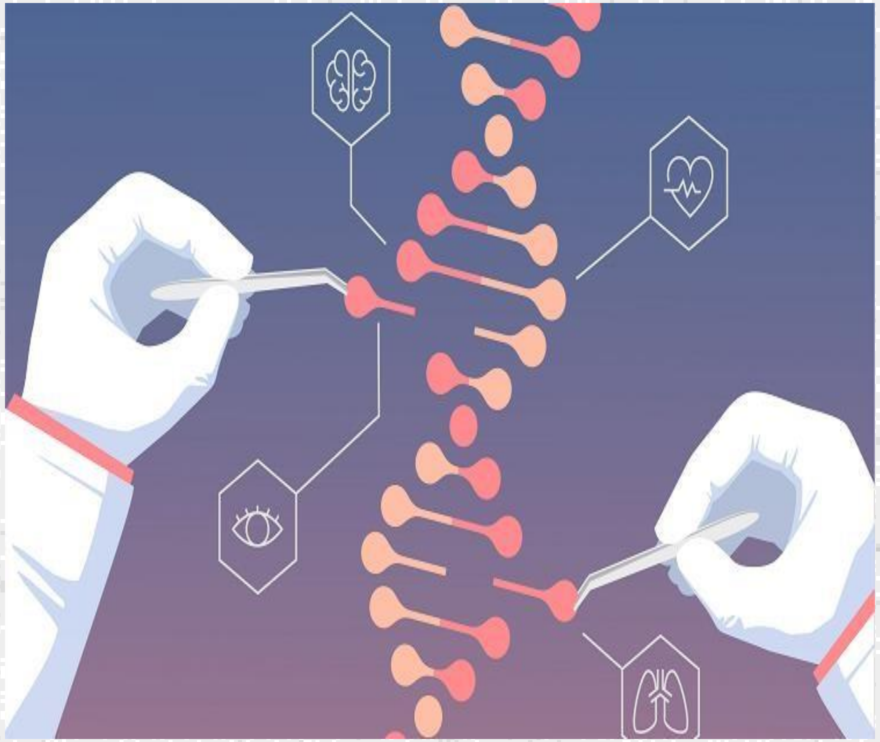
AMRITA
VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY UNDER SECTION 3 OF UGC ACT, 1956

DEEP LEARNING-BASED OFF-TARGET PREDICTION FOR CRISPR-CAS9



Team members:

1. KAILASH S. – CB.AI.U4AIM24017
2. SHREERAM M. – CB.AI.U4AIM2023
3. MAHADEV M. – CB.AI.U4AIM24025
4. SANJAY K. – CB.AI.U4AIM24038

SUBJECTS :

24AIM112 - Molecular biology and basic cellular physiology

24AIM115 - Ethics, innovative research businesses & IPR

INTRODUCTION

- CRISPR-Cas9 is an advanced gene-editing technology that allows precise modifications of DNA. It consists of a guide RNA (gRNA) that directs the Cas9 enzyme to specific target sequences.
- The Cas9 protein then cuts the DNA, enabling genetic modifications for medical and research purposes. CRISPR has shown great promise in treating genetic disorders and cancers by modifying oncogenes.
- However, a major challenge is the occurrence of off-target effects, where unintended DNA sequences are edited. These unintended modifications can cause harmful mutations, gene dysfunction, or even promote cancer progression.

ABSTRACT

- **CRISPR-Cas9** genome editing has revolutionized genetic engineering, but off-target effects, including mismatches and insertions/deletions (indels), pose significant challenges to its precision and safety.

- In this study, we present a deep learning-based off-target prediction model using a Feedforward Neural Network (FNN) trained on CRISPOR and GuideSeq datasets. Our model employs **8-bit one-hot encoding** to effectively represent genomic sequences, enabling the detection of subtle mismatches and indels that influence off-target activity.

- With extensive training and optimization, our FNN achieved **85% accuracy**, surpassing traditional alignment-based methods in predictive performance.

- This high-precision approach enhances CRISPR-Cas9 specificity, providing a robust tool for reducing unintended mutations and ensuring safer gene-editing applications

OBJECTIVE

- To design a machine learning-based system that accurately predicts potential off-target effects in CRISPR-Cas9 gene editing.
- To Convert 23-base DNA sequences into 8-bit binary encoded format.
- To Train and test multiple machine learning models (FNN, CNN, LSTM) for classification.
- To give the output (off-target or on-target) with confident score.

