TRƯỜNG ĐẠI HỌC LÂM NGHIỆP

**KHOA CƠ ĐIỆN VÀ CÔNG TRÌNH**

**----------o0o----------**

**ĐỒ Án**

**MÔN: Khai Phá Dữ Liệu**

**ĐỀ TÀI: “ Nhận Diện Loài Hoa Bằng Hình Ảnh”**

|  |  |
| --- | --- |
| *Giảng viên hướng dẫn*  *Thành viên nhóm*  *Mã Sinh Viên* | : Mai Hà An  : Nguyễn Đức Hạnh  Ngô Khoa Nguyên  : 2175801080530  2174801040026 |

Năm 2025

**H**

**À NỘI - 2024**



**MỤC LỤC**

[Chương 1. Mở đầu 1](#_Toc195205275)

[1.1 Đặt vấn đề 1](#_Toc195205276)

[1.2. Mục tiêu của đề tài 1](#_Toc195205277)

[1.3. Phạm vi và giới hạn 2](#_Toc195205278)

[a .Phạm vi nghiên cứu 2](#_Toc195205279)

[b. Giới hạn nghiên cứu 2](#_Toc195205280)

[c. Ứng dụng thực tế 2](#_Toc195205281)

[CHƯƠNG 2: TỔNG QUAN VỀ NHẬN DIỆN HÌNH ẢNH 3](#_Toc195205282)

[2.1 Khái niệm cơ bản 3](#_Toc195205283)

[2.2 Các kỹ thuật trong AI 5](#_Toc195205284)

[2.2.1 Học sâu (Deep Learning) 5](#_Toc195205285)

[2.3.2 Phân loại (Classification) 5](#_Toc195205286)

[2.3.3 Phân cụm (Clustering) 5](#_Toc195205287)

[CHƯƠNG 3: DỮ LIỆU ĐẦU VÀO CHO DỮ LIỆU NHẬN DIỆN HÌNH ẢNH 5](#_Toc195205288)

[3.1 Nguồn Dữ Liệu Hình Ảnh 5](#_Toc195205289)

[3.2 Cách Thức Thu Thập Và Chuẩn Bị Dữ Liệu 6](#_Toc195205290)

[**3.3 Tiền xử lý dữ liệu** 7](#_Toc195205291)

**LỜI MỞ ĐẦU**

Trong thời buổi hiện đại ngày nay, công nghệ thông tin cũng như những ứng dụng của nó không ngừng phát triển, lượng thông tin và cơ sở dữ liệu được thu thập và lưu trữ cũng tích lũy ngày một nhiều lên. Con người cũng vì thế mà cần có thông tin với tốc độ nhanh nhất để đưa ra quyết định dựa trên lượng dữ liệu khổng lồ đã có. Các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống ngày càng không đáp ứng được thực tế, vì thế, một khuynh hướng kỹ thuật mới là Kỹ thuật phát hiện tri thức và khai phá dữ liệu nhanh chóng được phát triển. Khai phá dữ liệu đã và đang được nghiên cứu, ứng dụng trong nhiều lĩnh vực khác nhau ở các nước trên thế giới. Ở Việt Nam, kỹ thuật này đang được nghiên cứu và dần đưa vào ứng dụng. Khai phá dữ liệu là một bước trong quy trình phất hiện tri thức. Hiện nay, mọi người không ngừng tìm tòi các kỹ thuật để thực hiện khai phá dữ liệu một cách nhanh nhất và có được kết quả tốt nhất. Trong bài tập lớn này, chúng em tìm hiểu và trình bày về một kỹ thuật trong khai phá dữ liệu để phân lớp dữ liệu cũng như tổng quan về khai phá dữ liệu, với đề tài “**nhận diện loài hoa bằng hình ảnh**”. Trong quá trình làm bài tập lớn này, em xin gửi lời cảm ơn đến thầy giáo. Em rất mong nhận được những góp ý từ thầy. Em xin chân thành cảm ơn!

