XÂY DỰNG BỘ DỮ LIỆU CHO ỨNG DỤNG GỢI Ý HOA

Lê Vy[1,2,3], **Lưu Bảo Uyên**[1,2,4], **Trần Lương Vân Nhi**[1,2,5], **Hồ Nguyễn Thiên Vũ**[1,2,6], **Trần Quốc Khánh**[1,2], **Nguyễn Gia Tuấn Anh**[1,2]

[1] *Trường Đại học Công nghệ Thông tin, Thành phố Hồ Chí Minh, Việt Nam*

[2] *Đại học Quốc gia, Thành phố Hồ Chí Minh, Việt Nam*

[3] *22521703@mail.edu.vn*

[4] *22521640@mail.edu.vn*

[5] *22521044@gm.edu.vn*

[6] *22521689@gm.edu.vn*

Abstract

Trong bài báo cáo này trình bày quá trình thu thập và tiền xử lý bộ dữ liệu hình ảnh và về các loại hoa trang trí. Bộ dữ liệu này nhằm mục đích phục vụ cho hệ khuyến nghị "gợi ý hoa trang trí". Nhóm tôi đã thực hiện thu thập một tập dữ liệu ban đầu, sau đó áp dụng các phương pháp tiền xử lý và cuối cùng huấn luyện các mô hình học máy để thực hiện dự đoán. Chúng tôi đề xuất các bộ lọc Sobel, Prewitt, Scharr kết hợp mô hình Random Forest và Support Vector Machine sau đó dùng các độ đo để đánh giá chất lượng. Kết quả cho thấy chúng tôi đã thành công ghi nhận dự đoán một bộ dữ liệu cơ bản bao gồm 14 loài hoa được gán nhãn tên khoa học và màu hoa với độ chính xác cao nhất là 60%. Quá trình này giúp cải thiện chất lượng bộ dữ liệu hình ảnh, có thể sử dụng cho các mục đích khai thác và phát triển tiếp theo.

Từ khóa - phân loại hoa, trích xuất đặc trưng.

1 Giới thiệu

Ngày nay, khi mức sống con người được nâng cao, đời sống tinh thần được quan tâm hơn thì nhu cầu sử dụng hoa tươi để trang trí nhà cửa, trang trí tiệc của khách hàng cũng đang ngày một tăng cao và đòi hỏi sự đa dạng. Trong bối cảnh này, ngành công nghiệp hoa tươi cũng không ngừng phát triển và đổi mới để đáp ứng nhu cầu ngày càng cao của thị trường. Tuy nhiên, việc lựa chọn hoa tươi phù hợp với trang trí là một thách thức đối với nhiều người tiêu dùng. Trang trí hoa đẹp không chỉ cần có thẩm mỹ mà còn phải hài hòa với tổng thể, góp phần tô điểm và nâng tầm cho không gian. Điều này đặt ra nhu cầu về thông tin và dữ liệu tham khảo đa dạng.

Để giải quyết vấn đề này, nhóm chúng tôi đề xuất phát triển ý tưởng về việc xây dựng một ứng dụng phân loại hoa nhằm giúp khách hàng và những người trong lĩnh vực hoa dễ dàng tìm kiếm hoa theo nhu cầu cụ thể của họ. Đối với đồ án môn học Tiền xử lý và Xây dựng bộ dữ liệu, chúng tôi sẽ tập trung trình bày những đề xuất về cáccác phương pháp phục vụ cho quá trình thu thập, tiền xử lý ảnh và các mô hình tương thích với bộ dữ liệu để phục vụ cho ứng dụng này.

2 Khái quát bài toán

Bài báo cáo gồm các nội dung: Ở phần 1 và phần 2 chúng tôi đã giới thiệu tổng quan về bài toán được đặt ra. Ở phần 3, chúng tôi cung cấp thêm về các công trình có liên quan tới bài toán mà chúng tôi đang thực hiện. Quá trình thu thập dữ liệu hình ảnh và thông tin của dữ liệu, kèm theo các bước lọc nhiễu và làm sạch dữ liệu sẽ được trình đầy đủ ở phần 4. Phần 5 sẽ trình bày các phương pháp đề xuất về viêc xử lý hình ảnh bao gồm điền khuyết điểm ảnh lỗi, phân đoạn ảnh và trích xuất đặc trưng. Phần 6 sẽ giới thiệu các mô hình máy học thông dụng và các độ đo chúng tôi dùng chọn làm tiêu chí để đánh giá. Việc thực hiện so sánh, đánh giá các kết quả và lựa chọn ra thuật toán phù hợp dựa trên kết quả độ đo sẽ được trình bày ở phần 7. Cuối cùng đưa ra kết luận và phân tích hướng phát triển trong phần . Đầu vào của bài toán: Ảnh của một loại hoa Đầu ra của bài toán: Tên loài hoa và màu sắc của loài hoa đó.