AI MODELS AND DEVELOPMENT

- **Title:** ***Optimizing Precision Genome Editing through Machine Learning***

  Summary:

  This study leverages Support Vector Machines (SVMs) to predict off-target effects in CRISPR-Cas9 gene editing. By analyzing sequence features and genomic context, the SVM model classifies potential off-target sites, thereby reducing unintended mutations. It achieved 64% recall rate

  Dataset Used: GUIDE-seq data

  Importance:

  Demonstrates the effectiveness of classical ML methods in improving guide RNA design, enhancing the overall precision and safety of CRISPR applications.

- **Title: Crispr2vec – Predicting Off-Target Cuts in CRISPR Systems Using Deep Metric Learning**

  Summary:

  Crispr2vec uses deep metric learning to predict off-target effects in CRISPR-Cas9 editing and use it to accurately identify potential off-target sites. It generalizes well across assays like GUIDE-seq and CIRCLE-seq, outperforming tools like CHOPCHOP and Cas-OFFinder. The model achieved Spearman and Pearson correlations of 0.60 and 0.55.

  Dataset Used: GUIDE-seq, CIRCLE-seq

  Importance:

  Demonstrates the power of deep learning in improving off-target prediction, boosting the precision and safety of CRISPR in both therapeutic and agricultural gene editing.

- **Title:** ***CRISPR-Net: A Recurrent Convolutional Network Quantifies CRISPR Off-target Activities with Indels and Mismatches***

  Summary:

  CRISPR-Net is a hybrid deep learning model that combines Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) to capture both spatial and sequential features in genomic sequences. This design allows the model to detect off-target effects, including subtle mismatches and indels.

  Dataset Used: CRISPOR and GUIDE-seq.

  Importance:

  Its architecture significantly enhances off-target detection by accommodating complex sequence variations, thereby improving the safety and efficacy of CRISPR-Cas9 editing.

- **Title:** ***CRISPR-M: Predicting sgRNA Off-Target Effect Using a Multi-View Convolutional Neural Network***

  Summary:

  CRISPR-M introduces a multi-view Convolutional Neural Network (CNN) that processes genomic data from several perspectives to predict single guide RNA (sgRNA) off-target effects. By integrating diverse genomic features, the model achieves robust performance in identifying potential off-target sites.

  Dataset Used: CRISPOR and GUIDE-seq, ensuring a wide representation of off-target events in its training process.

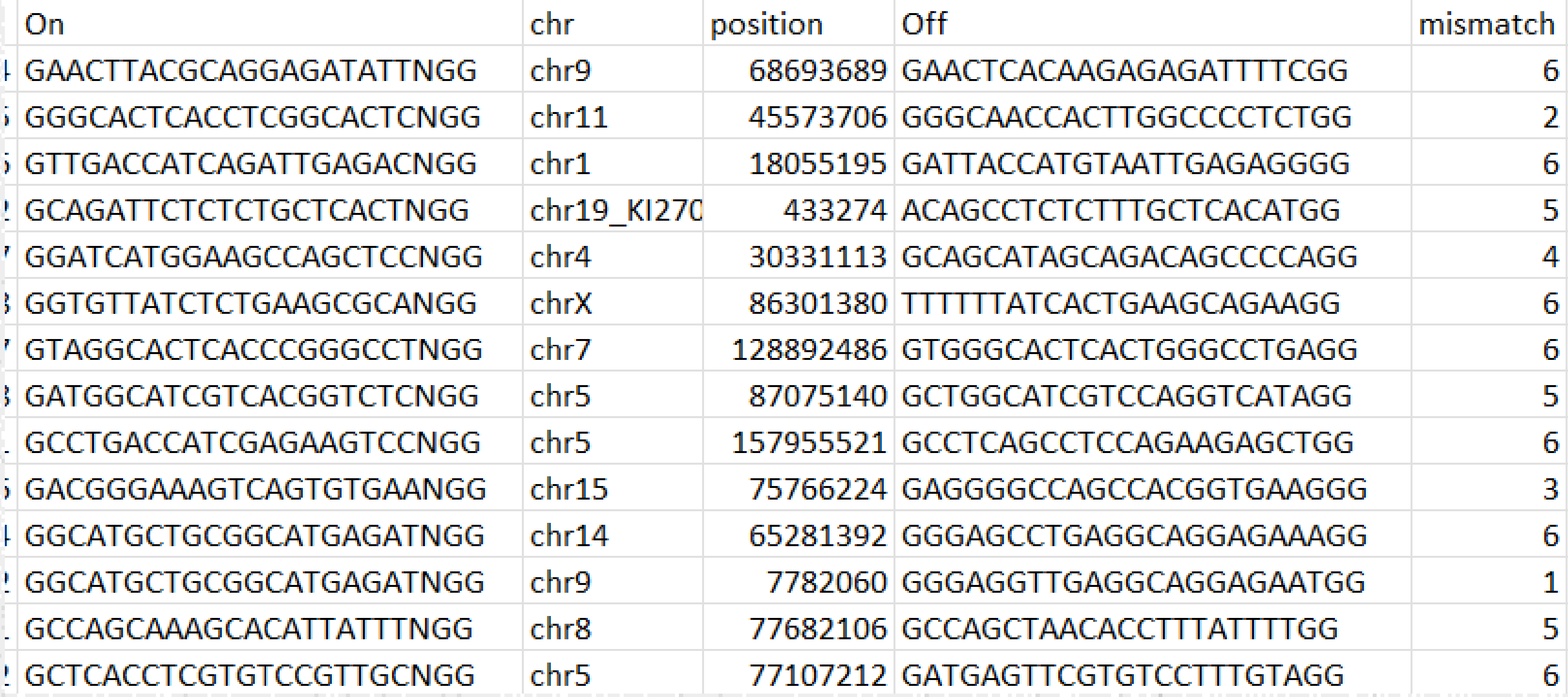
  Importance:

  This multi-view architecture provides a comprehensive analysis of off-target effects, leading to improved sgRNA design and ultimately enhancing the reliability of CRISPR-based gene editing.

DATASET

Our off-target prediction system is built and tested using two reliable and widely recognized datasets in CRISPR-Cas9 research.

- **CRISPOR Dataset**
- **Guide seq Dataset**



On	chr	position	Off	mismatch
GAACTTACGCAGGAGATATTNGG	chr9	68693689	GAACTCACAAGAGAGATTTTCGG	6
GGGCACTCACCTCGGCACTCNGG	chr11	45573706	GGGCAACCACTTGGCCCCTCTGG	2
GTTGACCATCAGATTGAGACNGG	chr1	18055195	GATTACCATGTAATTGAGAGGGG	6
GCAGATTCTCTCTGCTCACTNGG	chr19_KI270	433274	ACAGCCTCTCTTTGCTCACATGG	5
GGATCATGGAAGCCAGCTCCNGG	chr4	30331113	GCAGCATAGCAGACAGCCCCAGG	4
GGTGTTATCTCTGAAGCGCANGG	chrX	86301380	TTTTTTATCACTGAAGCAGAAGG	6
GTAGGCACTCACCCGGGCCTNGG	chr7	128892486	GTGGGCACTCACTGGGCCTGAGG	6
GATGGCATCGTCACGGTCTCNGG	chr5	87075140	GCTGGCATCGTCCAGGTCATAGG	5
GCCTGACCATCGAGAAGTCCNGG	chr5	157955521	GCCTCAGCCTCCAGAAGAGCTGG	6
GACGGGAAAGTCAGTGTGAANGG	chr15	75766224	GAGGGGCCAGCCACGGTGAAGGG	3
GGCATGCTGCGGCATGAGATNGG	chr14	65281392	GGGAGCCTGAGGCAGGAGAAAGG	6
GGCATGCTGCGGCATGAGATNGG	chr9	7782060	GGGAGGTTGAGGCAGGAGAATGG	1
GCCAGCAAAGCACATTATTTNGG	chr8	77682106	GCCAGCTAACACCTTTATTTTGG	5
GCTCACCTCGTGTCCGTTGCNGG	chr5	77107212	GATGAGTTCGTGTCCTTTGTAGG	6

