BỘ NÔNG NGHIỆP VÀ MÔI TRƯỜNG

PHÂN HIỆU TRƯỜNG ĐẠI HỌC THỦY LỢI

BỘ MÔN CÔNG NGHỆ THÔNG TIN

BÁO CÁO CUỐI KỲ

MÔN HỌC: Học Máy

ĐỀ TÀI: Ứng dụng Machine Learning trong dự đoán phân loại bệnh lá lúa

Giảng viên:	Ths. Vũ Thị Hạnh
Sinh viên thực hiện:	Nguyễn Lê Minh Hậu – 2351267261 Nguyễn Đức Huy - 2351267265
Lớp:	S26-65TTNT

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

TP. Hồ Chí Minh, ngày … tháng … năm 2026

Chữ ký của giảng viên

LỜI CẢM ƠN

Để hoàn thành báo cáo kết thúc môn học này, bên cạnh sự nỗ lực của từng các nhân thì nhóm em đã nhận được sự hướng dẫn tận tình và đầy tâm huyết từ giảng viên hướng dẫn.

Lời đầu tiên, nhóm em xin được bày tỏ lòng biết ơn sâu sắc và chân thành nhất đến cô Vũ Thị Hạnh, người đã trực tiếp giảng dạy, định hướng và đồng hành cùng nhóm em trong suốt quá trình học tập và thực hiện đồ án môn Học Máy này.

Môn Học Máy này không chỉ giúp chúng em tiếp cận những kiến thức nền tảng quan trọng về các thuật toán học máy, mô hình dự đoán và phân tích dữ liệu, mà còn rèn luyện cho chúng em tư duy logic, khả năng lập trình cũng như phương pháp tiếp cận và giải quyết các bài toán thực tế dựa trên dữ liệu. Chúng em đặc biệt trân trọng những

bài giảng rõ ràng, khoa học của cô, sự tận tâm trong việc giải đáp thắc mắc cũng như những định hướng chuyên môn quý báu, giúp chúng em từng bước hoàn thiện mô hình và hiểu sâu hơn bản chất của các thuật toán đã học.

Sự nghiêm túc trong học thuật, cùng với sự động viên và khích lệ đúng lúc của cô, chính là động lực lớn để chúng em không ngừng cố gắng, hoàn thiện bài báo cáo một cách chỉn chu và nghiêm túc nhất.

Bên cạnh đó, chúng em cũng xin gửi lời cảm ơn đến các nguồn tài liệu tham khảo và cộng đồng học thuật, đã cung cấp những kiến thức, dữ liệu và kinh nghiệm thực tiễn, giúp chúng em có cơ sở để vận dụng hiệu quả những nội dung đã được học vào bài toán cụ thể.

Do thời gian thực hiện và kiến thức còn hạn chế, bài báo cáo của nhóm em khó tránh khỏi những thiếu sót. Nhóm em rất mong nhận được những nhận xét và góp ý chân thành từ cô để chúng em có thể rút kinh nghiệm và hoàn thiện bản thân hơn trong quá trình học tập và nghiên cứu sau này.

Chúng em xin chân thành cảm ơn cô!

DANH MỤC

I. GIỚI THIỆU 11
1.1 Bối cảnh và lý do chọn đề tài 11
.1.2 Vai trò của Machine Learning trong nông nghiệp 11
1.3 Tổng quan các nghiên cứu liên quan 12
1.4 Cấu trúc dự án 15
II. MỤC TIÊU VÀ BÀI TOÁN ĐẶT RA 17
2.1 Mục tiêu chính 17
2.2 Bài toán đặt ra 18
2.3 Các mô hình được lựa chọn 19
III. MÔ TẢ DỮ LIỆU VÀ BƯỚC TIỀN XỬ LÝ 20
3.1 Nguồn và đặc điểm dữ liệu 20
3.2 Các đặc trưng gốc (Original Features) 21
3.3 Phân tích dữ liệu khám phá (EDA) 21
3.4 Pipeline tiền xử lý dữ liệu 22
IV. MÔ HÌNH HỌC MÁY SỬ DỤNG 25
4.1. Tổng quan quy trình xây dựng mô hình 25
4.2. Mô hình CNN cơ bản 27
4.3. Mô hình học EfficientNetB1 30
4.4. Trực quan hóa Grad-CAM 31
V. KẾT QUẢ VÀ ĐÁNH GIÁ MÔ HÌNH 32
5.1. Thiết lập thực nghiệm 32
5.2. Kết quả huấn luyện và đánh giá 33
5.3. Phân tích kết quả 33
5.4. Kết luận chương 34
VI. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 34
6.1. Kết luận chung 34
6.2. Những đóng góp chính 35
6.3. Những hạn chế 35

6.4. Hướng phát triển trong tương lai.....36

I. GIỚI THIỆU

1.1 Bối cảnh và lý do chọn đề tài

Trong những năm gần đây, sự phát triển mạnh mẽ của trí tuệ nhân tạo, đặc biệt là học sâu trong thị giác máy tính, đã tạo điều kiện thuận lợi để xây dựng các hệ thống có khả năng phân tích và nhận dạng hình ảnh một cách tự động. Nhờ đó, nhiều bài toán thực tế có thể được giải quyết nhanh hơn và chính xác hơn so với phương pháp thủ công.

Trong nông nghiệp, bệnh hại trên cây lúa là một trong những nguyên nhân phổ biến làm giảm năng suất và chất lượng. Nhiều loại bệnh biểu hiện trên lá có dấu hiệu khá giống nhau, trong khi ảnh chụp ngoài đồng thường chịu ảnh hưởng bởi ánh sáng, nền phức tạp và góc chụp, khiến việc nhận diện bằng mắt thường dễ nhầm lẫn. Vì vậy, việc ứng dụng học sâu để xây dựng mô hình phân loại bệnh lá lúa từ ảnh là hướng tiếp cận phù hợp, có tính ứng dụng cao.