3 Công trình liên quan

Năm 2014, Hossam M. Zawbaa và các cộng sự của mình đã đề xuất phương pháp phân loại hoa hiệu quả bằng thuật toán học máy. Họ chọn ra tám loại hoa đã tiêu biểu để phân tích để rút ra những đặc điểm của chúng. Phương pháp gồm ba giai đoạn: giai đoạn phân đoạn, giai đoạn trích xuất đặc trưng và giai đoạn phân loại. Nổi bật của phương pháp này là sử dụng Chuyển đổi tính năng bất biến tỷ lệ

(SIFT) và Phân tích kết cấu Fractal dựa trên phân đoạn (SFTA) được sử dụng để trích xuất các tính năng hoa. Trong giai đoạn phân đoạn, vùng hoa được phân đoạn để loại bỏ nền phức tạp khỏi tập dữ liệu hình ảnh. Sau đó các đặc điểm hình ảnh hoa được trích xuất. Cuối cùng, đối với giai đoạn phân loại, phương pháp đề xuất đã áp dụng thuật toán Support Vector Machine và Random Forest để phân loại các loại hoa khác nhau. Thử nghiệm đã được thực hiện trên tập dữ liệu gồm 215 hình ảnh bông hoa bằng cách sử dụng phương pháp đề xuất.

Đến năm 2019, Isha Patel và cộng sự của mình là Sanskruti Patel đã ứng dụng kỹ thuật thị giác máy tính và kỹ thuật học máy để thực hiện trích xuất thông tin có ý nghĩa từ hình ảnh. Bằng phương pháp lai sử dụng Multiple Kernel Learning – Support Vector Machine, họ đã thực hiện phân loại nhiều nhãn được thử nghiệm trên tập dữ liệu chứa 25000 hình ảnh bông hoa của 102 loài khác nhau. Nền cơ bản và đặc điểm hình thái bao gồm màu sắc, kích thước, kết cấu, kiểu cánh hoa, số lượng cánh hoa, hoa đĩa, vành nhật hoa, sự hình thành hoa và hoa lớp được trích xuất để tăng độ chính xác phân loại.

4 Bộ dữ liệu

4.1 Guideline

Bộ dữ liệu chỉ gồm hình ảnh về 14 loài hoa được chọn.
Mỗi hình ảnh được gán nhãn riêng theo từng loài hoa, đảm bảo chia theo thư mục riêng.
Ảnh cần rõ nét, không bị cắt xén quá nhiều, độ phân giải không được quá thấp (dưới 6 mp).
Ảnh cần chụp chính diện loài hoa hoặc chụp hình chiếu cạnh.
Không sử dụng ảnh chụp từ dưới lên, không chấp nhận ảnh hoa chưa nở/hoa tàn

4.2 Thu thập dữ liệu

Để số lượng hình ảnh được phong phú và đa dạng, chúng tôi thực hiện việc lấy ảnh từ website Wikimedia Commons, Pixels, Pixabay, Unsplash. Quá trình này chúng tôi sử dụng một thư viện Python là Beautifulsoup để hỗ trợ. Thư viện này dùng để lấy dữ liệu ra khỏi các file HTML và XML, giúp cho việc trích xuất thông tin từ các trang web trở nên dễ dàng hơn. Ngoài ra, chúng tôi đã tham khảo và kết hợp thêm bộ dữ liệu 102flowers của Đại học Oxford.

Sau công đoạn thu thập hình ảnh, chúng tôi thu thập được 17808 mẫu dữ liệu hình ảnh về 20 loài hoa, mỗi loài khoảng từ 300 đến 2500 ảnh.

4.3 Lọc dữ liệu

Để đảm bảo chất lượng bộ dữ liệu, chúng tôi tiến hành lọc các dữ liệu hình ảnh được thu về.

1 Đưa tất cả hình ảnh về cùng định dạng .jpg.

2 Xóa các ảnh sai thông tin mà thư viện lưu về trong quá trình crawl.

3 Xóa các ảnh nhiễu, ảnh mờ, ảnh trùng lặp.

→ Kết quả sau khi lọc ảnh, chúng tôi thu được 1674 ảnh, trung bình mỗi loài hoa từ 50 đến 200 ảnh.