# Chương 1. Mở đầu

## 1.1 Đặt vấn đề

**Tổng quan về nhận diện hình ảnh**:

Nhận diện hình ảnh là một lĩnh vực của trí tuệ nhân tạo (AI) và học máy Nhận diện hình ảnh là một lĩnh vực của trí tuệ nhân tạo (AI) và học máy (machine learning), nơi mà máy tính có khả năng phân tích và hiểu nội dung của hình ảnh. Công nghệ này đã trở thành một phần thiết yếu trong cuộc sống hiện đại, ảnh hưởng đến nhiều lĩnh vực từ an ninh đến y tế. Ví dụ, nhận diện khuôn mặt được sử dụng trong việc mở khóa điện thoại, giúp người dùng bảo mật thông tin cá nhân một cách nhanh chóng và tiện lợi. Bên cạnh đó, nhận diện đối tượng trong ảnh đang ngày càng phổ biến trong ngành bán lẻ, giúp phân loại và quản lý hàng hóa một cách tự động.

## 1.2. Mục tiêu của đề tài

* **Tìm hiểu cách kết hợp khai phá dữ liệu với nhận diện hình ảnh**: Nghiên cứu cách thức áp dụng các kỹ thuật khai phá dữ liệu để tối ưu hóa các bước xử lý dữ liệu hình ảnh, bao gồm cả xử lý trước (preprocessing) và trích xuất đặc trưng (feature extraction).
* **Nâng cao độ chính xác và tốc độ của hệ thống nhận diện**: Đề xuất các phương pháp khai phá dữ liệu nhằm giảm thiểu các lỗi nhận diện do nhiễu, thiếu ánh sáng hoặc các biến dạng trong hình ảnh. Bên cạnh đó, tăng tốc độ xử lý dữ liệu để hệ thống nhận diện hình ảnh có thể hoạt động trong thời gian thực.
* **Thử nghiệm và so sánh các phương pháp**: Đánh giá hiệu quả của các thuật toán khai phá dữ liệu khác nhau (như phân loại, phân cụm, hay giảm chiều) trong quá trình nhận diện hình ảnh. So sánh kết quả giữa việc sử dụng khai phá dữ liệu và không sử dụng trong các mô hình nhận diện.

## 1.3. Phạm vi và giới hạn

### a .Phạm vi nghiên cứu

* **Công cụ sử dụng**:
  + **Google Colab**: Chạy mã Python và huấn luyện mô hình với GPU miễn phí.
  + **Google Drive**: Lưu trữ tập dữ liệu và mô hình sau khi huấn luyện.
  + **Thư viện Ultralytics YOLOv8** (ultralytics).
* **Các bước thực hiện**:
  + **Kết nối Google Drive** với Google Colab để truy cập dữ liệu.
  + **Cài đặt YOLOv8** và các thư viện cần thiết.
  + **Chuẩn bị dữ liệu**: Tải dataset lên Google Drive, kiểm tra định dạng nhãn (YOLO, COCO, PASCAL VOC).
  + **Huấn luyện mô hình** trên Google Colab với tập dữ liệu tùy chỉnh.
  + **Đánh giá mô hình** và thử nghiệm nhận diện trên ảnh/video thực tế.
  + **Xuất và lưu mô hình** lên Google Drive để sử dụng sau này.

### b. Giới hạn nghiên cứu

* Google Colab có giới hạn về **thời gian chạy** và **tài nguyên GPU**, có thể ảnh hưởng đến quá trình huấn luyện mô hình lớn.
* Việc tải dữ liệu từ Google Drive có thể **mất thời gian** nếu kích thước dữ liệu lớn.
* Độ chính xác của mô hình phụ thuộc vào **chất lượng và số lượng dữ liệu huấn luyện**.

### c. Ứng dụng thực tế

* Ứng dụng vào **giám sát an ninh**, **nhận diện vật thể trong ảnh/video**, **tự động hóa trong công nghiệp**, v.v.
* Có thể triển khai mô hình đã huấn luyện trên **Raspberry Pi, Jetson Nano, hoặc Web API**.

