.1 Guide and Target Embedding

Both the guide and target sequences are independently embedded using an embedding layer, which converts the sequence of nucleotides into continuous-valued vectors.

5.4.2 Concatenation

The embedded guide and target sequences are concatenated to form a unified representation, which is then processed by convolutional layers.

5.4.3 CNN and BiLSTM

A Convolutional Neural Network (CNN) extracts local sequence patterns, followed by a Bidirectional Long Short-Term Memory (BiLSTM) layer, which learns temporal dependencies in the sequence. This combination allows the model to capture both local motifs and global sequence dependencies.

5.4.4 Capsule Network

After the CNN and BiLSTM layers, the data is processed by the Capsule Layer to capture hierarchical relationships between features.

5.4 NUMERICAL FEATURE PROCESSING

The model also includes numerical features that can significantly impact the CRISPR-Cas9 off-target prediction. These features could include the distance between guide and target, GC content, or other domain-specific features:

5.4.1 Dense Layer for Numerical Features

The numerical features are processed through a fully connected layer with ReLU activation and dropout regularization. This allows the model to learn complex relationships between numerical and sequence-based features.

The final output of the model is a binary classification, indicating whether a given guide-target pair will result in an off-target effect. The outputs of the Capsule Layer, numerical features, and PAM processing are concatenated. A dense layer with 64 units is applied to learn the joint representation of the sequence and numerical features. Dropout layers are applied throughout the network to prevent overfitting and ensure generalization. A final sigmoid activation function is used to produce the binary classification output (off-target vs. on-target).

The proposed architecture combines advanced deep learning techniques — Capsule Networks, Transformer layers, and CNN-BiLSTM — to enhance the prediction of CRISPR-Cas9 off-target effects. The addition of a PAM detection mechanism and the careful fusion of both sequence-based and numerical features ensure that the model captures all critical aspects of the problem, offering a comprehensive approach to off-target prediction. Through this methodology, we aim to improve the specificity and efficiency of CRISPR-Cas9 genome-editing applications.

6. EXPERIMENTAL RESULTS AND DISCUSSION

Off-Target Datasets for CRISPR-Cas9 used in this research are sourced from kaggle. These datasets contain DNA sequences and their corresponding off-target predictions, which are crucial for evaluating the efficiency and accuracy of CRISPR-Cas9 targeting.

The proposed model is compiled using the Adam optimizer with a learning rate of 0.001. The loss function is binary cross-entropy, which is appropriate for binary classification tasks. The model is evaluated based on multiple metrics, including:

Accuracy: Measures the overall correctness of the predictions.

AUC (Area Under the ROC Curve): Assesses the model's ability to discriminate between positive and negative classes.

Precision: Measures the proportion of true positives among all positive predictions.

Recall: Measures the proportion of true positives among all actual positives.

AUC-PR (Area Under the Precision-Recall Curve): Focuses on the model's performance when dealing with imbalanced classes.

6.1 TRAINING THE MODEL

To address the class imbalance inherent in the CRISPR-Cas9 off-target prediction task, the model is trained using a balanced dataset. Class weights are calculated manually to give higher importance to the minority class (off-target predictions), ensuring that the model does not become biased towards the majority class. The class weights are defined as

Class 0 (non-target): Weight = 1.0, Class 1 (off-target): Weight = 3.0

This weighting scheme is applied to account for the higher importance of correctly identifying off-target sites. The model is compiled using the following configuration:

Optimizer: The Adam optimizer is selected due to its adaptive learning rate properties, which are well-suited for training deep learning models on complex tasks such as off-target prediction. The learning rate is set to 0.001.

Loss Function: The model utilizes binary cross-entropy as the loss function, suitable for binary classification problems.

Evaluation Metrics: Several metrics are employed to assess the model's performance, with particular emphasis on metrics that are relevant for imbalanced datasets:

Accuracy: Measures the proportion of correct predictions.

Area Under the ROC Curve (AUC): Evaluates the model's ability to distinguish between classes.

Area Under the Precision-Recall Curve (PR AUC): Provides insight into the model's ability to predict the minority class (off-target predictions).

Precision and Recall: These metrics are crucial for evaluating the performance of the model in predicting the minority class.

6.2 HYPERPARAMETER TUNING AND OPTIMIZATION

The model is trained using the fit function, with the following parameters:

Input Data: The model receives two inputs: the guide-target sequences and the associated numerical features.

Epochs: The model is trained for 50 epochs to allow adequate learning and fine-tuning.

Batch Size: A batch size of 64 is used for efficient training.

