

# BÁO CÁO ĐỒ ÁN CUỐI KỲ

Lớp: CS114.(K21+K21.KHTN)

Môn: MÁY HỌC

GV: PGS.TS Lê Đình Duy - THS. Phạm Nguyễn Trường An  
Trường ĐH Công Nghệ Thông Tin, ĐHQG-HCM

# PHÂN LOẠI HOA TRONG ẢNH CHỤP BÔNG HOA

**Phạm Mạnh Tiến - 18520166 - CS114.K21.KHTN**

**Link github: <https://github.com/tienpm/CS114.K21.KHTN>**

# Tóm tắt

- Tên đề tài: Phân loại hoa trong ảnh chụp bông hoa
- Tóm tắt về đề án và kết quả đạt được

Đề án cố gắng giải bài toán phân loại các loài hoa trong ảnh chụp bông hoa trong môi trường tự nhiên hoặc được trồng trong các vườn hoa. Model thuật toán đang bị overfitting và cần phải được tinh chỉnh thêm

- Ảnh của các thành viên của nhóm



**Phạm Mạnh Tiến - MSSV: 18520166**

# I. Giới thiệu bài toán

Bài toán phân loại ảnh là một trong các bài toán trong Computer Vision đặt ra nhiều thách thức trong ứng dụng thực tế. Phân loại hoa trong ảnh có thể giúp cho việc giám sát sự phát triển cây hoa trong các vườn hoa cây cảnh, giúp cho việc tìm kiếm và phân loại hoa trong cửa hàng hoặc vườn hoa cảnh nhanh chóng hơn. Trong đồ án này, sinh viên tìm hiểu cách sử dụng những kiến thức đã học trong môn Máy học để thiết kế một hệ thống phân loại loài hoa dựa vào ảnh chụp bông hoa.

## II. MÔ TẢ TẬP DỮ LIỆU

- Tập dữ liệu được xây dựng bằng cách chụp ảnh bông hoa trong môi trường tự nhiên với bạn Nguyễn Quyết Thắng - MSSV: 18520166.
- Tập dữ liệu gồm 1200 thuộc 12 loài hoa: hoa Dừa Cạn, hoa Bông Trang, hoa Hồng, hoa Giấy, hoa Chi Cúc, hoa Lan Hồ Điệp, hoa Sứ, hoa Huỳnh Anh, hoa Cúc, hoa Mào Gà, hoa Chiều Tím, hoa Đồng Tiền, 100 ảnh cho mỗi loài hoa.
- Các điểm thu thập dữ liệu: Ký túc xá khu B - ĐHQG Hồ Chí Minh (hoa Chiều Tím); cổng trước UIT (hoa Giấy); các vườn hoa cây cảnh ở đường Phạm Văn Đồng (hoa Lan Hồ Điệp) và đường Kha Vạm Cân (hoa Dừa Cạn, hoa Bông Trang, hoa Hồng, hoa Chi Cúc) quận Thủ Đức; khu đô thị Vạn Phúc, Hiệp Bình Phước, quận Thủ Đức (hoa Sứ, hoa Huỳnh Anh); công viên Gia Định quận Phú Nhuận (hoa Mào Gà, hoa Cúc, hoa Đồng Tiền).
- Thách thức: dữ liệu được thu thập trong nhiều điều kiện thời tiết khác nhau đặc biệt là lúc trời mưa nhỏ hoặc mới mưa xong cánh hoa của bông hoa thường bị thấm nước hoặc nước mưa bám nhiều trên cánh hoa (tuy nhiên không được rửa nước ở cánh hoa, tránh việc bông hoa của chủ vườn hoa bị gãy hoặc tổn hại). Việc này gây ra nhiễu trong ảnh thu thập được và gây khó khăn cho việc rút trích đặc trưng

## II. MÔ TẢ TẬP DỮ LIỆU



Cánh hoa bị thấm nước mưa làm ảnh hưởng  
màu sắc của hoa



Cánh hoa còn đọng lại nhiều giọt nước mưa

# III. TIỀN XỬ LÝ VÀ RÚT TRÍCH ĐẶC TRƯNG

- Tiền xử lý ảnh: Ảnh chụp bằng điện thoại nên thường có kích thước lớn làm ảnh hưởng đến thời gian các thuật toán rút trích đặc trưng vì vậy nhóm đã resize ảnh lại thành kích thước 500 x 500. Các thuật toán xử lý ảnh để rút trích đặc trưng yêu cầu input đầu vào là ảnh Binary (ảnh trắng đen) hoặc ảnh theo dạng HSV nên ta phải chuyển hệ màu của ảnh trước khi đưa vào thuật toán

- Sử dụng các thuật toán xử lý ảnh để rút trích đặc trưng bao gồm:
  - Hu Moments: Để rút trích hình dạng của bông hoa
  - Halarick Texture: Để rút trích hình dạng vân của cánh hoa
  - Color Histogram: Để rút trích màu sắc của bông hoa