# CHƯƠNG 2: TỔNG QUAN VỀ NHẬN DIỆN HÌNH ẢNH

## 2.1 Khái niệm cơ bản

**Nhận diện hình ảnh** là một lĩnh vực quan trọng trong công nghệ thông tin và trí tuệ nhân tạo, với mục tiêu chính là xác định, nhận diện và phân loại các đối tượng hoặc đặc trưng cụ thể từ hình ảnh kỹ thuật số hoặc video. Nhận diện hình ảnh bao gồm nhiều giai đoạn, từ tiền xử lý dữ liệu, trích xuất đặc trưng, đến phân loại đối tượng. Đây là một bước tiến lớn trong việc tự động hóa nhiều quy trình công nghiệp và thương mại, đặc biệt là trong các ứng dụng như nhận diện khuôn mặt, nhận diện chữ viết tay, và phân loại các loại hình ảnh trong y tế.

Khai phá dữ liệu là một quá trình xác định các mẫu tiềm ẩn có tính hợp lệ, mới lạ, có ích và có thể hiểu được trong một khối dữ liệu rất lớn.

**Khai Phá Dữ Liệu**

Khai phá tri thức từ CSDL ( Knowledge Discovery in Database) Khai phá tri thức từ CSDL gồm 5 bước:

* B1: Lựa chọn CSDL
* B2: Tiền xử lý
* B3: Chuyển đổi
* B4: Khai phá dữ liệu
* B5: Diễn giải và đánh giá

Khai phá dữ liệu là 1 bước trong quá trình khai phá tri thức từ CSDL. Các ứng dụng của khai phá dữ liệu Phát hiện tri thức và khai phá dữ liệu liên quan đến nhiều ngành, nhiều lĩnh vực: thống kê, trí tuệ nhân tạo, cơ sở dữ liệu, thuật toán, tính toán song song và tốc độ cao, thu thập tri thức cho các hệ chuyên gia, quan sát dữ liệu... Đặc biệt phát hiện tri thức và khai phá dữ liệu rất gần gũi với lĩnh vực thống kê, sử dụng các phương pháp thống kê để mô hình dữ liệu và phát hiện các mẫu, luật ... Ngân hàng dữ liệu (Data Warehousing) và các công cụ phân tích trực tuyến (OLAP- On Line Analytical Processing) cũng liên quan rất chặt chẽ với phát hiện tri thức và khai phá dữ liệu. Khai phá dữ liệu có nhiều ứng dụng trong thực tế,

Các bước của quá trình khai phá dữ liệu Quy trình phát hiện tri thức thường tuân theo các bước sau:

* Bước thứ nhất: Hình thành, xác định và định nghĩa bài toán. Là tìm hiểu lĩnh vực ứng dụng từ đó hình thành bài toán, xác định các nhiệm vụ cần phải hoàn thành. Bước này sẽ quyết định cho việc rút ra được các tri thức hữu ích và cho phép chọn các phương pháp khai phá dữ liệu thích hợp với mục đích ứng dụng và bản chất của dữ liệu.
* Bước thứ hai: Thu thập và tiền xử lý dữ liệu. Là thu thập và xử lý thô, còn được gọi là tiền xử lý dữ liệu nhằm loại bỏ nhiễu (làm sạch dữ liệu), xử lý việc thiếu dữ liệu (làm giàu dữ liệu), biến đổi dữ liệu và rút gọn dữ liệu nếu cần thiết, bước này thường chiếm nhiều thời gian nhất trong toàn bộ qui trình phát hiện tri thức. Do dữ liệu được lấy từ nhiều nguồn khác nhau, không đồng nhất, … có thể gây ra các nhầm lẫn. Sau bước này, dữ liệu sẽ nhất quán, đầy đủ, được rút gọn và rời rạc hoá.
* Bước thứ ba: Khai phá dữ liệu, rút ra các tri thức. Là khai phá dữ liệu, hay nói cách khác là trích ra các mẫu hoặc/và các mô hình ẩn dưới các dữ liệu. Giai đoạn này rất quan trọng, bao gồm các công đoạn như: chức năng, nhiệm vụ và mục đích của khai phá dữ liệu, dùng phương pháp khai phá nào? Thông thường, các bài toán khai phá dữ liệu bao gồm: các bài toán mang tính mô tả - đưa ra tính chất chung nhất của dữ liệu, các bài toán dự báo - bao gồm cả việc phát hiện các suy diễn dựa trên dữ liệu hiện có. Tuỳ theo bài toán xác định được mà ta lựa chọn các phương pháp khai phá dữ liệu cho phù hợp.
* Bước thứ tư: Sử dụng các tri thức phát hiện được. Là hiểu tri thức đã tìm được, đặc biệt là làm sáng tỏ các mô tả và dự đoán. Các bước trên có thể lặp đi lặp lại một số lần, kết quả thu được có thể được lấy trung bình trên tất cả các lần thực hiện. Các kết quả của quá trình phát hiện tri thức có thể được đưa vào ứng dụng trong các lĩnh vực khác nhau do các kết quả có thể là các dự đoán.