Class Weights: The class weights are passed to the model to address class imbalance and ensure that the minority class is given higher importance during training.

To prevent overfitting and improve model performance, several callbacks are used:

EarlyStopping: This callback halts training if the validation PR AUC score does not improve for 10 consecutive epochs, and the best weights are restored.

ReduceLROnPlateau: If the validation PR AUC score plateaus for 5 epochs, the learning rate is reduced by a factor of 0.5 to facilitate convergence.

ModelCheckpoint: The best model, based on validation PR AUC, is saved during training to ensure that the optimal model is used for evaluation.

Figure 2 and Figure 3 shows the Precision Recall and Precision Recall AUC obtained through the model.

We assessed the performance of the hybrid deep learning model by tracking key metrics, including precision, recall, and PR-AUC (Precision-Recall Area Under the Curve), throughout the training and validation stages. As training progressed, we observed consistent improvements, and the evaluation metrics for both the training and validation sets

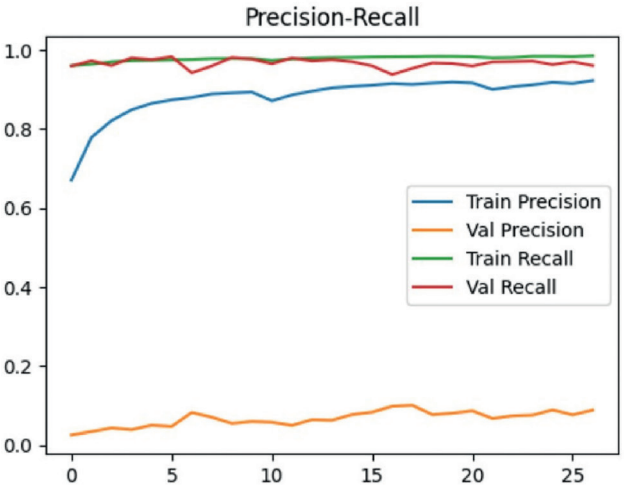

Figure 2. Precision recall curve of the proposed model

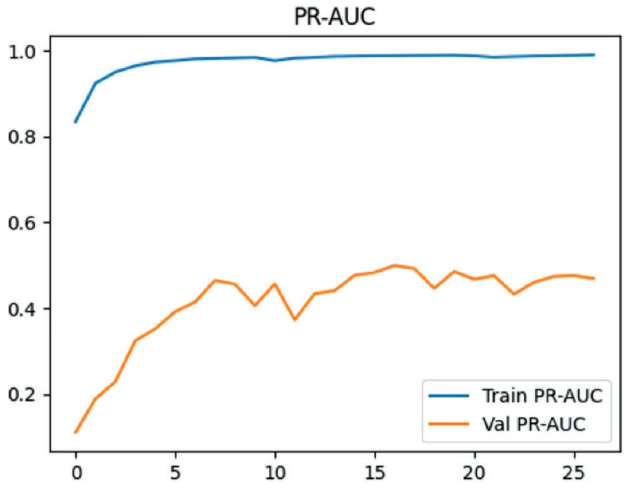

Figure 3. Precision recall - AUC curve of the proposed model

closely aligned. This suggests that the model generalised well, showing no significant overfitting despite the class imbalance in the dataset.

To address this imbalance, we adjusted the test set distribution so that class 0 samples outnumbered class 1 samples by 40%. This modification reflected real-world conditions where off-target edits are less frequent than true edits. As a result, the test set consisted of 3,487 samples from class 0 (58.3%) and 2,491 samples from class 1 (41.7%).

The model performed exceptionally well on the modified test set, with class 0 showing a precision of 0.94 and recall of 0.95. For class 1, the precision was 0.93, and the recall was 0.92. The macro-average F1-score came out at 0.93, and the test PR-AUC was 0.9836. These results indicate that the model captured the relevant sequence and numerical features for off-target prediction, even in the face of class imbalance. The high PR-AUC value highlights the model's strong ability to differentiate between the two classes, maintaining both sensitivity and specificity.

In addition to conventional evaluation metrics, we developed a biologically-constrained prediction wrapper to improve the interpretability of the model's predictions. This wrapper applies fundamental biological rules, such as checking for valid PAM sequences and filtering out perfect matches or targets with high mismatch counts. We tested this wrapper on a range of biologically diverse cases and found that it successfully filtered out irrelevant targets (e.g., perfect matches or invalid PAMs). The model then accurately handled more ambiguous cases, providing high-confidence predictions.