Fig. 1: Part of the Dataset

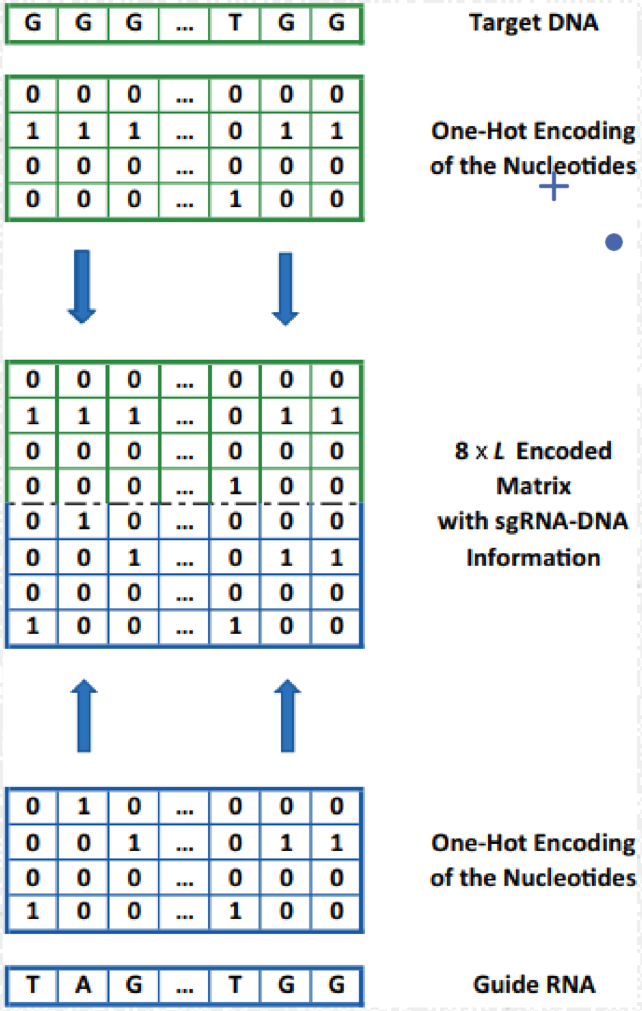
METHODOLOGY

- **Dataset Preparation**
  - Collected Guide seq and CRISPOR datasets.
  - One hot encoded the DNA sequences into 8x23 matrix.
  - Combined the datasets to improve the results.

- **Model Architectures**

1. **Feedforward Neural Network (FNN)**

   - Input: 184 features
   - 3 hidden Dense layers with ReLU
   - Output: Sigmoid node for binary classification
   - Loss: Binary Crossentropy
   - Optimizer: Adam



G	G	G	...	T	G	G

Target DNA

0	0	0	...	0	0	0
1	1	1	...	0	1	1
0	0	0	...	0	0	0
0	0	0	...	1	0	0

One-Hot Encoding of the Nucleotides

0	0	0	...	0	0	0
1	1	1	...	0	1	1
0	0	0	...	0	0	0
0	0	0	...	1	0	0
0	1	0	...	0	0	0
0	0	1	...	0	1	1
0	0	0	...	0	0	0
1	0	0	...	1	0	0

8 x *L* Encoded Matrix with sgRNA-DNA Information

0	1	0	...	0	0	0
0	0	1	...	0	1	1
0	0	0	...	0	0	0
1	0	0	...	1	0	0

One-Hot Encoding of the Nucleotides

T	A	G	...	T	G	G

Guide RNA

Fig. 2: 8-bit One hot encoding

2. Convolutional Neural Network (CNN)

- Input: (23, 8) encoded matrix
- Conv1D layers extract motif patterns
- MaxPooling and Dropout for regularization
- Dense layers for classification
- Effective in capturing local sequence patterns

3. Long Short-Term Memory (LSTM)

- Input: (23, 8) encoded matrix
- LSTM layers capture sequential dependencies
- Dense + Sigmoid output for classification
- Ideal for capturing long-range dependencies in DNA sequences

4. Random Forest (RF)

- Input: 184-dimensional vector
- Predicts class probabilities based on majority voting
- Provides feature importance analysis