## 2.2 Các kỹ thuật trong AI

### 2.2.1 Học sâu (Deep Learning)

* **Ưu điểm:**
  + Khả năng trích xuất đặc trưng tự động mà không cần thiết kế thủ công.
  + Hiệu quả cao với dữ liệu lớn và phức tạp.
* **Nhược điểm:**
  + Yêu cầu tài nguyên tính toán cao.
  + Đòi hỏi lượng lớn dữ liệu để huấn luyện.

### 2.3.2 Phân loại (Classification)

* **Ưu điểm:**
  + Áp dụng được trên tập dữ liệu có nhãn.
  + Các thuật toán như Decision Tree, Naïve Bayes dễ triển khai và giải thích.
* **Nhược điểm:**
  + Hiệu quả thấp nếu dữ liệu có nhiều nhiễu hoặc thiếu cân bằng.

### 2.3.3 Phân cụm (Clustering)

* **Ưu điểm:**
  + Tốt cho dữ liệu không có nhãn.
  + Phát hiện các cụm mới hoặc các mẫu không mong đợi.
* **Nhược điểm:**
  + Độ chính xác phụ thuộc vào thuật toán và đặc trưng của dữ liệu.

# CHƯƠNG 3: DỮ LIỆU ĐẦU VÀO CHO DỮ LIỆU NHẬN DIỆN HÌNH ẢNH

## 3.1 Nguồn Dữ Liệu Hình Ảnh

* **CIFAR-10**: Là một tập hợp dữ liệu hình ảnh gồm 60.000 hình ảnh màu kích thước 32x32 trong 10 lớp khác nhau. Đây là tập dữ liệu thách thức dành cho các mô hình nhận diện hình ảnh.
* **MNIST**: Gồm 70.000 hình ảnh đen trắng 28x28 pixel của các chữ số viết tay từ 0 đến 9, đây là một trong những tập dữ liệu chuẩn để đánh giá các mô hình nhận diện chữ số.
* **ImageNet**: Một tập hợp dữ liệu lớn với hơn 14 triệu hình ảnh thuộc hàng ngàn danh mục khác nhau, được sử dụng trong các cuộc thi như ImageNet Large Scale Visual Recognition Challenge (ILSVRC).

## 3.2 Cách Thức Thu Thập Và Chuẩn Bị Dữ Liệu

Để đảm bảo chất lượng dữ liệu cho hệ thống nhận diện hình ảnh, cần thực hiện các bước thu thập và chuẩn bị như sau:

### 3.2.1 Thu thập dữ liệu

* **Từ các bộ dữ liệu công khai:**  
  Tải về từ các nguồn trực tuyến như Kaggle, TensorFlow Datasets, hoặc các kho lưu trữ dữ liệu học thuật.
* **Tự chụp hình ảnh:**  
  Sử dụng máy ảnh hoặc điện thoại thông minh để chụp ảnh cây cối trong môi trường thực tế.
* **Lấy dữ liệu từ Internet:**  
  Thu thập hình ảnh từ các trang web hoặc cơ sở dữ liệu trực tuyến liên quan đến cây cối.

