

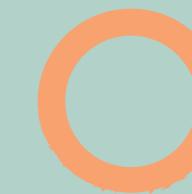


DS106.P11 - Tối ưu hóa & Ứng dụng

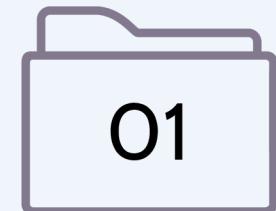
# Phân Loại Rối Loạn Nhịp Tim dựa trên GA Stacking trong Học Máy Kết hợp

Nhóm 05:

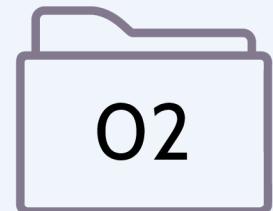
- Nguyễn Lê Vy - 21522811
- Nguyễn Thị Mai Trinh - 21522718



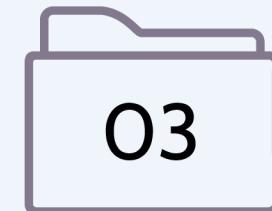
# NỘI DUNG BÁO CÁO:



GIỚI THIỆU



NGHIÊN CỨU LIÊN  
QUAN



BỘ DỮ LIỆU



PHƯƠNG PHÁP  
THỰC NGHIỆM



KẾT QUẢ THỰC  
NGHIỆM



KÊT LUẬN

BACK

NEXT

# I. GIỚI THIỆU BÀI TOÁN

Chúng tôi xây dựng một mô hình hybrid kết hợp CNN\_1D, Gradient Boosting, XGBoost, RandomForest và SVC, với meta-model là Logistic Regression được tối ưu hóa bằng GA/ GridSearch, để phân loại rối loạn nhịp tim. Nhóm kỳ vọng mô hình sẽ vượt trội hơn so với các phương pháp đơn lẻ, qua đó thể hiện tiềm năng của GA Stacking trong việc tối ưu hóa mô hình kết hợp.

- **INPUT:** Đặc trưng ECG từ bộ dữ liệu rối loạn nhịp tim.
- **OUTPUT:** Dự đoán các loại rối loạn nhịp tim (13 nhãn)

## II. NGHIÊN CỨU LIÊN QUAN

Các nghiên cứu trước đây sử dụng các mô hình riêng lẻ như CNN, SVM, LSTM để phân loại rối loạn nhịp tim nhưng chưa khai thác hết tiềm năng của các kỹ thuật học máy đối với các tập dữ liệu phức tạp.

Phương pháp kết hợp, như GA-Stacking, kết hợp các ưu điểm của các mô hình, giúp cải thiện độ chính xác và khả năng khái quát hóa. Các nghiên cứu như "Improved Stacked Ensemble with Genetic Algorithm for Automatic ECG Diagnosis" (2023) đã chứng minh hiệu quả vượt trội của GA-Stacking trong phát hiện bất thường ECG.

Dự án này áp dụng những tiến bộ trên nhằm nâng cao khả năng phát hiện rối loạn nhịp tim.

### III. BỘ DỮ LIỆU

#### BỘ DỮ LIỆU CARDIAC ARRHYTHMIA

Nhóm sử dụng bộ dữ liệu Rối loạn Nhịp Tim, cung cấp trên nền tảng UCI, được giới thiệu lần đầu trong bài báo “A Supervised Machine Learning Algorithm for Arrhythmia Analysis” (1997).

Bộ dữ liệu phục vụ mục đích xác định, nhận diện và phân loại các loại rối loạn nhịp tim khác nhau. Dữ liệu bao gồm 452 hồ sơ bệnh nhân với 279 đặc trưng, bao gồm các thông tin như tuổi, giới tính, chiều cao, cân nặng, nhịp tim và các thông số ECG như sóng Q, R, S trên các kênh DI, DII, DIII và các kênh khác. Khoảng 0,33% giá trị đặc trưng bị thiếu.