Xuất phát từ thực tiễn đó, đề tài "Hệ thống phân loại bệnh lá lúa " được lựa chọn nhằm xây dựng một quy trình hoàn chỉnh từ phân tích dữ liệu, tiền xử lý ảnh, huấn luyện và đánh giá mô hình, đến triển khai giao diện web giúp người dùng tải ảnh và nhận kết quả dự đoán nhanh chóng

.1.2 Vai trò của Machine Learning trong nông nghiệp

Machine Learning đã và đang được ứng dụng rộng rãi trong lĩnh vực nông nghiệp nhằm hỗ trợ người sản xuất trong việc phân tích dữ liệu, giám sát cây trồng và đưa ra quyết định canh tác. Thông qua việc học từ dữ liệu lịch sử và dữ liệu thực tế (hình ảnh, thời tiết, giống cây, điều kiện môi trường…), các mô hình Machine Learning có khả năng phát hiện những mối quan hệ tiềm ẩn mà các phương pháp quan sát và phân tích truyền thống khó nhận ra.

Trong bài toán phân loại bệnh lá lúa, Machine Learning, đặc biệt là học sâu trong thị giác máy tính, cho phép xử lý trực tiếp ảnh chụp ngoài đồng và tự động trích xuất các đặc trưng quan trọng như màu sắc, kết cấu bề mặt lá, hình dạng vết bệnh và

mức độ lan rộng. Nhờ đó, mô hình có thể nhận diện và phân loại nhiều nhóm bệnh khác nhau với độ chính xác cao hơn, đồng thời giảm sự phụ thuộc vào kinh nghiệm chủ quan của người quan sát. Việc kết hợp tiền xử lý ảnh và tăng cường dữ liệu cũng giúp mô hình hoạt động ổn định hơn trong điều kiện ánh sáng thay đổi, nền ảnh phức tạp hoặc ảnh có chất lượng không đồng đều.

Việc ứng dụng Machine Learning trong nông nghiệp không nhằm thay thế hoàn toàn vai trò của con người, mà đóng vai trò như một công cụ hỗ trợ ra quyết định, giúp phát hiện sớm nguy cơ bệnh, nâng cao hiệu quả quản lý mùa vụ và góp phần hướng tới nông nghiệp thông minh.

1.3 Tổng quan các nghiên cứu liên quan

Phân loại bệnh cây trồng nói chung và bệnh trên lá lúa nói riêng bằng Machine Learning và Deep Learning là hướng nghiên cứu được quan tâm mạnh trong khoảng hơn một thập kỷ trở lại đây. Cùng với sự phát triển của các thuật toán học sâu và sự sẵn có của nhiều bộ dữ liệu ảnh nông nghiệp công khai, nhiều phương pháp khác nhau đã được đề xuất nhằm nâng cao độ chính xác nhận diện bệnh trong điều kiện môi trường thực tế. Các nghiên cứu thường tập trung vào việc khai thác đặc trưng hình ảnh như màu sắc, kết cấu và hình dạng vết bệnh để xây dựng mô hình phân loại, hỗ trợ giám sát dịch bệnh và ra quyết định trong canh tác.

Bảng dưới đây tổng hợp một số hướng tiếp cận tiêu biểu liên quan đến bài toán phân loại bệnh lá, bao gồm lá lúa, bằng Machine Learning và Deep Learning.

Tác giả (Năm)	**Phương pháp**	**Dataset**	**Kết quả tốt nhất**	**Hạn chế**
Mohanty et al. (2016)	Deep Learning (AlexNet/GoogLeNet – Transfer Learning)	PlantVillage	Accuracy cao (≈ 99% trên bộ dữ liệu chuẩn)	Dữ liệu phòng thí nghiệm, nền đơn giản; khó

Tác giả (Năm)	Phương pháp	Dataset	Kết quả tốt nhất	Hạn chế
				khái quát ra ảnh ngoài đồng
Sladojevic et al. (2016)	CNN phân loại bệnh lá	Tập ảnh bệnh lá	Accuracy cao trên tập nghiên cứu	Chưa đánh giá sâu với điều kiện ánh sáng/nền phức tạp ngoài đồng
Ferentinos (2018)	CNN với nhiều kiến trúc khác nhau	PlantVillage	Accuracy cao (≈ 99%)	Chưa phản ánh tốt điều kiện thực tế; ít nhiễu môi trường
Too et al. (2019)	So sánh CNN (VGG, ResNet, DenseNet, Inception)	PlantVillage	Mô hình sâu cho hiệu năng tốt	Tốn tài nguyên; hiệu năng giảm khi ảnh thực tế bị nhiễu
Một số nghiên cứu gần đây	EfficientNet/Transfer Learning + Augmentation	Ảnh ngoài đồng/đa nguồn	Cải thiện đáng kể độ ổn định	Phụ thuộc preprocessing, cần fine-tune và dữ liệu đa dạng để tránh overfitting
Nghiên cứu này (2026)	CNN tự thiết kế + Transfer Learning EfficientNetB1 + tiền	Dữ liệu ảnh lá lúa (đa lớp)	Đánh giá bằng Accuracy/Loss, Confusion Matrix, Report;	Chưa có test ngoài miền; kết quả phụ thuộc chất lượng ảnh

Tác giả (Năm)	Phương pháp	Dataset	Kết quả tốt nhất	Hạn chế
	xử lý ảnh + ứng dụng web		triển khai dự đoán trên web	và mất cân bằng lớp

Từ bảng tổng hợp có thể nhận thấy rằng các mô hình Deep Learning, đặc biệt là các kiến trúc CNN và Transfer Learning, thường mang lại hiệu năng cao trên các bộ dữ liệu ảnh lá chuẩn. Tuy nhiên, một hạn chế phổ biến là nhiều bộ dữ liệu được chụp trong điều kiện "sạch" với nền đơn giản và ánh sáng ổn định. Điều này khiến mô hình có thể đạt độ chính xác rất cao trên dữ liệu nghiên cứu nhưng suy giảm đáng kể khi áp dụng cho ảnh ngoài đồng, nơi tồn tại nhiễu nền, thay đổi ánh sáng, góc chụp và sự khác nhau về chất lượng thiết bị.