Nhằm thống nhất và nâng cao chất lượng dữ liệu, ảnh sau khi lọc phải đáp ứng được các tiêu chuẩn sau:
- Ảnh chỉ có 1 loại hoa
- Hoa chỉ có 1 tone màu chủ đạo
- Kích thước chiều rộng: 180 (mm)
- Kích thước chiều dài: 200 – 250 (mm)

5 Tiền xử lý dữ liệu

Tiền xử lý hình ảnh là bước cơ bản được áp dụng để nâng cao chất lượng hình ảnh và loại bỏ những nhiễu không phù hợp xuất hiện trong hình ảnh. Cải thiện hình ảnh là một quá trình quan trọng nhằm khôi phục hình thức trực quan của hình ảnh. Nó được thực hiện để giúp cho việc biến đổi hình ảnh trong các giai đoạn sau được tốt hơn. Sử dụng các phương pháp thường nâng cao đặc trưng hình ảnh như là chia tỷ lệ hình ảnh, chuyển đổi không gian màu, và tăng cường độ tương phản. Sau khi áp dụng tính năng loại bỏ nhiễu, ta thực hiện việc chuyển đổi hình ảnh từ hệ màu RGB sang thang độ xám.

5.1 Phục hồi ảnh nhiễu

Hình ảnh đầu vào đang ở hệ màu RGB [Hình 1] – một không gian màu rất phổ biến trong kỹ thuật biểu diễn hình ảnh. Trong không gian màu RGB này, mỗi điểm ảnh sự kết hợp của 3 màu sắc cơ bản: màu đỏ (R, Red), xanh lục (G, Green) và xanh lam (B, Blue) để mô tả tất cả các màu sắc khác. Tuy vậy, hình ảnh đầu vào ở hệ RGB sẽ có thể gặp nhiều khó khăn trong việc xử lý hình ảnh. Vì hình ảnh ở không gian màu này thường bị ảnh hưởng bởi các loại nhiễu (noise) và mờ (blur) như Gaussian, Salt and Pepper, Poison,...

Để giải quyết vấn đề này, chúng tôi tiến hành sử dụng bộ lọc Median để làm mờ và biến đổi điền khuyết nhiễu ảnh. Sau khi áp dụng bộ lọc, hình ảnh sẽ trở nên rõ nét hơn [Hình 2].

Hình 1: Ảnh đầu vào hệ màu RGB

Hình 2: Ảnh lọc qua bộ lọc median

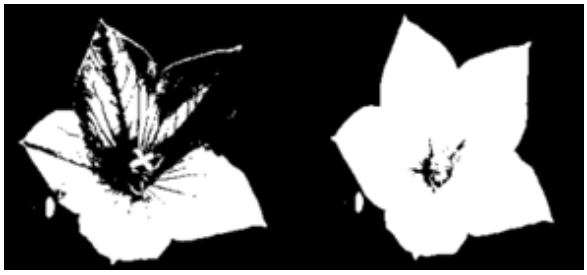
5.2 Phân đoạn ảnh

Tiếp theo, chúng tôi thực hiện việc phân đoạn trên hình ảnh. Hình ảnh có chứa hoa cũng chứa các bộ phận của cây, lá hoặc những chi tiết không liên quan trên nền ảnh. Việc loại bỏ phần nền không mong muốn này là mục tiêu của bước hai. Đây là bước cần thiết để tập trung trích xuất được chỉ đặc trưng màu và đặc trưng cạnh cần thiết.

Chúng tôi sử dụng kỹ thuật gom cụm và phân đoạn nhằm loại bỏ nền của hình ảnh và cải thiện chất lượng của hình ảnh hoa tiền cảnh. Bước đầu, hình ảnh được vào ở hệ màu RGB, được xử lý bằng thuật toán kmeans để màu trong ảnh đơn điệu hơn [Hình 3]

Hình 3: Ảnh được xử lý bằng kmeans

Sau đó, chuyển hình ảnh sang hệ màu Lab. Vì ảnh được biểu diễn bằng hệ màu RGB sẽ phụ thuộc vào từng thiết bị, hệ màu Lab đã khắc phục được khuyết điểm đó. Hệ màu Lab có sự phân tách biệt riêng về giá trị màu và độ sáng của màu sắc, giúp cho việc xử lý độ sáng và màu sắc được độc lập. Ngoài ra, không gian màu Lab giúp thể hiện rõ hơn độ tương phản của hình ảnh, điều này giúp ích rất nhiều trong việc phân tách nền ảnh. Tiếp đến, dựa vào khảo sát của [4], sử dụng phương pháp phân đoạn của Otsu, một phương pháp đơn giản và tiêu chuẩn nhất để chọn ngưỡng tự động. Mục tiêu của phương pháp này là tìm ra ngưỡng tốt nhất để phân loại các pixel thành hai lớp: lớp nền và lớp vật thể. Đây là bước biến đổi hình ảnh bông hoa thành thang độ xám nhị phân để giảm độ phức tạp của dữ liệu. Hơn nữa, để lưu trữ các ảnh được phân đoạn, các thao tác trích chọn đặc trưng.



