VIỆN NGHIÊN CỨU VÀ ĐÀO TẠO VIỆT - ANH, ĐẠI HỌC ĐÀ NẴNG
KHOA CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

BÁO CÁO CUỐI KỲ

DATA ANALYTICS FOR LIFE SCIENCE

<u>ĐỀ TÀI</u>: Ứng dụng deep learning (CNN-LSTM) và explainable ai (XAI) trong dự đoán hiệu quả cắt của hệ thống crispr-cas9

Giáo viên hướng dẫn: ThS. Nguyễn Chí Thiện
TS. Trần Thanh Hòa

Họ và tên sinh viên:

1. **Trần Xuân Cường……22040004**
2. **Nguyễn Vân Nhi……22040006**

MỤC LỤC

1. PROJECT OVERVIEW

1.1. Objectives

Công nghệ chỉnh sửa gen CRISPR-Cas9 đã tạo ra một cuộc cách mạng trong lĩnh vực sinh học phân tử, cho phép các nhà khoa học chỉnh sửa DNA với độ chính xác cao. Tuy nhiên, thách thức lớn nhất nằm ở việc thiết kế các đoạn RNA hướng dẫn (sgRNA) sao cho tối ưu hóa hiệu suất cắt tại đích (on-target efficiency) và giảm thiểu tác động ngoài mong muốn.
Dự án này được phát triển với các mục tiêu cốt lõi sau:

1. **Xây dựng mô hình Deep Learning hiệu suất cao:** Phát triển một kiến trúc mạng nơ-ron lai ghép (Hybrid Deep Learning) kết hợp CNN và LSTM để học các đặc trưng phức tạp từ chuỗi DNA, từ đó dự đoán chính xác điểm hiệu quả (Efficiency Score) của sgRNA.
2. **Tích hợp khả năng giải thích (Explainable AI - XAI):** Khắc phục nhược điểm "hộp đen" của các mô hình AI truyền thống bằng cách tích hợp kỹ thuật Saliency Maps. Hệ thống không chỉ đưa ra dự đoán mà còn chỉ rõ nucleotide nào đóng vai trò quyết định, giúp các nhà nghiên cứu hiểu rõ cơ chế sinh học đằng sau.
3. **Phát triển ứng dụng thực tế (Deployment):** Đóng gói mô hình vào một ứng dụng web (Streamlit App) với giao diện trực quan, cho phép người dùng nhập liệu dễ dàng, quét trình tự gen và nhận báo cáo phân tích tự động.
4. **Tự động hóa phân tích (Automated Insight):** Sử dụng kỹ thuật sinh ngôn ngữ tự nhiên (NLG) để chuyển đổi các tham số kỹ thuật thành các khuyến nghị sinh học dễ hiểu cho người dùng cuối.

1.2. Technology used

Dự án sử dụng hệ sinh thái Python với các thư viện tiên tiến nhất trong lĩnh vực Khoa học dữ liệu và Trí tuệ nhân tạo:

- **Ngôn ngữ lập trình:** Python 3.x.
- **Deep Learning Framework:** TensorFlow/Keras (xây dựng, huấn luyện và lưu trữ mô hình).
- **Xử lý & Phân tích dữ liệu:**

  - *Pandas & NumPy:* Thao tác trên khung dữ liệu bảng và tính toán ma trận số học.
  - *Scikit-learn:* Chia tập dữ liệu, chuẩn hóa và đánh giá các chỉ số thống kê (Spearman correlation).
  - *SciPy:* Hỗ trợ các tính toán khoa học chuyên sâu.
- **Web Framework:** Streamlit (Xây dựng giao diện web tương tác, Dashboard).
- **Visualization (Trực quan hóa):**
  - *Matplotlib & Seaborn:* Vẽ biểu đồ trong quá trình huấn luyện và đánh giá model.
  - *Altair:* Vẽ biểu đồ tương tác trên ứng dụng web.
- **Explainable AI:** GradientTape (tính toán đạo hàm để xây dựng Saliency Maps).

1.3. Project Structure

Cấu trúc dự án được tổ chức khoa học thành các module riêng biệt để đảm bảo tính duy trì và mở rộng:

1. **Data Processing & Training Module (CRISPR-Cas9.ipynb):**
   - Chịu trách nhiệm tải dữ liệu thô.
   - Tiền xử lý: Cắt chuỗi DNA (23bp), One-Hot Encoding.
   - Định nghĩa kiến trúc Model.
   - Huấn luyện, tinh chỉnh (Fine-tuning) và đánh giá mô hình.
2. **Application Logic & Interface (app.py):**
   - Tải mô hình đã huấn luyện (best_model.keras).
   - Xử lý logic quét chuỗi DNA (DNA Scanning Algorithm).
   - Tích hợp module XAI để giải thích dự đoán.
   - Xây dựng giao diện người dùng (UI/UX) theo phong cách "Dark Biotech".