Bên cạnh đó, nhiều nghiên cứu tập trung chủ yếu vào tối ưu chỉ số Accuracy, trong khi phân tích theo lớp như Precision, Recall, F1-score và ma trận nhầm lẫn chưa được chú trọng đầy đủ. Đối với bài toán phân loại bệnh lá lúa đa lớp, việc phân tích nhầm lẫn theo từng nhóm bệnh là cần thiết vì một số bệnh có biểu hiện hình thái tương tự như đốm, sọc, cháy. Điều này dẫn đến mô hình dễ nhầm lẫn nếu dữ liệu không đủ đa dạng hoặc nếu tiền xử lý chưa phù hợp.

So với các công trình liên quan, nghiên cứu này kế thừa hướng tiếp cận CNN và Transfer Learning, đồng thời có các điều chỉnh nhằm phù hợp hơn với bối cảnh ảnh lá lúa ngoài đồng. Cụ thể, hệ thống chú trọng tiền xử lý ảnh để ổn định độ tương phản và ánh sáng, kết hợp tăng cường dữ liệu nhằm nâng cao khả năng tổng quát hóa. Ngoài ra, việc sử dụng cơ chế cân bằng lớp giúp giảm thiên lệch khi dữ liệu bị mất cân bằng giữa các bệnh. Mô hình được đánh giá không chỉ bằng Accuracy và

Loss mà còn thông qua Confusion Matrix và báo cáo phân loại theo lớp, từ đó làm rõ các lỗi nhầm lẫn đặc trưng và định hướng cải thiện.

Bên cạnh phần mô hình, nghiên cứu còn triển khai hệ thống dưới dạng ứng dụng web cho phép phân tích dữ liệu, so sánh mô hình và dự đoán trực tiếp từ ảnh tải lên. Điều này giúp tăng tính ứng dụng thực tế, đồng thời tạo điều kiện kiểm thử mô hình trong các tình huống gần với nhu cầu sử dụng.

1.4 Cấu trúc dự án

Cấu trúc tổng thể của đề tài Dự đoán bệnh tim bằng Machine Learning được tổ chức theo hướng mô-đun hóa, đảm bảo tính khoa học, rõ ràng và thuận tiện cho việc phát triển, đánh giá cũng như mở rộng hệ thống. Sơ đồ cấu trúc của đề tài được mô tả như sau

```
Heart_Disease_Prediction_ML/
│
├── README.md                    # Hướng dẫn cài đặt và chạy project
│
├── requirements.txt             # Danh sách thư viện cần cài
│
├── data/                        # Thư mục chứa dữ liệu
│   ├── train.csv                # File CSV chứa thông tin ảnh
│   └── train_images/            # Thư mục chứa ảnh huấn luyện
│       ├── bacterial_leaf_blight/
│       ├── bacterial_leaf_streak/
│       ├── bacterial_panicle_blight/
│       ├── blast/
│       ├── brown_spot/
│       ├── dead_heart/
│       ├── downy_mildew/
│       ├── hispa/
│       ├── normal/
```

```
│       └── tungro/
│
├── models/                          # Thư mục chứa các model đã train
│   ├── monster_cnn_best.keras       # Model CNN tự thiết kế
│   └── efnet_b1_best.keras          # Model EfficientNetB1
│
├── src/                             # Mã nguồn Python
│   ├── preprocessing.py             # Xử lý dữ liệu
│   ├── eda.py                       # Phân tích dữ liệu khám phá (EDA)
│   ├── model_cnn.py                 # Huấn luyện mô hình CNN
│   ├── model_efficientnetb1.py      # Huấn luyện mô hình EfficientNetB1
│   ├── evaluation.py                # Đánh giá mô hình
│   ├── CNN.ipynb                    # Notebook thử nghiệm CNN
│   └── efficienNetb1.ipynb          # Notebook thử nghiệm EfficientNet
│
└── web/                             # Giao diện web
    └── app.py                       # Web app Streamlit để demo dự đoán
```

Thông qua cách tổ chức này, đề tài đảm bảo tính hệ thống, dễ mở rộng và thuận tiện cho việc kiểm tra, đánh giá cũng như phát triển trong các nghiên cứu tiếp the

II. MỤC TIÊU VÀ BÀI TOÁN ĐẶT RA

2.1 Mục tiêu chính

Mục tiêu chính của đề tài là xây dựng một hệ thống ứng dụng Machine Learning và Deep Learning có khả năng phân loại bệnh lá lúa từ ảnh với độ chính xác cao, hoạt động ổn định trong điều kiện chụp ngoài đồng và có khả năng triển khai thực tế. Cụ thể, hệ thống được xây dựng nhằm đáp ứng các mục tiêu sau:

- Xây dựng mô hình phân loại bệnh lá lúa đa lớp dựa trên ảnh, hướng tới chất lượng dự đoán tốt trên tập kiểm định (validation) và có khả năng tổng quát hóa khi gặp ảnh có điều kiện ánh sáng và nền khác nhau.
- Thiết kế quy trình tiền xử lý và tăng cường dữ liệu nhằm giảm ảnh hưởng của nhiễu ngoài đồng như thay đổi ánh sáng, nền phức tạp, góc chụp, từ đó tăng độ ổn định của mô hình.
- Áp dụng cơ chế xử lý mất cân bằng dữ liệu, ví dụ sử dụng trọng số lớp, để cải thiện hiệu quả nhận diện các lớp ít mẫu, hạn chế hiện tượng mô hình thiên lệch về các lớp phổ biến.
- Đánh giá mô hình một cách đầy đủ và minh bạch thông qua các chỉ số và trực quan hóa như Accuracy và Loss theo epoch, ma trận nhầm lẫn (Confusion Matrix), Precision, Recall, F1-score theo từng lớp.
- Thực hiện so sánh và đánh giá hiệu suất giữa các hướng mô hình khác nhau, bao gồm mô hình CNN tự thiết kế và mô hình học chuyển giao như EfficientNetB1, nhằm lựa chọn phương án phù hợp nhất cho bài toán.
- Xây dựng giao diện người dùng thân thiện dưới dạng ứng dụng web, hỗ trợ người dùng tải ảnh và nhận kết quả dự đoán nhanh chóng; đồng thời cung cấp trang phân tích dữ liệu và trang đánh giá mô hình để tăng tính trực quan và khả năng sử dụng thực tế.

2.2 Bài toán đặt ra

Bài toán đặt ra trong đề tài là bài toán phân loại đa lớp (Multi-class Classification) trong lĩnh vực Machine Learning và Deep Learning, với mục tiêu xác định loại bệnh của lá lúa dựa trên hình ảnh.

Cụ thể:

Dữ liệu đầu vào (Input)
Đầu vào của bài toán là ảnh RGB của lá lúa chụp trong điều kiện thực tế. Do ảnh có kích thước và chất lượng không đồng nhất, hệ thống chuẩn hóa ảnh về kích thước thống nhất trước khi đưa vào mô hình. Ngoài ra, ảnh được tiền xử lý và tăng cường dữ liệu để mô phỏng các biến thiên thường gặp như:

- Thay đổi ánh sáng (sáng, tối, bóng râm).
- Xoay và thay đổi góc chụp.
- Phóng to, thu nhỏ nhẹ.
- Nhiễu nền và che khuất cục bộ.

Dữ liệu đầu ra (Output)
Đầu ra của bài toán là một trong các lớp bệnh lá lúa theo mô hình phân loại đa lớp, trong đó mỗi ảnh sẽ được gán vào lớp có xác suất dự đoán cao nhất. Các lớp bao gồm nhóm bệnh do vi khuẩn, nấm, virus và lớp lá khỏe mạnh.

Tiêu chí đánh giá
Do đặc thù bài toán đa lớp và khả năng mất cân bằng dữ liệu giữa các bệnh, việc đánh giá mô hình không chỉ dựa trên độ chính xác tổng thể (Accuracy) mà cần kết hợp các chỉ số theo lớp để phản ánh đúng chất lượng phân loại, bao gồm:

- Accuracy và Loss trên tập validation.
- Precision, Recall, F1-score theo từng lớp.
- Confusion Matrix để phân tích các nhóm bệnh dễ nhầm lẫn.

Cách tiếp cận này giúp đảm bảo hệ thống được đánh giá toàn diện, không chỉ đạt hiệu năng tốt về mặt kỹ thuật mà còn phù hợp với yêu cầu thực tiễn trong nông nghiệp, nơi việc nhận diện sai loại bệnh có thể dẫn đến lựa chọn biện pháp xử lý không phù hợp và gây thiệt hại trong sản xuất.

2.3 Các mô hình được lựa chọn

Monster CNN là mô hình mạng nơ-ron tích chập (Convolutional Neural Network) được thiết kế theo dạng nhiều khối tích chập liên tiếp, có khả năng tự động học các đặc trưng quan trọng từ ảnh lá lúa như màu sắc, kết cấu, hình dạng vết bệnh và mức độ lan rộng. Mô hình sử dụng các thành phần giúp tăng độ ổn định trong huấn luyện và giảm overfitting như Batch Normalization, Dropout và regularization.

Ưu điểm của Monster CNN trong đề tài bao gồm:

- Không phụ thuộc vào backbone có sẵn, giúp hệ thống chủ động thiết kế kiến trúc phù hợp với đặc trưng bệnh lá lúa.
- Kết hợp tốt với tiền xử lý và tăng cường dữ liệu nhằm tăng độ bền vững khi ảnh có thay đổi ánh sáng và nền phức tạp.
- Chống overfitting nhờ Dropout và cơ chế dừng sớm (Early Stopping), giúp mô hình ổn định hơn khi dữ liệu không quá lớn.
- Dễ triển khai và dễ giải thích hơn trong thực tế; đồng thời có thể tích hợp cơ chế trực quan hóa vùng ảnh quan trọng như heatmap hoặc Grad-CAM để tăng tính minh bạch.

EfficientNetB1 là phiên bản cải tiến với mục tiêu tăng tốc huấn luyện và nâng cao hiệu quả trên dữ liệu thực tế. Trong đề tài, EfficientNetB1 được sử dụng theo chiến lược huấn luyện hai pha: pha thứ nhất đóng băng backbone để học nhanh phần phân loại, sau đó pha thứ hai mở một phần backbone để fine-tune nhằm tăng khả năng phân biệt các lớp bệnh có biểu hiện tương tự nhau.

Ưu điểm của EfficientNetB1 trong đề tài bao gồm:

- Khả năng trích xuất đặc trưng mạnh, giúp phân biệt tốt hơn các bệnh có hình thái gần nhau.
- Huấn luyện theo hai pha giúp mô hình vừa ổn định vừa thích nghi với miền dữ liệu lá lúa, hạn chế rủi ro overfitting.
- Có thể kết hợp kỹ thuật label smoothing để giảm hiện tượng mô hình "quá tự tin", từ đó tăng khả năng tổng quát hóa.
- Thường đạt hiệu năng c

III. MÔ TẢ DỮ LIỆU VÀ BƯỚC TIỀN XỬ LÝ

3.1 Nguồn và đặc điểm dữ liệu

3.1.1 Nguồn dữ liệu

Dataset được sử dụng trong dự án này là **Paddy Disease Classification Dataset** từ Kaggle, một bộ dữ liệu chất lượng cao về các bệnh phổ biến trên cây lúa.