### III. BỘ DỮ LIỆU

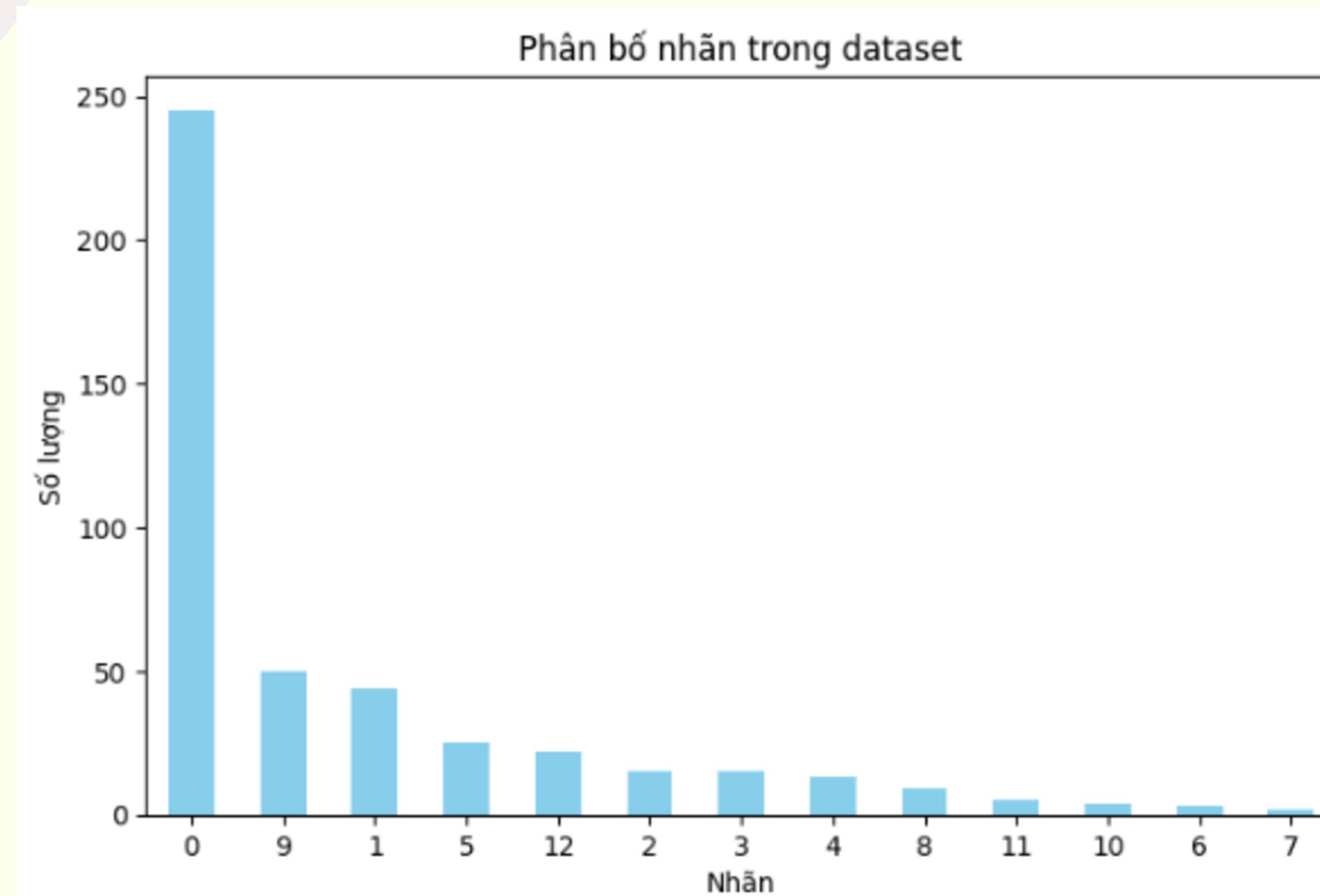
#### TIỀN XỬ LÝ DỮ LIỆU

Sau bước tiền xử lý, bộ dữ liệu giữ lại 452 hàng và 279 đặc trưng, 1 nhãn. Biến mục tiêu "diagnosis" đã được giảm từ 16 xuống còn 13 danh mục:

- '0': Chỉ số ECG bình thường
- '1': Biến đổi do bệnh động mạch vành
- '2': Nhồi máu cơ tim vùng trước
- '3': Nhồi máu cơ tim vùng sau
- '4': Nhịp xoang nhanh
- '5': Nhịp xoang chậm
- '6': Ngoại tâm thu thất
- '7': Ngoại tâm thu nhĩ
- '8': Block nhánh trái
- '9': Block nhánh phải
- '10': Phì đại thất trái
- '11': Rung nhĩ/ cuồng nhĩ
- '12': Các trường hợp khác

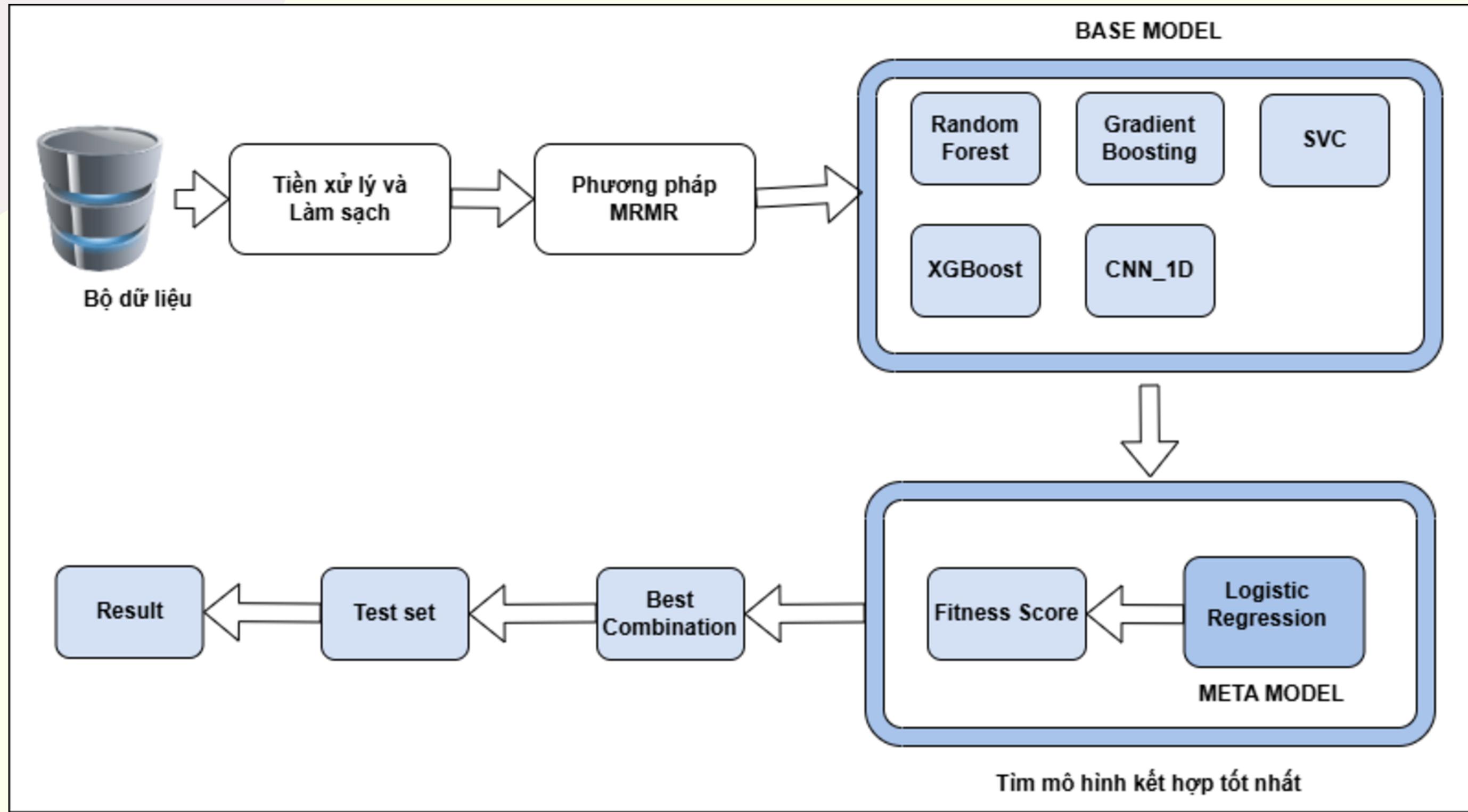
### III. BỘ DỮ LIỆU

Bộ dữ liệu bị mất cân bằng, số lượng các nhãn chênh lệch nhau khá lớn. Để giải quyết vấn đề này nhóm sử dụng phương pháp Đánh trọng số Nhãn.



# IV. PHƯƠNG PHÁP THỰC NGHIỆM

## QUY TRÌNH



# PHƯƠNG PHÁP MRMR

- MAXIMUM RELEVANCE (TỐI ĐA HÓA TÍNH LIÊN QUAN):

$$\text{Relevance} = \frac{1}{|S|} \sum_{x_i \in S} I(x_i, y)$$

Trong đó:

- $I(x_i, y)$ : Mutual Information giữa đặc trưng  $x_i$  và biến mục tiêu  $y$ .

- MINIMUM REDUNDANCY (TỐI THIỂU HÓA TÍNH DƯ THỪA):

$$\text{Redundancy} = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j)$$

Trong đó:

- $I(x_i, x_j)$ : Mutual Information giữa hai đặc trưng  $x_i$  và  $x_j$ .

# PHƯƠNG PHÁP MRMR

## • HÀM MỤC TIÊU MRMR

MIQ (Mutual Information Quotient):

$$\text{Score} = \frac{\text{Relevance}}{\text{Redundancy}}$$

Công thức của Mutual Information giữa hai biến rời rạc  $x_i$  và  $y$  là:

$$I(x_i, y) = \sum_{x_i \in X} \sum_{y \in Y} P(x_i, y) \log \frac{P(x_i, y)}{P(x_i)P(y)}$$

Trong đó:

- $P(x_i, y)$  là xác suất đồng thời của  $x_i$  và  $y$ , tức là xác suất mà  $x_i$  và  $y$  đồng thời xảy ra.
- $P(x_i)$  là xác suất biên của  $x_i$ , tức là xác suất của  $x_i$  mà không quan tâm đến  $y$ .
- $P(y)$  là xác suất biên của  $y$ , tức là xác suất của  $y$  mà không quan tâm đến  $x_i$ .