These results demonstrate not only the accuracy of the model but also its ability to respect essential biological constraints. This makes the model particularly suitable for applications in experimental screening and CRISPR-Cas9 off-target prediction, where both prediction accuracy and biological relevance are crucial.

In this study, we introduced a comprehensive deep learning framework designed to improve the prediction of CRISPR-Cas9 off-target effects, which is a major challenge that continues to limit the safe clinical use of gene editing. Our hybrid model combines convolutional layers, bidirectional LSTMs, transformer blocks and capsule networks to learn both local and global patterns in guide and off-target DNA sequences.

We enhanced the model's interpretability and biological relevance by incorporating features such as mismatch counts, GC content and PAM motif presence, and applied rule-based biological checks to ensure the predictions align with established genomic behaviour. Using k-mer encoding and a multi-channel feature fusion strategy, the model captured sequence complexity effectively. We also addressed class imbalance through a combination of SMOTE oversampling and undersampling techniques, which helped the model generalise better across classes.

Our results show that this multi-representational, biologically informed approach improves both sensitivity and specificity in off-target prediction, offering a practical tool to help prioritise off-target candidates for experimental validation. By embedding biological insight within a powerful deep learning architecture, our method supports the more efficient and safer design of CRISPR-based experiments. Looking ahead, we aim to expand the model to consider cell-type-specific chromatin environments and to validate its predictions on experimentally confirmed off-target datasets, moving it closer to clinical and therapeutic applications.

7. CONCLUSION

In this study, we introduced a comprehensive deep learning framework designed to improve the prediction of CRISPR-Cas9 off-target effects, which is a major challenge that continues to limit the safe clinical use of gene editing. Our hybrid model combines convolutional layers, bidirectional LSTMs, transformer blocks and capsule networks to learn both local and global patterns in guide and off-target DNA sequences.

We enhanced the model's interpretability and biological relevance by incorporating features such as mismatch counts, GC content and PAM motif presence, and applied rule-based biological checks to ensure the predictions align with established genomic behaviour. Using k-mer encoding and a multi-channel feature fusion strategy, the model captured sequence complexity effectively. We also addressed class imbalance through a combination of SMOTE oversampling and undersampling techniques, which helped the model generalise better across classes.

Our results show that this multi-representational, biologically informed approach improves both sensitivity and specificity in off-target prediction, offering a practical tool to help prioritise off-target candidates for experimental validation. By embedding biological insight within a powerful deep learning architecture, our method supports the more efficient and safer design of CRISPR-based experiments. Looking ahead, we aim to expand the model to consider cell-type-specific chromatin environments and to validate its predictions on experimentally confirmed off-target datasets, moving it closer to clinical and therapeutic applications.

8. REFERENCES

1. JINEK, M., CHYLINSKI, K., FONFARA, I., HAUER, M., DOUDNA, J.A. and CHARPENTIER, E., (2012). A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science*, *337*(6096), 816–821.