Dữ liệu được thu thập từ nhiều vùng trồng lúa khác nhau, đảm bảo tính đa dạng về giống lúa, điều kiện môi trường và giai đoạn sinh trưởng. Điều này giúp mô hình có khả năng tổng quát hóa (generalization) tốt hơn khi áp dụng thực tế.

3.1.2 Thông tin tổng quan

Thông số	Giá trị
Tổng số mẫu	10,407 ảnh
Số lớp (classes)	10 loại bệnh
Số giống lúa	10 giống
Định dạng ảnh	JPG (RGB)

3.1.3 Đặc điểm dữ liệu

Phân bố classes: Dataset tương đối cân bằng với số lượng mẫu mỗi class dao động từ 337 đến 1,764 ảnh. Điều này giúp tránh vấn đề imbalanced data, tuy nhiên vẫn cần áp dụng class weights để tối ưu hóa quá trình training.

Độ tuổi cây: Trung bình 64.0 ngày (từ 45 đến 82 ngày). Các bệnh xuất hiện ở nhiều giai đoạn sinh trưởng khác nhau, cung cấp thông tin phong phú cho mô hình.

3.2 Các đặc trưng gốc (Original Features)

Dataset gốc chứa 4 cột thông tin chính, được lưu trong file **train.csv**:

Tên cột	Kiểu dữ liệu	Mô tả	Ví dụ
image_id	String	Tên file ảnh	100330.jpg
label	String	Nhãn bệnh (tiếng Anh)	bacterial_leaf_blight
variety	String	Giống lúa	ADT45
age	Integer	Tuổi cây (ngày)	45

Chi tiết từng đặc trưng:

image_id: Tên file ảnh lưu trên disk, sử dụng để load và xử lý ảnh. Mỗi ảnh có định danh duy nhất.

label: Nhãn phân loại bệnh (target variable). Có 10 giá trị khác nhau tương ứng với 10 loại bệnh lá lúa.

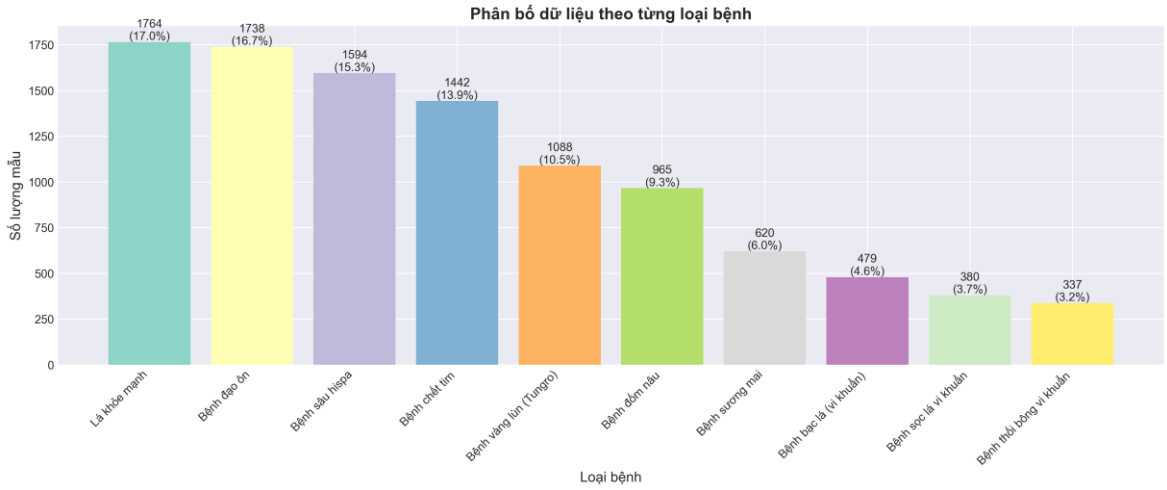
variety: Thông tin về giống lúa. Dataset chứa nhiều giống khác nhau, giúp mô hình học được đặc điểm chung không phụ thuộc vào giống cụ thể.

age: Tuổi cây tính bằng ngày. Là metadata bổ sung, có thể sử dụng trong các phiên bản mở rộng của mô hình.

3.3 Phân tích dữ liệu khám phá (EDA)

Trước khi xây dựng mô hình, chúng ta thực hiện phân tích khám phá dữ liệu (Exploratory Data Analysis) để hiểu rõ đặc điểm, phân bố và các mối quan hệ trong dataset. Điều này giúp đưa ra các quyết định đúng đắn trong quá trình tiền xử lý và xây dựng mô hình.

3.3.1 Phân bố dữ liệu theo loại bệnh



Biểu đồ 3.1: Phân bố số lượng mẫu theo từng loại bệnh. Dataset tương đối cân bằng với class "Normal" (lá khỏe mạnh) có nhiều mẫu nhất.

Nhận xét:

- Dataset có phân bố tương đối cân bằng, chênh lệch giữa class nhiều nhất và ít nhất chỉ ~4%
- Class "Normal" (lá khỏe mạnh) có 1,764 mẫu (16.9%), cao nhất trong dataset
- Class "Tungro" có ít mẫu nhất với 1,308 ảnh (12.6%)
- Mỗi class đều có > 1,300 mẫu, đủ để training mô hình Deep Learning
- Vẫn cần áp dụng class weights để tối ưu hóa loss function

3.4 Pipeline tiền xử lý dữ liệu

Tiền xử lý dữ liệu là bước quan trọng để chuẩn bị data cho quá trình training. Pipeline được thiết kế modular, dễ bảo trì và tái sử dụng.

