

# BÁO CÁO ĐỒ ÁN CUỐI KỲ

0

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

# **NHẬN DIỆN HOA TRONG ẢNH CHỤP BÔNG HOA**

**Nguyễn Quyết Thắng - 18520152 -  
CS114.K21.KHTN**

**Link Github:**

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

# Tóm tắt

- Đề tài: Nhận diện hoa trong ảnh chụp bông hoa
- Sử dụng các thuật toán tiền xử lý, rút trích đặc trưng và máy học để giải quyết bài toán.
- Ảnh của các thành viên của nhóm



Nguyễn Quyết Thắng

# Bài toán

Thi giác máy tính máy tính được ứng dụng rất nhiều trong thực tế và một trong những ứng dụng của nó là phân loại hình ảnh. Đò án nhận diện hoa trong ảnh chụp bông hoa thuộc vào ứng dụng phân loại ảnh của thi giác máy tính. Input đầu vào sẽ là hình ảnh một loài hoa và output sẽ là tên loài hoa tương ứng.

# Mô tả dữ liệu

Bộ dữ liệu được em và Phạm Mạnh Tiến thu thập ở rải rác các khu vực trong quận Thủ Đức và Thành phố Hồ Chí Minh. Bao gồm 1200 ảnh tương ứng với 12 loài hoa(mỗi loài có 100 tấm) khác nhau gồm: hoa dừa cạn, hoa bông trang, hoa hồng, hoa giấy, hoa chi cúc, hoa lan hồ điệp, hoa sứ, hoa huỳnh anh, hoa mào gà, hoa cúc, hoa đồng tiền, hoa chiều tím được thu thập ở đường Kha Vạn Cân, đường Phạm Văn Đồng, khu đô thị Vạn Phúc, công viên Gia Định, trước cổng trường và kí túc xá.



# Mô tả dữ liệu

Khó khăn: các bông hoa được thu thập vào các điều kiện thời tiết khác nhau. Đặc biệt là lúc trời mưa khiến cánh hoa bị thấm nước hoặc có nước đọng làm sai lệch màu sắc của hoa.



# Tiền xử lý dữ liệu và rút trích đặc trưng

Tiền xử lý dữ liệu:

- Chuyển ảnh từ ảnh RGB sang HSV.
- Resize lại ảnh về kích thước 500\*500.

Trích xuất đặc trưng:

- Sử dụng Hu moment để lấy hình dáng của bông hoa.
- Sử dụng Color histogram để lấy màu hoa.
- Sử dụng Haralick texture để lấy vân hoa.

# Phân chia dataset

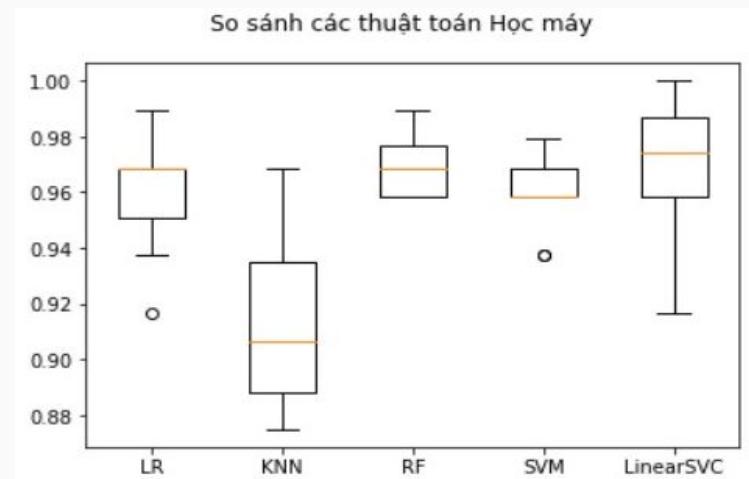
Em chia data thành 2 phần:

- Phần 1 gồm 960 ảnh dùng làm training set.
- Phần 2 gồm 240 ảnh dùng làm test set.

# Chọn mô hình thuật toán máy học

Với bài toán Classifications, ta có nhiều thuật toán phân loại như Linear Classifier, SVM, SRC,... Với đồ án này, em chọn các thuật toán gồm Logistic Regression, KNN, Random Forest, SVM, Linear SVC để giải quyết bài toán.

Sau khi train thì em so sánh giữa các thuật toán và thấy thuật toán Random Forest cho ra kết quả tốt nhất với tỷ lệ đoán đúng là gần 97% trên tập test.



# Kết quả và đánh giá

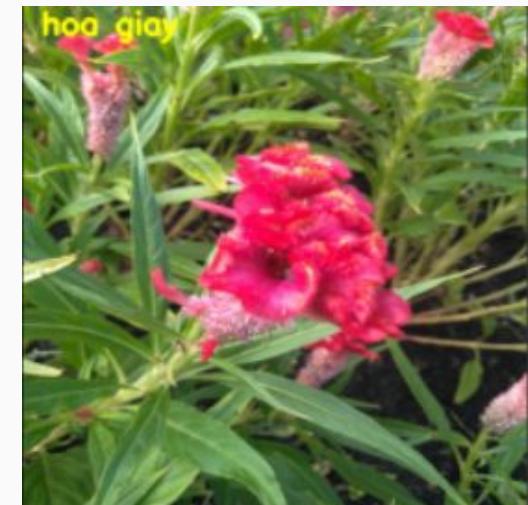
Em sử dụng model mới được train để chạy trên tập dữ liệu mới được xây dựng bằng cách crop lại ảnh. Kết quả sau khi em sử dụng model của Random Forest để chạy thì mô hình bị overfitting, cho kết quả sai trên đa số các ảnh trong testset.



kết quả đúng



kết quả sai



# Kết quả và đánh giá

Lí do bị overfitting là do:

- Dữ liệu còn quá ít
- Rút trích đặc trưng vẫn chưa ổn

Vì vậy nên dự định tiếp theo để thuật toán chính xác hơn ta cần:

- Thu thập thêm lượng lớn dữ liệu hoa nhằm tăng độ đa dạng của các loài hoa
- Tìm hiểu thêm về cách thuật toán rút trích đặc trưng.