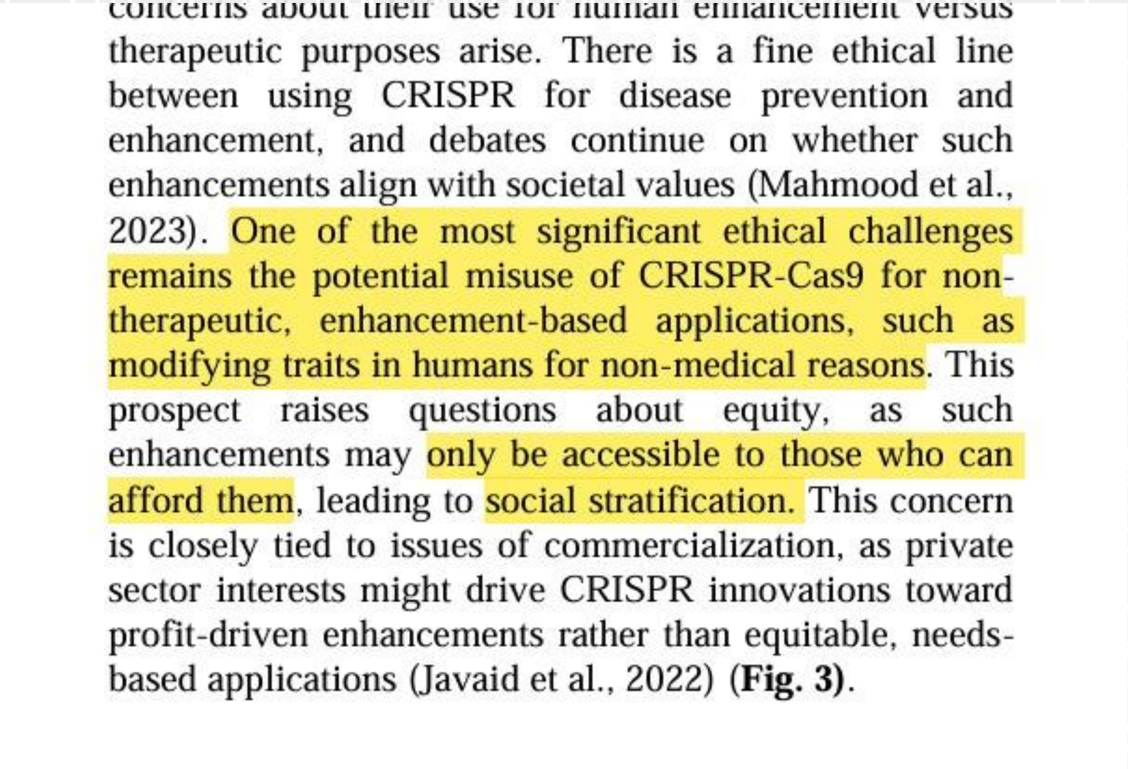
INTELLECTUAL PROPERTY RIGHTS (IPR)

FEATURES	US 10,000,772(Doudna et al.)	US 8,697,359 (Zhang)
HOLDERS	The Regents of the University of California , Oakland , CA (US) ; University of Vienna , Vienna (AT)	The Broad Institute Inc., Cambridge, MA (US); Massachusetts Institute of Technology, Cambridge, MA (US)
KEY INNOVATION	1. Introduction of the *single guide RNA (sgRNA)* which simplifies the CRISPR-Cas9 system. 2. Methods for *transcriptional modulation* using a catalytically inactive Cas9 (dCas9), allowing for regulation of gene expression without cutting the DNA.	1. Adapting and applying the CRISPR-Cas9 system to *eukaryotic cells*. 2. Demonstrating its use in altering gene expression and genome editing within *eukaryotic systems*, including mammalian cells.

INTELLECTUAL PROPERTY RIGHTS(IPR)

FEATURES	US 10,000,772(Doudna et al.)	US 8,697,359 (Zhang)
APPLICATIONS OF THE PATENT	1. Precise DNA modification 2. Regulation of gene expression through transcriptional modulation 3. Genetically modified cells and organisms	1. Genome editing in eukaryotic cells 2. Control of gene expression in eukaryotic cells
CELL TYPE	No specific restriction to cell type, includes both prokaryotes and eukaryotes	Focuses primarily on eukaryotic cells, particularly mammalian cells. The patent demonstrates the effectiveness of the CRISPR-Cas9 system for gene editing and regulation in these cell types.

ETHICAL CONCERNS



concerns about their use for human enhancement versus therapeutic purposes arise. There is a fine ethical line between using CRISPR for disease prevention and enhancement, and debates continue on whether such enhancements align with societal values (Mahmood et al., 2023). One of the most significant ethical challenges remains the potential misuse of CRISPR-Cas9 for non-therapeutic, enhancement-based applications, such as modifying traits in humans for non-medical reasons. This prospect raises questions about equity, as such enhancements may only be accessible to those who can afford them, leading to social stratification. This concern is closely tied to issues of commercialization, as private sector interests might drive CRISPR innovations toward profit-driven enhancements rather than equitable, needs-based applications (Javaid et al., 2022) (**Fig. 3**).