Hình 4: Sử dụng phương pháp Otsu lên hai hệ ảnh

Bên trái là dữ liệu hình ảnh sau khi sử dụng phương pháp Otsu lên hệ màu RGB, bên phải là dữ liệu hình ảnh sau khi sử dụ phương pháp Otsu lên hệ màu Lab. Có thể thấy, ảnh phải cho kết quả khả quan hơn Cuối cùng, sử dụng hàm mask để được hình ảnh phân đoạn như mong muốn [Hình 5]

Hình 5

5.3 Trích xuất đặc trưng

Sau khi áp dụng phân đoạn hình ảnh, cần xác định các thuộc tính đặc trưng nhất định. Trích xuất đặc

trưng là quá trình áp dụng cho một hình ảnh để chuyển đổi thông tin hình ảnh của nó thành không gian vector. Tính đa chiều và khối lượng lớn dữ liệu là một thách thức đối với việc huấn luyện mô hình, vậy nên chọn tính năng đặc trưng cho bộ dữ liệu là một bước quan trọng vì nó loại bỏ các thuộc tính không liên quan nhằm mục đích tăng độ chính xác của dự đoán.

Cần phải tìm ra các lựa chọn thể hiện được nét đặc trưng riêng để suy ra thông tin của hình ảnh bông hoa và tạo ra một bộ phân loại theo cách có thể phân biệt giữa các loài khác nhau. Trong báo cáo này, chúng tôi chọn ra 2 đặc trưng nổi bật nhất là đặc trưng màu và đặc trưng cạnh.

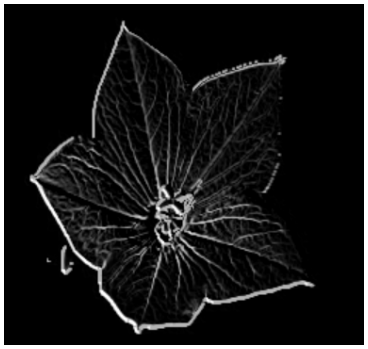
5.3.1 Đặc trưng màu

Để trích xuất đặc điểm màu từ hình ảnh bông hoa, hai mô hình màu được sử dụng là không gian màu HSV và Lab. Không gian màu HSV thường thể hiện tốt hơn về cách mọi người liên quan đến màu sắc so với mô hình màu RGB và nó cũng tạo ra đồ họa chất lượng cao. Trong không gian màu HSV, Hue, Saturation và Value lần lượt đề cập đến tông màu, sắc thái và giá trị. Để ước tính tầm nhìn của con người và tăng thêm độ chính xác cho việc diễn giải giá trị đặc trưng màu, một mô hình nổi bật khác là không gian màu Lab cũng được sử dụng.

5.3.2 Đặc trưng cạnh

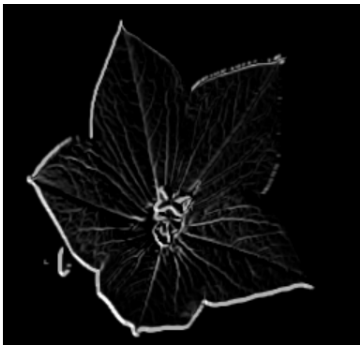
Hình ảnh hoa được chọn sẽ tại thời điểm hoa nở và có độ bung đẹp nhất. Chúng tôi sẽ tìm kết quả đặc trưng cạnh của hoa bằng ba bộ lọc là bộ lọc Sobel, bộ lọc Prewitt và bộ lọc Scharr.

Bộ lọc Sobel [Hình 6] hoạt động dựa trên việc áp dụng các bộ lọc gradient Sobel theo hai hướng (theo trục x và y) để xác định độ lớn của gradient (độ biến đổi) của ảnh theo hai hướng này. Trong đó, gradient là độ chênh lệch giữa các giá trị pixel lân cận trong hình ảnh.