2. ANTSBAMBOOSYSTEM (HỆ THỐNG KIẾN TRÚC)

2.1. Overall Architecture

Hệ thống được thiết kế theo luồng xử lý tuần tự khép kín (End-to-End Pipeline), đảm bảo dữ liệu đi từ dạng thô đến kết quả phân tích cuối cùng một cách mượt mà:

- **Bước 1 - Input Interface:** Người dùng cung cấp chuỗi gen đích (Target DNA Sequence).
- **Bước 2 - Pattern Matching Engine:** Hệ thống sử dụng thuật toán cửa sổ trượt (sliding window) để tìm kiếm motif PAM ("GG"). Tại mỗi vị trí tìm thấy, hệ thống trích xuất 23 nucleotide liền trước để tạo thành ứng viên sgRNA.
- **Bước 3 - Vectorization:** Các chuỗi ứng viên (A, T, G, C) được chuyển đổi sang dạng ma trận số học thông qua One-Hot Encoding.
- **Bước 4 - Inference Engine:** Ma trận đầu vào được đưa qua mô hình Deep Learning đã huấn luyện để tính toán xác suất hiệu quả (Efficiency Score).
- **Bước 5 - Interpretation Layer (XAI):** Đối với các kết quả được chọn, hệ thống kích hoạt lớp giải thích để tính toán trọng số ảnh hưởng của từng vị trí nucleotide.
- **Bước 6 - Visualization & Reporting:** Kết quả được tổng hợp thành bảng xếp hạng, biểu đồ nhiệt và báo cáo văn bản hiển thị trên Streamlit.

2.2. Models Used

Trái tim của hệ thống là một mô hình **Hybrid CNN-BiLSTM-Attention**, được thiết kế chuyên biệt để xử lý dữ liệu chuỗi sinh học.

- **Input Layer:** Chấp nhận đầu vào có kích thước (23, 4), tương ứng với chiều dài chuỗi 23bp (20bp spacer + 3bp PAM) và 4 kênh one-hot.
- **Convolutional Block (Multi-scale CNN):**
    - Sử dụng 3 nhánh Convolution 1D song song với các kích thước hạt nhân (kernel size) lần lượt là 3, 5, và 7.
    - *Mục đích:* Nhánh kernel nhỏ (3) học các motif cục bộ, trong khi nhánh kernel lớn (7) học các mẫu ngữ cảnh rộng hơn.

- Mỗi nhánh đi kèm lớp BatchNormalization để ổn định quá trình học và hàm kích hoạt ReLU để tăng tính phi tuyến.
- Kết quả từ 3 nhánh được gộp lại bằng lớp Concatenate.

- **Sequence Modeling Block (Bi-LSTM):**
  - Sử dụng lớp Bidirectional LSTM với 128 đơn vị ẩn (units).
  - *Mục đích:* Học sự phụ thuộc xa và ngữ cảnh hai chiều của chuỗi DNA (tương tác giữa các nucleotide ở đầu và cuối chuỗi).
- **Attention Mechanism (Simplified):**
  - Sử dụng lớp GlobalAveragePooling1D để nén đặc trưng, tập trung vào các thông tin quan trọng nhất trên toàn bộ chuỗi thời gian.
- **Prediction Head (Fully Connected):**
  - Các lớp Dense (128 units, 64 units) kết hợp với Dropout (0.3) để giảm thiểu hiện tượng Overfitting.
  - Lớp Output sử dụng hàm kích hoạt Sigmoid (hoặc Linear tùy chỉnh trong file huấn luyện) để trả về một giá trị thực (Regression) đại diện cho điểm hiệu quả.

3. MODEL TRAINING PROCESS

Quá trình huấn luyện mô hình được thực hiện chi tiết và nghiêm ngặt trong Jupyter Notebook, bao gồm các bước sau:

1. Data Preparation:

- Dữ liệu đầu vào bao gồm các chuỗi DNA dài 30bp. Quá trình xử lý cắt lọc lấy vùng quan trọng nhất (index 4 đến 27) để tạo thành đầu vào chuẩn 23bp.
- Thực hiện One-Hot Encoding: Chuyển đổi mỗi nucleotide thành vector 4 chiều (A=[1,0,0,0], C=[0,1,0,0], ...).

2. Data Splitting:

- Dữ liệu được chia thành 3 tập độc lập:
  - **Training Set (80%):** Dùng để cập nhật trọng số mô hình.
  - **Validation Set (10%):** Dùng để tinh chỉnh tham số và quyết định điểm dừng sớm.