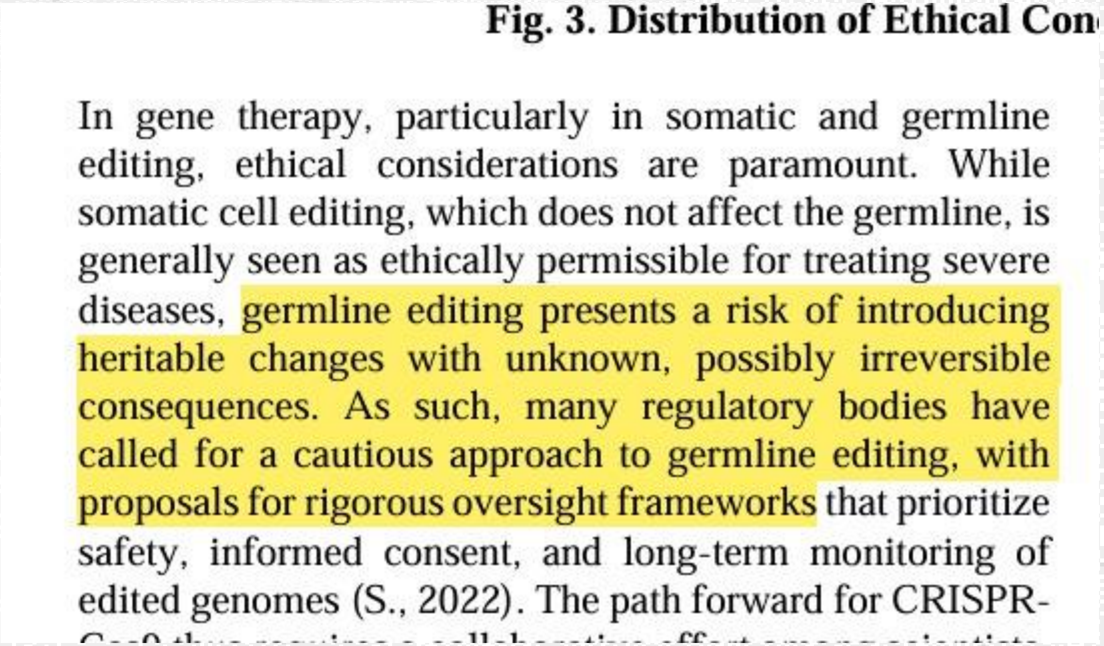
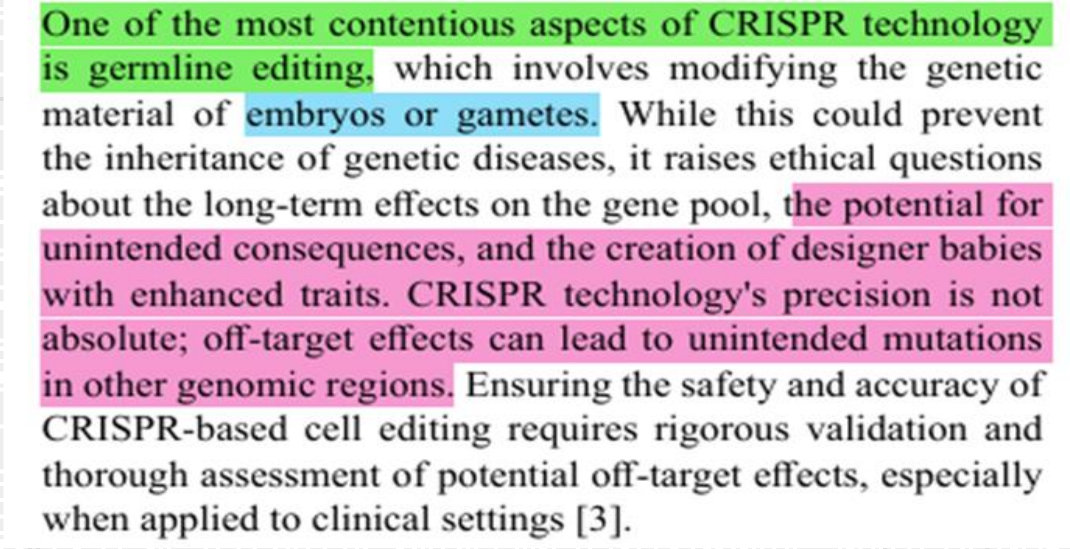


Fig. 3. Distribution of Ethical Con

In gene therapy, particularly in somatic and germline editing, ethical considerations are paramount. While somatic cell editing, which does not affect the germline, is generally seen as ethically permissible for treating severe diseases, germline editing presents a risk of introducing heritable changes with unknown, possibly irreversible consequences. As such, many regulatory bodies have called for a cautious approach to germline editing, with proposals for rigorous oversight frameworks that prioritize safety, informed consent, and long-term monitoring of edited genomes (S., 2022). The path forward for CRISPR-Cas9 thus requires a collaborative effort among scientists



One of the most contentious aspects of CRISPR technology is germline editing, which involves modifying the genetic material of embryos or gametes. While this could prevent the inheritance of genetic diseases, it raises ethical questions about the long-term effects on the gene pool, the potential for unintended consequences, and the creation of designer babies with enhanced traits. CRISPR technology's precision is not absolute; off-target effects can lead to unintended mutations in other genomic regions. Ensuring the safety and accuracy of CRISPR-based cell editing requires rigorous validation and thorough assessment of potential off-target effects, especially when applied to clinical settings [3].