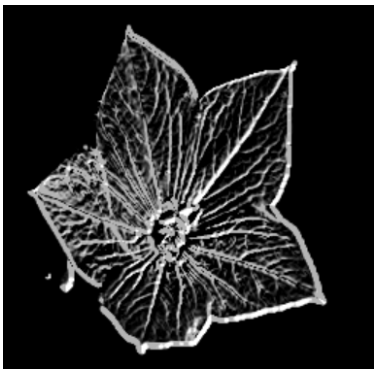
Hình 6: Ảnh qua bộ lọc Sobel

Bộ lọc Prewitt [Hình 7] hoạt động dựa trên nguyên tắc tính toán gradient của cường độ ánh sáng trong ảnh để phát hiện các cạnh. Bộ lọc Prewitt sử dụng hai ma trận kernel để tính toán gradient theo hai hướng x và y. Các ma trận này có giá trị cố định và được thiết kế để phát hiện các cạnh bằng cách tính toán sự thay đổi cường độ ánh sáng theo hai hướng.



Hình 7: Ảnh qua bộ lọc Prewitt

Bộ lọc Scharr [Hình 8] là một bộ lọc phát hiện biên cạnh tương tự như bộ lọc Prewitt và Sobel, nhưng nó có độ nhạy cao hơn trong việc phát hiện biên cạnh. Bộ lọc Scharr được thiết kế để cải thiện độ chính xác của gradient so với Sobel và Prewitt, đặc biệt là đối với các cạnh mịn và chi tiết. Giống như các bộ lọc gradient khác, bộ lọc Scharr sử dụng hai ma trận convolution (kernel) để tính toán gradient theo hướng x và y.



Hình 8: Ảnh qua bộ lọc Scharr

6 Huấn luyện dữ liệu

6.1 Các mô hình học máy

6.1.1 Random Forest

Thuật toán Random Forest (RF) là một phương pháp học máy được xây dựng dựa trên việc tạo ra một tập hợp các cây quyết định (Decision Trees),

với mỗi cây được xây dựng một cách ngẫu nhiên và độc lập từ nhau. Ý tưởng cơ bản của Random Forest là kết hợp dự đoán từ nhiều cây quyết định khác nhau để tạo ra một dự đoán tổng hợp.

Trong quá trình huấn luyện, mỗi cây quyết định trong Random Forest được xây dựng theo các bước sau: Lựa chọn mẫu ngẫu nhiên, xây cây quyết định và chọn thuộc tính tốt nhất để chia nhánh. Sau khi đã xây dựng một tập hợp các cây quyết định, quá trình dự đoán cho một điểm dữ liệu mới được thực hiện bằng cách tổng hợp dự đoán từ tất cả các cây quyết định trong Random Forest. Dự đoán cuối cùng thường được tính dựa trên đa số phiếu bầu hoặc trung bình của các dự đoán từ các cây con. Điều này giúp cân bằng giữa sự chính xác và tính ổn định của mô hình.

6.1.2 Support Vector Machine

Support Vector Machine (SVM) là một trong những thuật toán học máy có giám sát phổ biến nhất, được sử dụng cho các bài toán phân loại. Mục tiêu của thuật toán SVM là tạo đường hoặc ranh giới quyết định tốt nhất có thể tách không gian n chiều thành các lớp để chúng ta có thể dễ dàng đặt điểm dữ liệu mới vào đúng danh mục trong tương lai. Ranh giới quyết định tốt nhất này được gọi là siêu phẳng. SVM chọn các điểm/vector cực hạn giúp tạo siêu phẳng.

SVM là một thuật toán học có giám sát để sắp xếp dữ liệu thành hai loại. Nó được đào tạo với một loạt dữ liệu đã được phân loại thành hai loại, xây dựng mô hình như nó được đào tạo ban đầu. Nhiệm vụ của thuật toán SVM là xác định loại điểm dữ liệu mới thuộc về loại nào. Điều này làm cho SVM trở thành một loại bộ phân loại tuyến tính không nhị phân. Một thuật toán SVM không chỉ nên đặt các đối tượng vào các danh mục mà còn phải đặt lề giữa chúng trên một biểu đồ càng rộng càng tốt.

6.2 Độ đo đánh giá

Khi thực hiện bài toán phân loại, có 4 trường hợp của dự đoán có thể xảy ra:

- True Positive (TP): đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Positive (dự đoán đúng)
- True Negative (TN): đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Negative (dự đoán đúng)
- False Positive (FP): đối tượng ở lớp Negative, mô hình phân đối tượng vào lớp Positive (dự đoán sai)
- False Negative (FN): đối tượng ở lớp Positive, mô hình phân đối tượng vào lớp Negative (dự đoán sai)

Để đo lường hiệu quả của các thuật toán, nhóm chúng tôi sẽ sử dụng các độ đo sau:

6.2.1 Accuracy

Accuracy được định nghĩa là tỷ lệ phần trăm dự đoán đúng cho dữ liệu thử nghiệm. Nó có thể được tính toán bằng cách chia số lần dự đoán đúng cho tổng số lần dự đoán.

$$accuracy = \frac{correct\ predictions}{all\ predictions}$$

Nhược điểm của cách đánh giá này là chỉ cho ta biết được bao nhiêu phần trăm lượng dữ liệu được phân loại đúng mà không chỉ ra được cụ thể mỗi loại được phân loại như thế nào, lớp nào được phân loại đúng nhiều nhất hay dữ liệu của lớp nào thường bị phân loại nhầm nhất vào các lớp khác

6.2.2 Precision

Như đã nói phía trên, sẽ có rất nhiều trường hợp thước đo Accuracy không phản ánh đúng hiệu quả của mô hình. Vì vậy chúng ta cần một độ đo có thể khắc phục được những yếu điểm này. Precision là một trong những độ đo có thể khắc phục được, công thức như sau:

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

Precision sẽ cho chúng ta biết thực sự có bao nhiêu dự đoán Positive là thật sự True

6.2.3 Recall

Bên cạnh đó cũng là một độ đo quan trọng, nó đo lường tỷ lệ dự báo chính xác các trường hợp Positive trên toàn bộ các mẫu thuộc nhóm Positive. Công thức của Recall như sau:

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

Recall cao đồng nghĩa với việc True Positive Rate cao, tức là tỷ lệ bỏ sót các điểm thực sự là Positive là thấp.

6.2.4 F1-Score

F1-score là một độ đo kết hợp cả Recall và PPrecision

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

Một mô hình có chỉ số F1-score cao chỉ khi cả 2 chỉ số Precision và Recall đều cao. Một trong 2 chỉ số này thấp đều sẽ kéo điểm F1-score xuống. Trường hợp xấu nhất khi 1 trong hai chỉ số Precison và Recall bằng 0 sẽ kéo điểm F1-score về 0. Trường hợp tốt nhất khi cả điểm chỉ số đều đạt giá trị bằng 1, khi đó điểm F1-score sẽ là 1.

7 Kết quả thực nghiệm và đánh giá

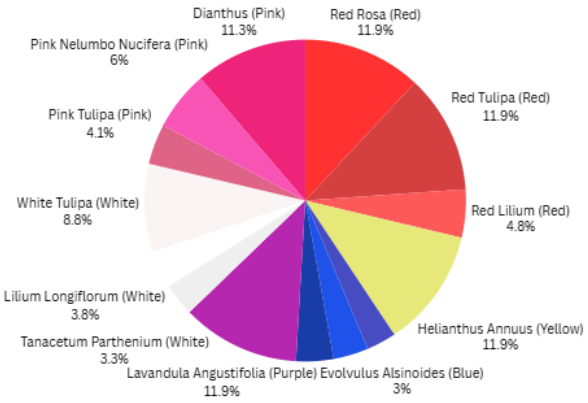
7.1 Kết quả thực nghiệm

Kết quả sau khi chúng tôi áp dụng mô hình Random Forest và mô hình SVM lên tập test, thu được kết quả dự đoán của lần lượt các nhãn như sau:

SVM			Random Forest		
P	R	F1	P	R	F1
0.33	0.08	0.12	0.50	0.15	0.24
0.43	0.83	0.57	0.63	0.89	0.74
0.00	0.00	0.00	0.67	0.22	0.33
0.55	0.38	0.45	0.84	0.90	0.87
0.41	0.74	0.53	0.52	0.87	0.64
0.00	0.00	0.00	0.50	0.17	0.25
0.33	0.06	0.10	0.67	0.24	0.35
0.00	0.00	0.00	0.40	0.12	0.19
0.00	0.00	0.00	0.40	0.18	0.25
0.00	0.00	0.00	1.00	0.05	0.10
0.19	0.33	0.33	0.57	0.57	0.57
0.39	0.39	0.39	0.53	0.75	0.62
0.00	0.00	0.00	0.00	0.00	0.00
0.14	0.17	0.16	0.55	0.70	0.62
Accuracy		0.34	Accuracy		0.60

Bảng 1: Kết quả sau khi thực thi tập dữ liệu Sobel

Tuy nhiên mô hình học máy không đạt được kết quả như chúng tôi mong đợi, đặc biệt là mô hình SVM cho kết quả rất thấp. Vậy nên chúng tôi tiến hành việc rà soát dữ liệu để tìm hiểu nguyên nhân. Qua quá trình này, chúng tôi nhận thấy sự chênh lệch đáng kể về số lượng mẫu giữa các lớp trong dữ liệu huấn luyện, dẫn đến tình trạng mất cân bằng dữ liệu. Điều này có thể là nguyên nhân khiến cho mô hình không hoạt động hiệu quả trong việc dự đoán và phân loại các lớp.