### 3.2.2 Chuẩn bị dữ liệu

1. **Làm sạch dữ liệu:**
   * Loại bỏ các hình ảnh bị mờ, nhiễu hoặc không liên quan.
   * Xử lý các hình ảnh có kích thước không đồng đều.
2. **Chuyển đổi định dạng:**
   * Đưa tất cả hình ảnh về định dạng chuẩn như JPEG hoặc PNG.
   * Chuyển đổi kích thước ảnh về cùng một kích thước (ví dụ: 32x32 pixel).
3. **Dán nhãn dữ liệu:**
   * Đối với dữ liệu tự thu thập, cần gán nhãn thủ công cho từng hình ảnh dựa trên các đặc trưng của cây (loại cây, màu sắc, hình dạng lá).

## 3.3 Tiền xử lý dữ liệu

Tiền xử lý là bước quan trọng giúp cải thiện chất lượng đầu vào và tăng hiệu suất của mô hình nhận diện hình ảnh. Các bước chính bao gồm:

### 3.3.1 Chuyển đổi kích thước

* Hình ảnh từ các nguồn khác nhau có thể có kích thước không đồng đều. Do đó, cần chuyển đổi tất cả hình ảnh về một kích thước cố định.
* **Ví dụ:** Đưa tất cả ảnh về kích thước 32x32 pixel hoặc 640x640 pixel (dành cho mô hình học sâu).

### 3.3.2 Bình thường hóa giá trị pixel

* Chuyển đổi giá trị pixel từ khoảng [0, 255] về khoảng [0, 1] hoặc [-1, 1]. Điều này giúp mô hình hội tụ nhanh hơn trong quá trình huấn luyện.
* **Công thức:**  
  Pixel bình phương hóa=Pixel gốc/255​

### 3.3.3 Loại bỏ nhiễu

* Sử dụng các bộ lọc để giảm nhiễu trong hình ảnh, ví dụ:
  + **Bộ lọc Gaussian:** Làm mờ hình ảnh để giảm chi tiết không mong muốn.
  + **Bộ lọc Median:** Loại bỏ nhiễu mà vẫn giữ được cạnh của đối tượng.

### 3.3.4 Tăng cường dữ liệu (Data Augmentation)

* Tăng kích thước tập dữ liệu bằng cách tạo ra các phiên bản khác nhau của cùng một hình ảnh.
* Các kỹ thuật thường dùng:
  + Lật ảnh ngang/dọc.
  + Xoay ảnh (góc 15°, 30°, 45°,...).
  + Phóng to hoặc thu nhỏ ảnh (Zoom in/out).
  + Dịch chuyển ảnh theo chiều ngang/dọc.

### 3.3.5 Chia tách tập dữ liệu

* Chia dữ liệu thành 3 phần chính:
  1. **Tập huấn luyện (Training Set):** 70% dữ liệu, dùng để dạy mô hình.
  2. **Tập kiểm tra (Validation Set):** 20% dữ liệu, dùng để đánh giá mô hình trong quá trình huấn luyện.
  3. **Tập kiểm thử (Test Set):** 10% dữ liệu, dùng để đánh giá hiệu suất cuối cùng của mô hình.

## 3.4 Các thách thức trong xử lý dữ liệu đầu vào

* **Nhiễu và độ mờ của hình ảnh:** Hình ảnh thu thập thực tế có thể chứa nhiễu từ môi trường (ánh sáng yếu, hình ảnh bị rung).
* **Dữ liệu không cân bằng:** Một số lớp dữ liệu (ví dụ: cây thông) có thể nhiều hơn các lớp khác (ví dụ: cây xoài), dẫn đến sai lệch trong mô hình.
* **Định dạng và kích thước khác nhau:** Hình ảnh từ nhiều nguồn cần được chuẩn hóa trước khi sử dụng.

# Chương 4: Mô hình