# IV. PHƯƠNG PHÁP THỰC NGHIỆM

## THAM SỐ CỦA CÁC MÔ HÌNH

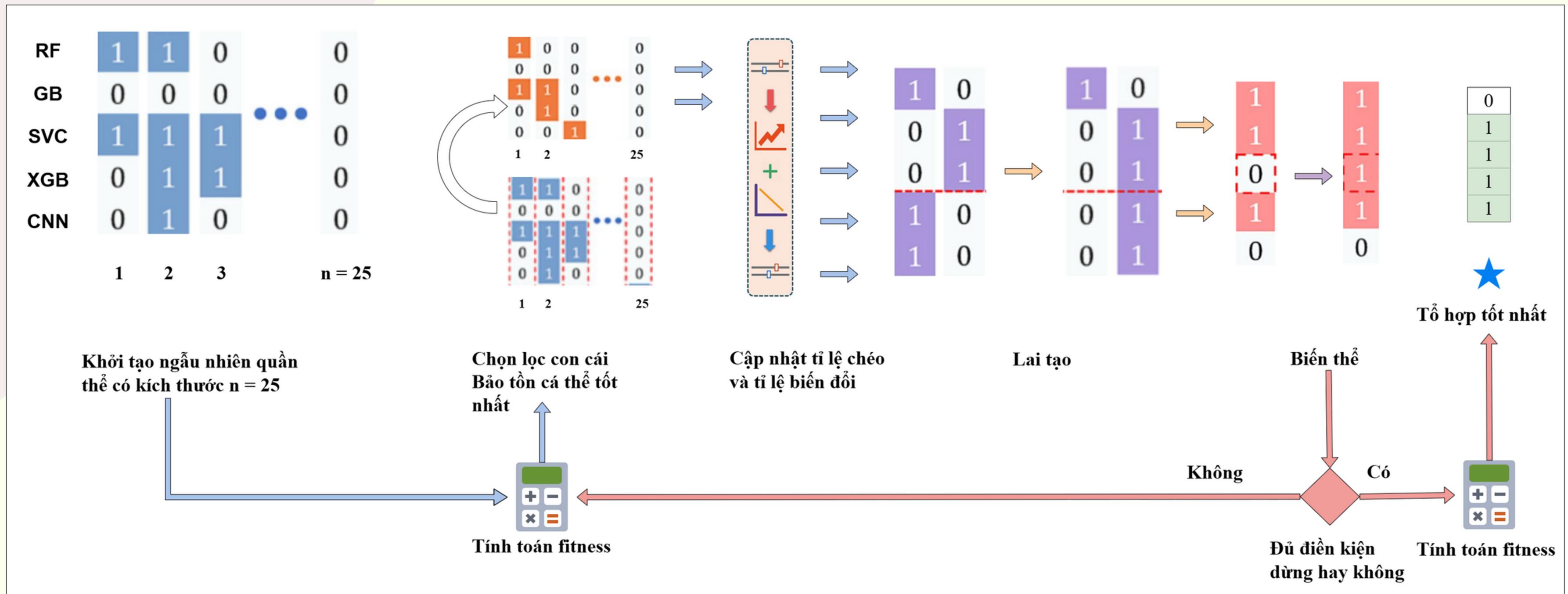
- **Random Forest:** n\_estimators=100, random\_state=42
- **Gradient Boosting:** Sử dụng tham số mặc định
- **SVC:** probability=True, random\_state=42
- **XGBoost:** objective='multi:softmax', num\_class=13
- **CNN:** optimizer Adam, lr=0.001, weight\_decay=1e-4 và loss function CrossEntropyLoss với trọng số lớp (class\_weights)

**GA:** population\_size=25, max\_generations=10, mutation\_rate=0.1, crossover\_rate=0.7

# IV. PHƯƠNG PHÁP THỰC NGHIỆM

## TÌM MÔ HÌNH KẾT HỢP TỐT NHẤT DỰA VỚI GA

Fitness: được tính dựa trên độ đo F1-Score.



## IV. PHƯƠNG PHÁP THỰC NGHIỆM

### TÌM MÔ HÌNH KẾT HỢP TỐT NHẤT VỚI GRID SEARCH

Grid Search là phương pháp tìm kiếm toàn diện để xác định tổ hợp mô hình tối ưu cho bài toán. Nó thử tất cả các kết hợp có thể của các mô hình trong không gian mô hình đã định trước và đánh giá hiệu suất của mỗi tổ hợp thông qua cross-validation. Mục tiêu là tìm ra sự kết hợp mô hình cho bài toán mang lại kết quả tốt nhất, ví dụ: Giá trị fitness (tính dựa trên chỉ số F1-score) cao nhất.