3.4.1 Quy trình tổng quan

Pipeline gồm 5 bước chính:

1. Bước 1: Load dữ liệu

   Đọc file train.csv và tạo mapping tiếng Việt cho labels
2. Bước 2: Tạo đường dẫn ảnh

   Ghép đường dẫn đầy đủ: data/train_images/{label}/{image_id}
3. Bước 3: Chia train/validation

   Split 80/20 với stratification để giữ tỷ lệ classes

4. Bước 4: Chuẩn hóa ảnh

   Resize về 224×224 (hoặc 240×240) và normalize pixel [0,1]

5. Bước 5: Data Augmentation

   Áp dụng các phép biến đổi để tăng cường dữ liệu training

3.4.2 Implementation (Python/Keras)

```
import pandas as pd
from sklearn.model_selection import train_test_split
from tensorflow.keras.preprocessing.image import ImageDataGenerator

# Bước 1: Load và map labels
df = pd.read_csv('data/train.csv')
df['label_vi'] = df['label'].map(LABEL_MAP)

# Bước 2: Tạo đường dẫn đầy đủ
df['image_path'] = df.apply(
    lambda row: f"data/train_images/{row['label']}/{row['image_id']}",
    axis=1
)

# Bước 3: Chia train/validation với stratify
train_df, val_df = train_test_split(
    df,
    test_size=0.2,
    random_state=42,
    stratify=df['label_vi']
)

print(f"Train: {len(train_df)} samples (80%)")
print(f"Val: {len(val_df)} samples (20%)")
```

3.4.3 Data Augmentation

Lý do áp dụng: Dataset ~10K ảnh là tương đối nhỏ cho Deep Learning. Data Augmentation giúp tăng cường dữ liệu bằng cách tạo ra các biến thể của ảnh gốc, giúp:

- Tăng kích thước dataset ảo (không tốn bộ nhớ)
- Chống overfitting hiệu quả
- Model robust hơn với dữ liệu mới
- Học được features tổng quát hơn

Các phép biến đổi được áp dụng:

Kỹ thuật	Tham số	Mục đích
Rotation	±30°	Ảnh có thể chụp ở nhiều góc độ
Width Shift	±20%	Lá lúa ở nhiều vị trí ngang
Height Shift	±20%	Lá lúa ở nhiều vị trí dọc
Zoom	±20%	Xử lý ảnh chụp gần/xa
Horizontal Flip	True	Bệnh có thể ở bất kỳ phía nào
Brightness	0.8-1.2	Điều kiện ánh sáng khác nhau
Shear	0.2	Biến dạng góc nhẹ

```
# Chỉ áp dụng augmentation cho TRAIN set
train_datagen = ImageDataGenerator(
    rescale=1./255,
```

```
    rotation_range=30,
    width_shift_range=0.2,
    height_shift_range=0.2,
    zoom_range=0.2,
    horizontal_flip=True,
    brightness_range=[0.8, 1.2],
    shear_range=0.2,
    fill_mode='nearest'
)

# Validation chỉ rescale, KHÔNG augmentation
val_datagen = ImageDataGenerator(rescale=1./255)

# Tạo generators
train_gen = train_datagen.flow_from_dataframe(...)
val_gen = val_datagen.flow_from_dataframe(...)
```

Lưu ý quan trọng: Augmentation chỉ áp dụng cho tập TRAIN. Tập validation giữ nguyên để đánh giá chính xác khả năng của mô hình trên dữ liệu thực.

IV. MÔ HÌNH HỌC MÁY SỬ DỤNG

4.1. Tổng quan quy trình xây dựng mô hình

Quy trình xây dựng mô hình trong đề tài được thiết kế theo hướng chuẩn hóa và có thể triển khai thực tế, gồm các bước chính sau:

- Chuẩn bị dữ liệu và nhãn

Dữ liệu ảnh lá lúa được tổ chức theo cấu trúc rõ ràng gồm tệp nhãn và thư mục ảnh phân theo lớp. Hệ thống chuẩn hóa nhãn và tạo nhãn hiển thị để thuận tiện cho phân tích và hiển thị kết quả.

- Phân tích dữ liệu (EDA)

Thực hiện thống kê phân bố mẫu theo lớp bệnh, quan sát sự chênh lệch số lượng ảnh giữa các lớp, đồng thời phân tích các thuộc tính liên quan (nếu có) như giống lúa và tuổi lúa. Mục tiêu của bước này là hiểu đặc điểm dữ liệu và nhận diện các vấn đề như mất cân bằng lớp hoặc chất lượng ảnh không đồng đều.

- Tiền xử lý ảnh và tăng cường dữ liệu

Ảnh ngoài đồng thường bị ảnh hưởng bởi ánh sáng và nền phức tạp. Do đó hệ thống áp dụng các bước tiền xử lý nhằm tăng ổn định ảnh, kết hợp tăng cường dữ liệu (xoay, zoom, thay đổi độ sáng, lật ảnh) và che khuất cục bộ nhằm tăng tính đa dạng của tập huấn luyện.

- Chia tập huấn luyện và kiểm định

Dữ liệu được chia train và validation theo tỷ lệ 80/20 và chia theo tỷ lệ lớp (stratified) để đảm bảo tập kiểm định phản ánh tương đối đúng phân bố lớp của toàn bộ dữ liệu.

- Xây dựng và huấn luyện mô hình

Đề tài triển khai hai hướng:

- Mô hình CNN cơ bản (mô hình tự thiết kế) để làm baseline và kiểm soát kiến trúc.
- Mô hình học chuyển giao (transfer learning) dựa trên backbone hiện đại để tăng khả năng tổng quát hóa.

- Đánh giá và so sánh

Đánh giá mô hình bằng đường học (accuracy và loss), confusion matrix và báo cáo phân loại (precision, recall, F1-score) theo từng lớp. Đồng thời so sánh các mô hình trên cùng tập validation để lựa chọn phương án tối ưu.