REFERENCE:

- file:///C:/Users/Sanja/AppData/Local/Microsoft/Windows/INetCache/IE/CO2IQ155/ethical-considerations-in-crisprbased-cell-editing-navigating-risks-and-rewards[2].pdf

RESULTS AND COMPARISION

Model	Accuracy	Precision	Recall	F1 - Score
CNN	86 %	0.85	0.85	0.85
LSTM	84 %	0.83	0.84	0.83
FNN	87 %	0.87	0.85	0.85
Random Forest	84 %	0.83	0.83	0.83

Table 1 : Comparision between the 4 models

This Table compares the performance of four machine learning models using four key metrics.
All values are macro averages, meaning they are averaged across all classes, treating each class equally.

- FNN is the best-performing model for this task.
- CNN and Random Forest are solid alternatives.
- LSTM may be less suitable for this particular dataset or problem.

RESULTS

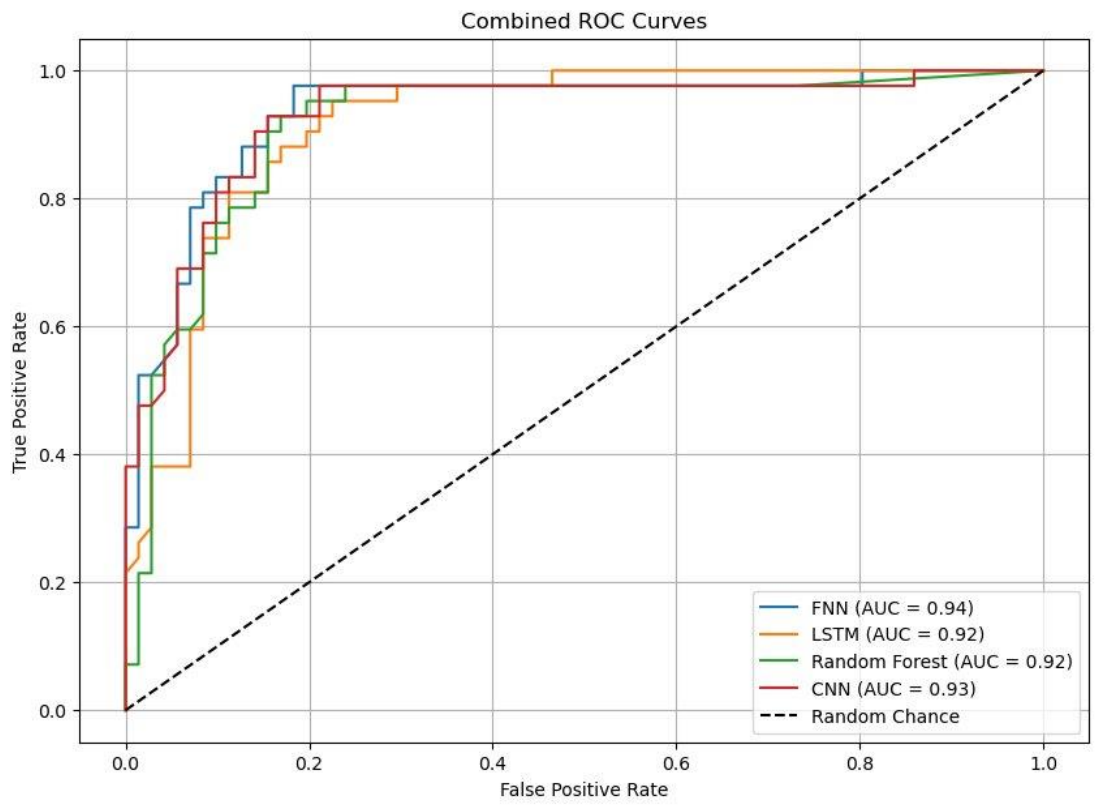


Fig. 3: Comparison of ROC curves for the four models

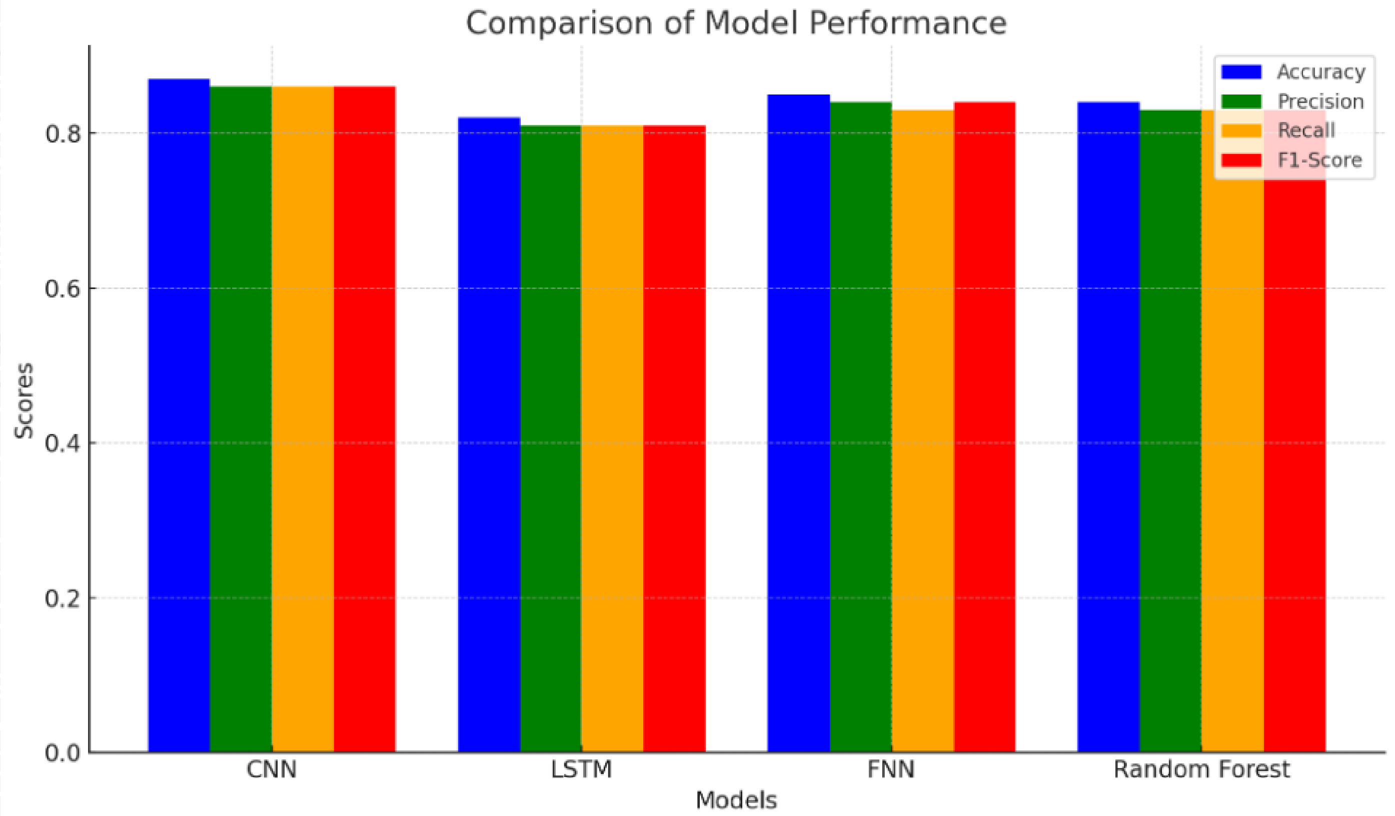


Fig. 4: Comparison of classification for the four models

CONCLUSION

Through this project, we explored how deep learning can help predict off-target effects in CRISPR-Cas9 gene editing, especially for cancer therapy. By combining real-world datasets and testing different models like FNN, LSTM, CNN, and Random Forest, we found that the FNN model gave the highest accuracy overall. Each model brought its own strengths, with LSTM and CNN showing promise in capturing sequence patterns. These results show that machine learning can play a powerful role in making gene editing safer and more precise. With more data and continued development, this approach can become an even more reliable tool in real clinical settings.

FUTURE SCOPE

- **Model Optimization**

  Fine-tune the FNN model to improve accuracy.

  Implement hyperparameter tuning and advanced regularization techniques.

- **Expanding Dataset for Cancer Therapy**

  Integrate cancer-specific CRISPR datasets from DepMap, BioGRID, TCGA, and COSMIC.

  Enhance model training using real-world oncogene mutation data.

- **Web/App Development**

  Develop a user-friendly web platform or mobile app for CRISPR off-target prediction.

  Enable researchers and doctors to input genetic sequences and receive risk assessments.