Hình 9: Tỷ lệ các nhãn

Để giải quyết vấn đề này, chúng tôi sử dụng phương pháp Oversampling. Oversampling tăng cường số lượng mẫu trong các lớp thiểu số bằng cách sao chép, tái sử dụng hoặc tạo ra các mẫu dữ liệu giả mạo từ dữ liệu huấn luyện hiện có. Mục đích của Oversampling là cân bằng lại tỷ lệ mẫu giữa các lớp để mô hình có thể học và dự đoán chính xác trên cả các lớp thiểu số.

Điều quan trọng là phải đánh giá lại hiệu quả của phương pháp Oversampling sau khi áp dụng. Trong nghiên cứu này, sau khi tăng cường mẫu dữ liệu thì mô hình đã đạt được hiệu suất tốt hơn.

SVM			Random Forest		
P	R	F1	P	R	F1
0.33	0.08	0.12	0.71	0.38	0.50
0.44	0.52	0.48	0.72	0.83	0.77
0.17	0.22	0.19	0.50	0.22	0.31
0.65	0.36	0.46	0.88	0.86	0.87
0.43	0.72	0.54	0.54	0.79	0.65
0.08	0.17	0.11	0.33	0.17	0.22
0.20	0.29	0.24	0.50	0.41	0.45
0.15	0.25	0.19	0.75	0.38	0.50
0.25	0.27	0.26	0.36	0.45	0.40
0.18	0.11	0.13	0.50	0.05	0.10
0.27	0.23	0.25	0.48	0.50	0.49
0.39	0.34	0.37	0.54	0.68	0.60
0.00	0.00	0.00	00.75	0.33	0.46
0.36	0.17	0.24	0.55	0.70	0.62
Accuracy		0.34	Accuracy		0.61

Bảng 2: Kết quả sau khi thực thi tập dữ liệu Sobel đã được Oversampling

Tiếp tục sử dụng phương pháp Oversampling lên bộ dữ liệu đã được xử lý qua bộ lọc Prewitt [Bảng 3]

Tiếp tục sử dụng phương pháp Oversampling lên

SVM			Random Forest		
P	R	F1	P	R	F1
0.40	0.40	0.40	0.60	0.30	0.40
0.36	0.44	0.39	0.58	0.74	0.65
0.25	0.27	0.26	0.60	0.27	0.37
0.67	0.48	0.56	0.84	0.92	0.88
0.38	0.58	0.46	0.56	0.81	0.66
0.00	0.00	0.00	0.13	0.20	0.16
0.13	0.22	0.17	0.33	0.11	0.17
0.30	0.35	0.33	0.80	0.40	0.53
0.20	0.23	0.21	0.54	0.54	0.54
0.11	0.08	0.10	0.00	0.00	0.00
0.17	0.17	0.17	0.45	0.42	0.44
0.38	0.32	0.34	0.54	0.55	0.55
0.00	0.00	0.00	0.60	0.21	0.32
0.21	0.10	0.14	0.44	0.53	0.48
Accuracy		0.32	Accuracy		0.56

Bảng 3: Kết quả sau khi thực thi tập dữ liệu Prewitt đã được Oversampling

SVM			Random Forest		
P	R	F1	P	R	F1
0.15	0.30	0.20	0.40	0.40	0.40
0.40	0.53	0.46	0.59	0.76	0.67
0.33	0.27	0.30	0.80	0.36	0.50
0.69	0.36	0.47	0.76	0.96	0.85
0.50	0.42	0.46	0.64	0.70	0.67
0.04	0.10	0.06	0.11	0.20	0.14
0.14	0.22	0.17	0.60	0.33	0.43
0.28	0.25	0.26	0.54	0.35	0.42
0.20	0.23	0.21	0.60	0.46	0.52
0.00	0.00	0.00	0.00	0.00	0.00
0.19	0.20	0.20	0.59	0.55	0.57
0.26	0.29	0.27	0.63	0.68	0.66
0.14	0.07	0.10	0.40	0.14	0.21
0.25	0.20	0.22	0.53	0.57	0.55
Accuracy		0.34	Accuracy		0.59

Bảng 4: Kết quả sau khi thực thi tập dữ liệu Scharr đã được Oversampling

bộ dữ liệu đã được xử lý qua bộ lọc Scharr [Bảng 4]