- **Test Set (10%):** Dùng để đánh giá khách quan hiệu năng cuối cùng.

3. Training Setup:

- **Optimizer:** Sử dụng thuật toán Adam với tốc độ học (learning rate) khởi tạo là 0.0003 - mức tối ưu để mô hình hội tụ ổn định mà không bị dao động quá mạnh.
- **Loss Function:** Sử dụng Mean Squared Error (MSE) vì đây là bài toán hồi quy (dự đoán giá trị liên tục).
- **Evaluation Metric:** Sử dụng Mean Absolute Error (MAE) để theo dõi sai số trung bình thực tế.

4. Callbacks Strategy:

- **ModelCheckpoint:** Tự động lưu lại bộ trọng số có val_loss thấp nhất, đảm bảo mô hình cuối cùng là phiên bản tốt nhất chứ không phải phiên bản ở epoch cuối cùng.
- **EarlyStopping:** Theo dõi val_loss, nếu sau 15 epochs mà lỗi không giảm, quá trình huấn luyện sẽ dừng lại để tiết kiệm tài nguyên và tránh Overfitting.

5. Result:

- Mô hình được huấn luyện qua 100 epochs (thực tế dừng sớm hơn nhờ EarlyStopping).
- Biểu đồ Loss cho thấy sự hội tụ tốt giữa tập Train và Validation, không có dấu hiệu phân kỳ (Overfitting).

4. BUILDING THE STREAMLIT APP

4.1. Application Architecture

Ứng dụng được xây dựng trên nền tảng **Streamlit**, đóng vai trò là cầu nối giữa mô hình AI phức tạp và người dùng cuối.

- **Frontend Layer:** Sử dụng Custom CSS để tạo giao diện "Dark Mode" chuyên nghiệp. Các thành phần như nút bấm, bảng hiển thị được thiết kế với hiệu ứng Neon, tạo cảm giác công nghệ cao (Biotech/Cyberpunk style).
- **Backend Layer:** Chạy trên môi trường Python, quản lý việc tải TensorFlow model vào bộ nhớ (Caching) để tăng tốc độ phản hồi cho các lần dự đoán sau.

4.2. Main Functions

4.2.1. Batch CSV Upload (Chức năng nhập liệu hàng loạt)

Lưu ý: Trong phiên bản hiện tại, chức năng này được tích hợp thông qua việc quét chuỗi Gen dài (DNA Scanning).

- Hệ thống cho phép người dùng nhập một đoạn DNA dài (Gene Sequence) thay vì từng chuỗi ngắn lẻ tẻ.
- Thuật toán bên dưới sẽ hoạt động như một trình xử lý hàng loạt (Batch Processor): Tự động cắt chuỗi dài thành hàng trăm đoạn con (sgRNA candidates) dựa trên vị trí PAM.
- Toàn bộ danh sách này được đưa vào mô hình xử lý song song, trả về kết quả dưới dạng bảng dữ liệu có thể sắp xếp và lọc.

4.2.2. Single SMILES (Chức năng nhập liệu đơn lẻ - DNA Sequence)

Lưu ý: Dựa trên bản chất sinh học của dự án CRISPR, mục này đề cập đến việc xử lý "Single DNA Sequence".

- Người dùng nhập trực tiếp một chuỗi DNA 23bp cụ thể.
- Hệ thống kiểm tra tính hợp lệ của chuỗi (độ dài, ký tự cho phép A/T/G/C).
- Mô hình thực hiện dự đoán tức thì và trả về điểm số cụ thể kèm theo phân loại (High/Medium/Low).
- Tại đây, chức năng XAI được kích hoạt ngay lập tức để giải thích cho kết quả đơn lẻ này.

4.3. XAI Integration

- Ứng dụng tích hợp một module trực quan hóa riêng biệt cho XAI.
- Khi người dùng chọn một kết quả dự đoán, ứng dụng sẽ gọi hàm get_saliency_map.
- Kết quả trả về là một mảng giá trị trọng số, được tô màu trực tiếp lên chuỗi DNA hiển thị trên màn hình (HTML/CSS injection).

- Màu càng đậm (đỏ/cam) thể hiện nucleotide đó càng quan trọng đối với quyết định của mô hình.

5. PIPELINE VÀ EXPLAINABLE AI

5.1. End-to-End Screening Pipeline

Quy trình sàng lọc (Screening Pipeline) trong hệ thống được tự động hóa hoàn toàn:

1. **Raw Gene Input:** Nhận chuỗi gen thô (ví dụ: trình tự gen TP53 hoặc MYC).
2. **Preprocessing & Validity Check:** Loại bỏ ký tự xuống dòng, khoảng trắng, chuyển về in hoa. Kiểm tra tính hợp lệ sinh học.
3. **Candidate Discovery:** Quét tìm motif 'GG' (PAM). Lấy lùi về trước 21 ký tự để tạo chuỗi 23bp.
4. **Model Inference:** Chạy mô hình Hybrid CNN-LSTM.
5. **Post-processing:**
    - Chuyển đổi giá trị dự đoán (0-1) thành thang điểm phần trăm.
    - Xếp hạng (Ranking) dựa trên điểm số.
    - Gán nhãn chất lượng (Excellent, Good, Average, Poor).