# III. TIỀN XỬ LÝ VÀ RÚT TRÍCH ĐẶC TRƯNG

Color  
Color Histogram



Texture  
Haralick



Shape  
Moments



Minh họa cho các thuật toán xử lý ảnh dùng để rút trích đặc trưng



# IV. MÔ HÌNH MACHINE LEARNING

- Vì tập dữ liệu nhỏ chỉ 1200 ảnh nên nhóm chia tập dữ liệu thành 80% test và 20% test kết hợp với việc sử dụng phương pháp

Cross-validation, cách thường dùng sử dụng là chia tập training ra k tập con không có phần tử chung, có kích thước gần bằng nhau. Tại mỗi lần kiểm thử, được gọi là *run*, một trong số k tập con được lấy ra làm validate set. Mô hình sẽ được xây dựng dựa vào hợp của k-1 tập con còn lại. Mô hình cuối được xác định dựa trên trung bình của các train error và validation error. Cách làm này còn có tên gọi là k-fold cross validation. Nhóm sử dụng phương pháp này với hy vọng tránh được việc model bị overfitting do tập dữ liệu huấn luyện nhỏ.

# IV. MÔ HÌNH MACHINE LEARNING

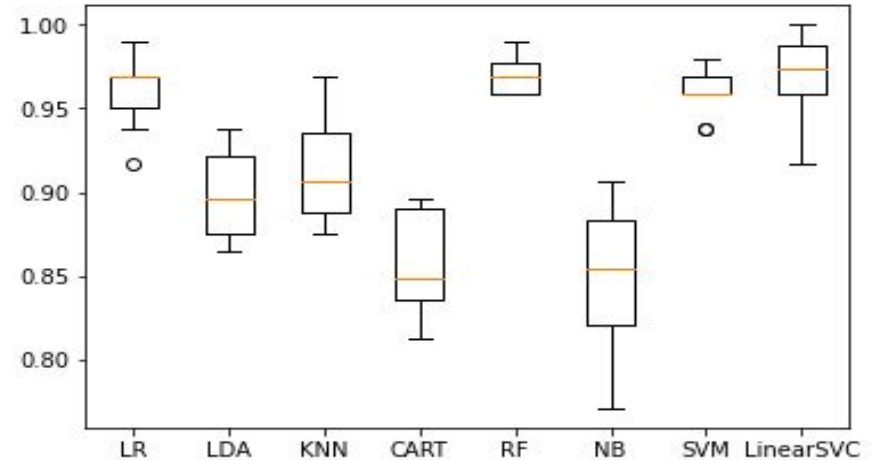
- Nhóm tiến hành huấn luyện tập dữ liệu trên nhiều model classification như: Logistic Regression, Linear Discriminant Analysis, K-Neighbors Classifier, Decision Tree, Random Forest, Naive Bayes, SVM, LinearSVC để xác định độ chính xác (accuracy) của từng model trên tập train, nhằm chọn ra model có độ chính xác cao nhất để đánh giá trên tập dữ liệu mới

# V. KẾT QUẢ

- Random Forest Classifier cho độ chính xác (accuracy) lớn nhất ~97%.  
Vì vậy nhóm đã tiến hành dùng Random Forest Classifier dự đoán trên tập dữ liệu mới chưa được đưa vào huấn luyện.

LR: 0.960417 (0.020199)  
LDA: 0.896875 (0.026125)  
KNN: 0.912500 (0.030262)  
CART: 0.858333 (0.028792)  
RF: 0.968750 (0.010417)  
NB: 0.847917 (0.044488)  
SVM: 0.959375 (0.012715)  
LinearSVC: 0.968750 (0.023292)

Machine Learning algorithm comparison



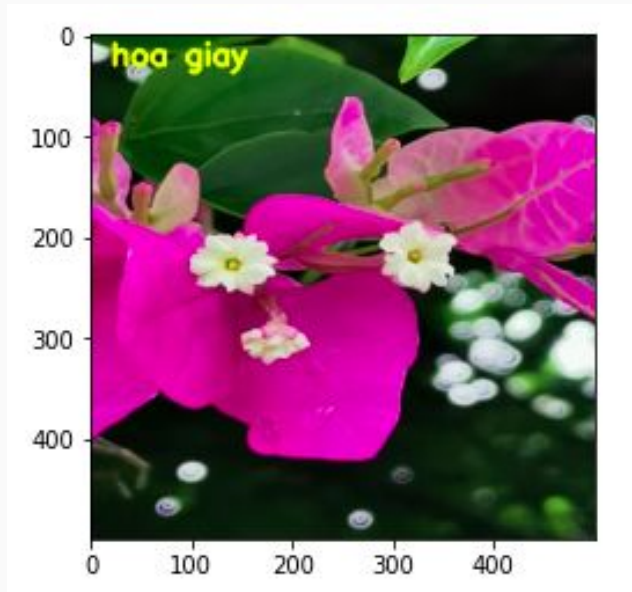
# V. KẾT QUẢ

- Tập dữ liệu mới được xây dựng bằng cách cắt một phần ảnh và xoay ảnh trong tập dữ liệu thu thập được. Tập dữ liệu này cũng có 12 class với 120 ảnh với 10 ảnh mỗi class. Kết quả Model chỉ có thể dự đoán chính xác một số ảnh của 3 loài hoa là hoa Dừa Cạn, hoa Hồng và hoa Giấy trên tổng số 12 loài hoa. Chứng tỏ mô hình bị overfitting và cần phải tinh chỉnh để có kết quả tốt hơn.