7.2 Đánh giá

7.2.1 Mô hình

Mô hình Random Forest và Support Vector Machine đều là những công cụ mạnh mẽ cho bài toán phân loại. Tuy nhiên, trong bài toán này, mô hình Random Forest đều cho kết quả tốt hơn SVM qua 3 bộ dữ liệu. Điều này có thể đến từ các vấn đề của bộ dữ liệu. Random Forest có cơ chế xử lý nhiều biến tốt nhờ vào việc trung bình hóa dự đoán từ nhiều cây quyết định khác nhau, giúp giảm ảnh hưởng của nhiễu và sai lệch trong dữ liệu. SVM có thể gặp khó khăn khi mô hình hóa bộ dữ liệu cócó số lượng tính năng rất lớn và có nhiều đặc trưng không liên quan. Ngoài ra RF có khả năng xử lý dữ liệu mất mát tốt hơn bằng cách sử dụng chiến lược như dự đoán giá trị bị thiếu dựa trên những cây khác nhau.

7.2.2 Bộ lọc

Bộ lọc Sobel cho kết quả hình ảnh trích xuất các cạnh được nhấn mạnh rõ ràng, độ dài các cạnh có xu hướng dày hơn, các đường biên rõ ràng và mạnh mẽ hơn.

Bộ lọc Prewitt vẫn trích xuất được các cạnh nhưng mỏng nên không rõ ràng và mạnh mẽ. Ảnh sau khi được trích xuất trông và ít nhiễu hơn.

Bộ lọc Scharr cho hình ảnh trích xuất nổi bật và rõ ràng các đường biên cạnh và gân hoa. Tuy nhiên vì bộ lọc quá nhạy cảm với nhiễu nên hình ảnh kết quả quá nhiều chi tiết.

Trong bài toán này, khả năng làm nổi bật cạnh mạnh mẽ của Sobel cân bằng với độ nhạy cảm nhiễu là yếu tố quan trọng giúp cho bộ dữ liệu được xử lý bằng bộ lọc Sobel cho kết quả tốt nhất. Hình ảnh được trích xuất cạnh rõ ràng với các chi tiết đặc trưng giúp cho mô hình học máy dễ dàng dự đoán kết quả tốt. Bộ lọc Prewitt tuy xử lý tốt ảnh nhiễu nhưng trọng số của bộ lọc ít mạnh mẽ hơn nên nó giảm khả năng phát hiện các cạnh và chi tiết nhỏ. Ngược lại, bộ lọc Scharr có thể trích xuất được quá nhiều chi tiết nhỏ, bao gồm cả nhiễu dẫn đến hình ảnh có thể khó hiểu hơn làm tăng độ phức tạp cho bài toán học máy. Điều này cũng làm giảm hiệu quả của thuật toán học máy ở bước dự đoán phân loại tiếp theo.

8 Kết luận và công việc tương lai

Trong bài báo cáo này, chúng tôi đã đề xuất các phương pháp tiền xử lý để nâng cao chất lượng một bộ dữ liệu hình ảnh 1674 ảnh với 14 nhãn đặc trưng. Các công đoạn được thực hiện thành công bằng việc áp dụng các kỹ thuật của thị giác máy tính. Đồng thời, kết quả của bài toán cũng mang đến cho người đọc một cái nhìn tổng quan hơn về các phương pháp tiền xử lý để có thể áp dụng cho những bộ dữ liệu cá nhân trong từng trường hợp.

Trong tương lai tới, chúng tôi có thể tiếp tục tìm thêm các phương pháp tiền xử lý dữ liệu và cải

thiện nâng cao các mô hình học máy. Nếu kết quả từ các thử nghiệm cho thấy các phương pháp có tiềm năng cao trong việc cải thiện độ chính xác của dự đoán, chúng tôi sẽ mở rộng bộ dữ liệu: tăng số lượng và loại hoa, đa dạng hóa điều kiện chụp,...

9 References

1 Sanskruti Patel and Isha Patel, Flower Identification and Classification using Computer Vision and Machine Learning Techniques. 2019 International Journal of Engineering and Advanced Technology.

2 Hossam M. Zawbaa, Mona Abbass, Sameh H. Basha and Maryam Hazman, Automatic Flower Classification Approach Using Machine Learning Algorithms, 2014 International Conference on Advances in Computing.

3 Bùi Đức Hành, Các hệ màu cơ bản trong xử lý ảnh, 2019.

4 Sunil L. Bangare, Amruta Dubal, Pallavi S. Bangare, Dr. S. T. Patil, Reviewing Otsu's Method For Image hresholding, 2015.