5.2. Explainable AI (XAI)

Đây là tính năng đột phá giúp hệ thống trở thành một "Oracle" (Nhà tiên tri) thực thụ thay vì chỉ là một cỗ máy tính toán.

- **Cơ chế Gradient-based Saliency:**
    - Hệ thống sử dụng tf.GradientTape để theo dõi quá trình lan truyền ngược (Backpropagation) từ đầu ra (Score) về đầu vào (Input Embedding).
    - Công thức: $S = \left| \frac{\partial y}{\partial x} \right|$. Trong đó y là điểm hiệu quả, x là vector one-hot đầu vào.
    - Giá trị tuyệt đối của đạo hàm cho biết sự thay đổi nhỏ tại nucleotide đó sẽ làm thay đổi kết quả dự đoán nhiều như thế nào.
- **Sinh báo cáo tự động (NLG):**

- Hệ thống phân tích bản đồ nhiệt để sinh ra các câu giải thích ngôn ngữ tự nhiên.
- **Phân tích vùng Seed (13-20):** Nếu trọng số tập trung cao ở đây, hệ thống báo cáo: *"Mô hình đánh giá cao khả năng bắt cặp bổ sung tại vùng Seed, đảm bảo độ đặc hiệu."*
- **Phân tích vùng PAM (21-23):** Nếu trọng số cao ở 'GG', hệ thống xác nhận: *"Tín hiệu PAM rõ ràng, đảm bảo Cas9 nhận diện vị trí cắt chính xác."*

6. RESULTS AND EVALUATION

Kết quả đánh giá trên tập dữ liệu kiểm thử (Test Set) cho thấy hiệu năng ấn tượng của mô hình:

- **Độ chính xác tương quan (Spearman Correlation):**
  - Mô hình đạt chỉ số Spearman $\rho \approx 0.66$ trên tập Test.
  - Trong lĩnh vực dự đoán CRISPR off-target/on-target, chỉ số $\rho > 0.6$ được coi là kết quả tốt, cho thấy mô hình học được thứ tự xếp hạng của các sgRNA rất sát với thực nghiệm sinh học.
- **Biểu đồ phân tán (Scatter Plot):**
  - Trục hoành (True Values) và trục tung (Predicted Values) cho thấy sự phân bố điểm tập trung quanh đường chéo $y=x$. Điều này chứng tỏ mô hình không bị lệch (bias) quá nhiều về phía giá trị cao hoặc thấp.
- **Case Study thực tế:**
  - Thử nghiệm trên gen đích thực tế, hệ thống đã lọc ra được Top 5 vị trí cắt có điểm số > 0.8. Các vị trí này khi phân tích XAI đều hiển thị vùng Seed Region rất "sáng" (trọng số cao), phù hợp với lý thuyết sinh học về cơ chế cắt của Cas9.

7. CONCLUDE

Dự án **ANTSBAMBOOSYSTEM - CRISPR Efficiency Oracle** đã hoàn thành xuất sắc các mục tiêu đề ra ban đầu, mang lại một giải pháp công nghệ trọn vẹn cho bài toán thiết kế sgRNA.

Tổng kết các đóng góp chính:

1. Đề xuất thành công kiến trúc mạng **Hybrid CNN-BiLSTM**, chứng minh khả năng học đặc trưng chuỗi gen vượt trội so với các phương pháp truyền thống.
2. Giải quyết vấn đề minh bạch hóa AI thông qua **XAI (Saliency Maps)**, giúp xây dựng niềm tin cho người dùng là các nhà sinh học/nghiên cứu viên.
3. Ứng dụng **Streamlit** hoạt động ổn định, giao diện thân thiện, tích hợp đầy đủ quy trình từ nhập liệu đến báo cáo, sẵn sàng cho việc triển khai thực tế.

Hướng phát triển trong tương lai:

- Mở rộng tập dữ liệu huấn luyện để bao gồm nhiều loại tế bào và điều kiện thí nghiệm khác nhau nhằm nâng cao độ chính xác tổng quát.
- Tích hợp thêm module dự đoán **Off-target** (cắt nhầm) để cung cấp cái nhìn toàn diện hơn về độ an toàn của sgRNA.
- Phát triển phiên bản API để tích hợp vào các pipeline tin sinh học lớn hơn.