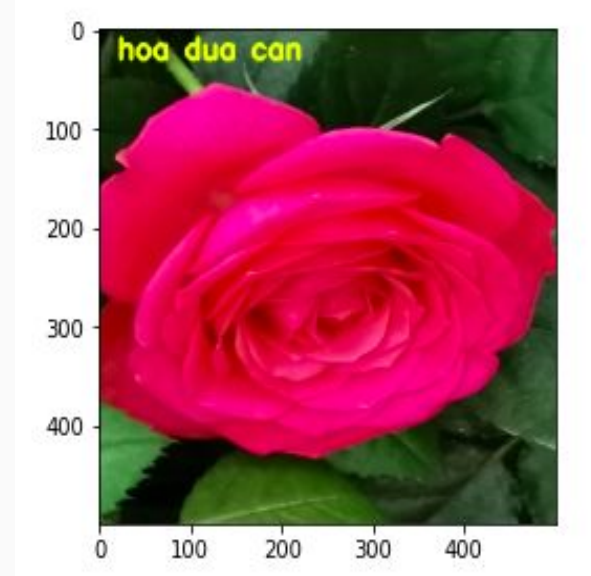
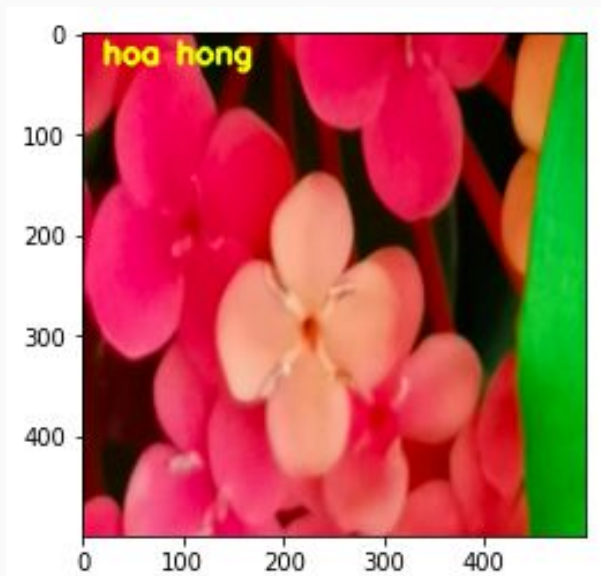
|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0            | 0.00      | 0.00   | 0.00     | 10      |
| 1            | 0.00      | 0.00   | 0.00     | 10      |
| 2            | 0.00      | 0.00   | 0.00     | 10      |
| 3            | 0.00      | 0.00   | 0.00     | 10      |
| 4            | 0.00      | 0.00   | 0.00     | 10      |
| 5            | 0.18      | 0.20   | 0.19     | 10      |
| 6            | 0.10      | 0.80   | 0.17     | 10      |
| 7            | 0.07      | 0.20   | 0.11     | 10      |
| 8            | 0.00      | 0.00   | 0.00     | 10      |
| 9            | 0.00      | 0.00   | 0.00     | 10      |
| 10           | 0.00      | 0.00   | 0.00     | 10      |
| 11           | 0.00      | 0.00   | 0.00     | 10      |
| accuracy     |           |        | 0.10     | 120     |
| macro avg    | 0.03      | 0.10   | 0.04     | 120     |
| weighted avg | 0.03      | 0.10   | 0.04     | 120     |

Kết quả sau khi fit dữ liệu mới chưa từng được train vào model Random Forest Classifier đã được train trước đó

# V. KẾT QUẢ



Mô hình dự đoán đúng với ảnh được đưa vào ngẫu nhiên (trường hợp rất hiếm)



Mô hình dự đoán sai

## VI. KẾT LUẬN

- Model Random Forest được chọn bị overfittinng hoàn toàn. Vì giá trị precision và recall của rất nhiều class có giá trị 0 (dự đoán sai với toàn bộ ảnh trong test set)
- Kết quả này là vô nghĩa nên hệ thống không thể áp dụng trong thực tế. Vì vậy đề án cần phải được cải tiến thêm.

## VII. HƯỚNG PHÁT TRIỂN

- Mở rộng tập dữ liệu với nhiều ảnh hơn
- Nghiên cứu các phương pháp rút trích đặc trưng bằng CNN và các phương pháp xử lý ảnh khác.
- Chính sửa tham số của các thuật toán máy học.