2. HSU, P.D., SCOTT, D.A., WEINSTEIN, J.A., RAN, F.A., KONERMANN, S., AGARWALA, V., LI, Y., FINE, E.J., WU, X., SHALEM, O. and CRADICK, T.J., (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology*, *31*(9), 827–832.
3. FU, Y., FODEN, J.A., KHAYTER, C., MAEDER, M.L., REYON, D., JOUNG, J.K. and SANDER, J.D., (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature Biotechnology*, *31*(9), 822–826.
4. ZHANG, X.H., TEE, L.Y., WANG, X.G., HUANG, Q.S. and YANG, S.H., (2015). Off-target effects in CRISPR/Cas9-mediated genome engineering. *Molecular therapy Nucleic acids*, *4,* e264.
5. YEH, C.D., RICHARDSON, C.D. and CORN, J.E., (2019). Advances in genome editing through control of DNA repair pathways. *Nature Cell Biology*, *21*(12), 1468–1478
6. K. VINAY KUMAR, SUMANASWINI PALAKURTHY, SRI HARSHA BALIJADADDANALA, SHARMILA REDDY PAPPULA, and ANIL KUMAR LAVUDYA. (2024). Early detection and diagnosis of oral cancer using deep neural network. *Journal of Computer Allied Intelligence*, *2*(2), 22–34.
7. KOSICKI, M., TOMBERG, K. and BRADLEY, A., (2018). Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. *Nature Biotechnology*, *36*(8), 765–771.
8. TSAI, S.Q., ZHENG, Z., NGUYEN, N.T., LIEBERS, M., TOPKAR, V.V., THAPAR, V., WYVEKENS, N., KHAYTER, C., IAFRATE, A.J., LE, L.P. and ARYEE, M.J., (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature biotechnology*, *33*(2), 187–197.
9. SWAPNA SATURI, and ARUN KUMAR SILIVERY. (2024). Computer allied intelligence in the education resource-sharing based incontract deep learning. *Journal of Computer Allied Intelligence*, *2*(4), 51–69.
10. WU, X., SCOTT, D.A., KRIZ, A.J., CHIU, A.C., HSU, P.D., DADON, D.B., CHENG, A.W., TREVINO, A.E., KONERMANN, S., CHEN, S. and JAENISCH, R., (2014). Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nature Biotechnology*, *32*(7), 670–676
11. STEMMER, M., THUMBERGER, T., DEL SOL KEYER, M., WITTBRODT, J. and MATEO, J.L., (2015). CCTop: An intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. *PloS One*, *10*(4), e0124633.
12. LABUN, K., MONTAGUE, T.G., GAGNON, J.A., THYME, S.B. and VALEN, E., (2016). CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Research*, *44*(W1), W272-W276.
13. SRINIVASA SAI ABHIJIT CHALLAPALLI. (2024). Optimizing dallas-fort worth bus transportation system using any logic. *Journal of Sensors, IoT and Health Sciences*, *2*(4), 40–55.
14. TSAI, S.Q., ZHENG, Z., NGUYEN, N.T., LIEBERS, M., TOPKAR, V.V., THAPAR, V., WYVEKENS, N., KHAYTER, C., IAFRATE, A.J., LE, L.P. and ARYEE, M.J., (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature Biotechnology*, *33*(2), 187–197.
15. TSAI, S.Q., NGUYEN, N.T., MALAGON-LOPEZ, J., TOPKAR, V.V., ARYEE, M.J. and JOUNG, J.K., (2017). CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. *Nature Methods*, *14*(6), 607–614
16. CHUAI, G., MA, H., YAN, J., CHEN, M., HONG, N., XUE, D., ZHOU, C., ZHU, C., CHEN, K., DUAN, B. and GU, F., (2018). DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biology*, *19*, 1–18.
17. SRINIVASA SAI ABHIJIT CHALLAPALLI. (2024). Sentiment analysis of the twitter dataset for the prediction of sentiments. *Journal of Sensors, IoT and Health Sciences*, *2*(4), 1–15.
18. ZHANG, G., DAI, Z. and DAI, X., (2020). C-RNNCrispr: Prediction of CRISPR/Cas9 sgRNA activity using convolutional and recurrent neural networks. *Computational and Structural Biotechnology Journal*, *18*, 344–354.
19. VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A.N., KAISER, Ł. and POLOSUKHIN, I., (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.
20. PARK, S., KOH, Y., JEON, H., KIM, H., YEO, Y. and KANG, J., (2020). Enhancing the interpretability of transcription factor binding site prediction using attention mechanism. *Scientific Reports*, *10*(1), 13413.
21. ZHOU, J., THEESFELD, C.L., YAO, K., CHEN, K.M., WONG, A.K. and TROYANSKAYA, O.G., (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, *50*(8), 1171–1179
22. KOO, P.K. and EDDY, S.R., (2019). Representation learning of genomic sequence motifs with convolutional neural networks. *PLoS Computational Biology*, *15*(12), e1007560.

23. NG, P., (2017). dna2vec: Consistent vector representations of variable-length k-mers. *arXiv Preprint arXiv:1701.06279*.
24. SUN, J., GUO, J. and LIU, J., (2024). CRISPR-M: Predicting sgRNA off-target effect using a multi-view deep learning network. *PLOS Computational Biology*, *20*(3), e1011972.
25. NIU, R., PENG, J., ZHANG, Z. and SHANG, X., (2021). R-CRISPR: A deep learning network to predict off-target activities with mismatch, insertion and deletion in CRISPR-Cas9 system. *Genes*, *12*(12), 1878
26. ZHANG, Z.R. and JIANG, Z.R., (2022). Effective use of sequence information to predict CRISPR-Cas9 off-target. *Computational and Structural Biotechnology Journal*, *20*, 650–661
27. TOUFIKUZZAMAN, M., HASSAN SAMEE, M.A. and SOHEL RAHMAN, M., (2024). CRISPR-DIPOFF: An interpretable deep learning approach for CRISPR Cas–9 off-target prediction. *Briefings in Bioinformatics*, *25*(2), bbad530.