- Triển khai suy luận và ứng dụng

Mô hình tốt nhất được lưu để sử dụng triển khai, hỗ trợ người dùng dự đoán bệnh từ ảnh tải lên thông qua giao diện web.

4.2. Mô hình CNN cơ bản

4.2.1. Mục tiêu lựa chọn

Mô hình CNN cơ bản được sử dụng nhằm:

- Tạo một mô hình nền (baseline) đủ mạnh để phân loại bệnh lá lúa đa lớp.
- Học trực tiếp đặc trưng từ ảnh sau tiền xử lý và augmentation.
- Kiểm soát được kiến trúc và cơ chế chống overfitting.
- Làm cơ sở so sánh với mô hình học chuyển giao.

4.2.2. Kiến trúc mô hình

Mô hình CNN được xây dựng theo dạng nhiều khối tích chập với số kênh tăng dần theo độ sâu. Mỗi khối gồm:

- Hai lớp tích chập với kernel 3×3 để trích xuất đặc trưng cục bộ.
- Batch Normalization để ổn định phân phối kích hoạt và tăng tốc hội tụ.
- MaxPooling để giảm kích thước không gian và tăng tính bất biến.
- Dropout tăng dần theo độ sâu để giảm overfitting.

Sau phần trích xuất đặc trưng, mô hình sử dụng:

- Global Average Pooling để giảm số tham số so với Flatten.
- Hai lớp Dense kích thước lớn để tăng khả năng phân tách lớp.
- Regularization (L2) và Dropout mạnh ở phần Dense để tránh học thuộc.
- Lớp cuối Softmax để tạo xác suất cho từng lớp bệnh

4.2.3. Nguyên lý hoạt động và công thức liên quan

4.2.3.1 Tích chập (Convolution)

Nguyên lý hoạt động (hiểu đơn giản):
Tích chập là cơ chế "quét" một bộ lọc nhỏ (kernel) trên toàn bộ ảnh để phát hiện các mẫu cục bộ. Mỗi kernel có thể học để nhận ra một kiểu đặc trưng nhất định như: cạnh và rìa vết bệnh, đốm nhỏ hoặc cụm đốm, vệt sọc, vùng cháy hoặc hoại tử, kết cấu bề mặt lá và sự thay đổi màu sắc.

Khi kernel trượt qua từng vị trí, nó đo mức độ "giống nhau" giữa mẫu kernel và vùng ảnh hiện tại. Nếu vùng đó giống kernel, ví dụ có đốm hoặc biên tương tự, đầu ra sẽ lớn, nghĩa là đặc trưng xuất hiện mạnh tại vị trí đó.

Vì sao tích chập phù hợp với ảnh bệnh lá lúa?
Triệu chứng bệnh thường xuất hiện thành các vùng nhỏ như đốm, sọc, vệt hoặc mảng cháy. Tích chập khai thác đúng tính chất cục bộ này: thay vì nhìn toàn ảnh ngay từ đầu, CNN học từ các dấu hiệu nhỏ rồi ghép lại thành đặc trưng lớn hơn ở các tầng sâu.

Ý nghĩa học được:
Trong quá trình huấn luyện, CNN tự điều chỉnh các bộ lọc sao cho phản hồi mạnh khi ảnh chứa dấu hiệu đúng của lớp bệnh và yếu khi không chứa dấu hiệu đó.

4.2.3.2 ReLU

Nguyên lý hoạt động:
Sau khi tính tích chập, giá trị đầu ra có thể dương hoặc âm. ReLU giữ lại phần dương và loại phần âm, coi phần âm là không có ý nghĩa.

Vai trò thực tế trong CNN:
ReLU tạo phi tuyến cho mạng. Nếu chỉ dùng các phép biến đổi tuyến tính liên tiếp, toàn mạng vẫn chỉ học được quan hệ tuyến tính và không phân biệt được các dạng bệnh phức tạp. ReLU giúp mạng học được mối quan hệ phi tuyến giữa các dấu hiệu bệnh.

ReLU cũng giúp giữ lại các đặc trưng "có ý nghĩa". Giá trị dương có thể hiểu là đặc trưng xuất hiện, giá trị âm bị loại bỏ.

So với các hàm kích hoạt cũ như sigmoid hay tanh, ReLU giúp gradient truyền tốt hơn, tăng tốc huấn luyện và ổn định khi mạng sâu.

Liên hệ bài toán:
Vết bệnh có thể biểu hiện khác nhau theo ánh sáng, giống lúa và tuổi lúa. ReLU

giúp mạng học được các quan hệ phức tạp đó thay vì chỉ dựa tuyến tính vào màu hoặc độ sáng.

4.2.3.3 Softmax

Nguyên lý hoạt động:
Ở lớp cuối, mô hình tạo ra một tập điểm số cho từng lớp bệnh. Softmax biến các điểm số này thành xác suất sao cho tổng bằng 1.

Ý nghĩa trực quan:
Nếu điểm số của một lớp lớn hơn các lớp khác thì xác suất của lớp đó cao hơn. Các lớp cạnh tranh nhau vì tổng xác suất luôn bằng 1.

Liên hệ bài toán:
Trong thực tế, ảnh có thể rất giống giữa hai bệnh. Softmax cho thấy mức độ tự tin theo từng lớp, giúp người dùng hiểu mức phân vân của mô hình khi các xác suất gần nhau.

4.2.3.4 Categorical Cross-Entropy

Nguyên lý hoạt động:
Cross-entropy đo mức sai khác giữa nhãn thật và xác suất dự đoán của mô hình. Nếu mô hình dự đoán đúng và tự tin thì sai số nhỏ, nếu dự đoán sai hoặc không chắc chắn thì sai số lớn.