Nhóm thực hiện đồng thời cả hai phương pháp: Grid Search và GA trong giai đoạn tìm kiếm tổ hợp mô hình tối ưu, nhằm so sánh hiệu suất của chúng và lựa chọn phương pháp tốt nhất cho nhiệm vụ này.

# V. KẾT QUẢ THỰC NGHIỆM

## ĐỘ ĐO ĐÁNH GIÁ

- **Accuracy:** Đo lường tỷ lệ các trường hợp được phân loại đúng so với tổng số trường hợp.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Instances}$$

- **Precision:** Đánh giá tỷ lệ dự đoán đúng (true positive) trong tất cả các dự đoán dương tính.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- **Recall:** Còn được gọi là Độ nhạy, đo lường khả năng nhận diện tất cả các trường hợp liên quan.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- **F1-Score:** Giá trị trung bình điều hòa của Precision và Recall, cung cấp sự cân bằng giữa chúng.

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

# PHƯƠNG PHÁP LỰA CHỌN MODEL

## GA

Cấu hình tối ưu gồm Gradient Boosting, SVC, XGBoost, và CNN. Thuật toán di truyền đã xác định tổ hợp này với giá trị fitness cao nhất 0.9708, chứng minh hiệu quả và độ chính xác cao của mô hình kết hợp.

```
Generation 0: Best Fitness = 0.9202508710875916
Generation 1: Best Fitness = 0.9217869118938206
Generation 2: Best Fitness = 0.9261922566997628
Generation 4: Best Fitness = 0.9483358014505993
Generation 5: Best Fitness = 0.9547231617662518
Generation 6: Best Fitness = 0.9612712206594701
Generation 7: Best Fitness = 0.9674576786065011
Generation 8: Best Fitness = 0.970751385929383
Best Solution Found: [0 1 1 1 1]
Best Fitness Achieved: 0.970751385929383
```

- 
- 

## PHƯƠNG PHÁP LỰA CHỌN MODEL

### GRID SEARCH

Cấu hình tối ưu gồm cả 5 mô hình: RandomForest, Gradient Boosting, SVC, XGBoost, và CNN. Mô hình Grid Search đã xác định tổ hợp này với giá trị fitness cao nhất 0.723.

```
Grid Search - Best Solution: (1, 1, 1, 1, 1)
Grid Search - Best Fitness: 0.7235154091138928
Thời gian chạy: 231.514844 giây
```

# KẾT QUẢ THỰC NGHIỆM

Bảng 1: Kết quả các mô hình trên tập test

Mô hình	RF	GB	SVC	XGB	CNN	Ensemble model GA	Ensemble model GS
Accuracy	0.71	0.72	0.60	0.72	0.61	0.70	<u>0.71</u>
Precision	0.40	0.49	0.22	0.50	0.33	0.45	<u>0.48</u>
F1-Score	0.34	0.42	0.16	0.43	0.29	0.37	<u>0.40</u>
Recall	0.36	0.44	0.15	0.45	0.28	0.39	<u>0.43</u>

## VI. KẾT LUẬN

Chúng tôi đã áp dụng GA-Stacking cùng GridSearch để xây dựng mô hình kết hợp nhằm phân loại rối loạn nhịp tim trên bộ dữ liệu Cardiac Arrhythmia, với mục tiêu phân loại chính xác các loại rối loạn nhịp tim. Mô hình kết hợp CNN\_1D, Gradient Boosting, XGBoost, Random Forest và SVC, cùng với meta-model (LR) và GA Stacking/ GridSearch để tối ưu hóa các dự đoán.

Kết quả hiện tại chưa đạt như kỳ vọng, nhóm sẽ tiếp tục cải tiến mô hình. Mục tiêu là tận dụng GA-Stacking để cải thiện việc trích xuất đặc trưng, giảm lỗi phân loại và nâng cao hiệu suất mô hình.

**CẢM ƠN THẦY GIÁO VÀ  
CÁC BẠN ĐÃ CHÚ Ý LẮNG  
NGHE!**