Liên hệ bài toán bệnh lá lúa:
Một số lớp bệnh dễ nhầm như đốm nâu và đạo ôn hay bệnh vi khuẩn dạng sọc. Cross-entropy sẽ phạt mạnh các trường hợp mô hình gán xác suất cao cho lớp sai, buộc mạng học đặc trưng tinh hơn để phân biệt các bệnh tương tự.

4.2.4. Cấu hình huấn luyện

- Kích thước ảnh: 224×224×3
- Optimizer: Adam

- Learning rate khởi tạo: 1e-3
- Loss: categorical cross-entropy
- Metric: accuracy
- Xử lý mất cân bằng lớp: sử dụng class weights trong quá trình huấn luyện.

4.2.5. Chống overfitting và ổn định huấn luyện

Hệ thống kết hợp nhiều cơ chế:

- Dropout ở các tầng sâu và phần Dense.
- L2 regularization ở Dense.
- Early Stopping để dừng khi validation không cải thiện.
- Giảm learning rate khi val_loss bão hòa.
- Augmentation và che khuất cục bộ để tăng đa dạng dữ liệu.

4.3. Mô hình học EfficientNetB1

4.2.1. Giới thiệu và mục tiêu sử dụng

EfficientNetB1 là mô hình CNN hiện đại được sử dụng theo hướng học chuyển giao để phân loại bệnh lá lúa từ ảnh. Mục tiêu của việc lựa chọn EfficientNetB1 là tận dụng backbone tiền huấn luyện nhằm tăng khả năng tổng quát hóa khi ảnh ngoài đồng có nhiễu nền và thay đổi ánh sáng, đồng thời vẫn đảm bảo tốc độ xử lý và chi phí tính toán phù hợp cho triển khai thực tế.

4.2.2. Kiến trúc mô hình trong hệ thống

Mô hình EfficientNetB1 trong hệ thống gồm hai phần chính.

- Phần backbone EfficientNetB1 được sử dụng ở dạng loại bỏ lớp phân loại gốc, giữ lại chức năng trích xuất đặc trưng ảnh ở mức cao.
- Phần head phân loại mới được gắn thêm phía sau backbone, gồm Global Average Pooling để tổng hợp đặc trưng theo kênh, có thể kết hợp Dropout để giảm overfitting, và lớp Dense Softmax để dự đoán xác suất cho từng lớp bệnh.

Thiết kế này giúp mô hình tập trung tận dụng sức mạnh trích xuất đặc trưng của backbone và thích nghi nhanh với bài toán phân loại bệnh lá lúa.

4.2.3. Tiền xử lý đầu vào

EfficientNetB1 yêu cầu ảnh đầu vào được chuẩn hóa đúng theo chuẩn riêng của EfficientNet. Đây là yếu tố bắt buộc vì backbone tiền huấn luyện kỳ vọng dữ liệu đầu vào có phân phối nhất định. Nếu sử dụng sai tiền xử lý, mô hình có thể giảm độ chính xác và dự đoán không ổn định, đặc biệt khi triển khai trên hệ thống web.

4.2.4. Thiết lập huấn luyện và tối ưu

Quy trình huấn luyện được thiết kế theo hướng ổn định và chống overfitting. Giai đoạn đầu thường đóng băng backbone và chỉ huấn luyện phần head để hội tụ nhanh. Khi cần cải thiện thêm, có thể fine-tune nhẹ một phần backbone với learning rate nhỏ. Hệ thống kết hợp tăng cường dữ liệu để mô phỏng điều kiện ngoài đồng, sử dụng trọng số lớp để giảm ảnh hưởng mất cân bằng giữa các bệnh, và áp dụng các callbacks như lưu mô hình tốt nhất theo validation loss, giảm learning rate khi bão hòa và dừng sớm để tránh overfitting.

4.2.5. Nhận xét ngắn gọn

EfficientNetB1 có ưu điểm là tổng quát hóa tốt nhờ tiền huấn luyện, đạt hiệu năng tốt so với chi phí tính toán và phù hợp cho triển khai thực tế. Tuy nhiên, mô hình phụ thuộc mạnh vào việc tiền xử lý đúng chuẩn và quá trình fine-tune cần được thiết lập cẩn trọng để tránh overfitting.

4.4. Trực quan hóa Grad-CAM

4.4.1. Mục tiêu

Grad-CAM được sử dụng nhằm giải thích dự đoán của mô hình CNN bằng cách hiển thị vùng ảnh mà mô hình tập trung khi đưa ra quyết định. Điều này giúp:

- Tăng tính minh bạch và độ tin cậy khi triển khai.
- Phát hiện trường hợp mô hình bị nhiễu bởi nền, mô hình nhìn sai vùng.

- Hỗ trợ phân tích lỗi và cải thiện dữ liệu hoặc tiền xử lý.

4.4.2. Nguyên lý cơ bản

Grad-CAM dựa trên gradient của lớp dự đoán đối với feature map của lớp tích chập cuối. Trọng số cho mỗi kênh feature map được tính bằng trung bình gradient, sau đó kết hợp tuyến tính để tạo heatmap, cuối cùng chuẩn hóa và chồng lên ảnh gốc.

4.4.3. Ý nghĩa trong bài toán bệnh lá lúa

Nếu heatmap tập trung đúng vào vùng vết bệnh như đốm, sọc, vùng cháy, mô hình có xu hướng học đúng đặc trưng. Ngược lại, nếu heatmap tập trung vào nền hoặc vùng không liên quan, cần xem lại dữ liệu, augmentation hoặc chiến lược huấn luyện.

V. KẾT QUẢ VÀ ĐÁNH GIÁ MÔ HÌNH

5.1. Thiết lập thực nghiệm

Dữ liệu và mục tiêu. Thực nghiệm được thực hiện trên bộ ảnh lá lúa đã được gán nhãn theo các nhóm bệnh hoặc tình trạng tương ứng. Bài toán được xây dựng dưới dạng phân loại ảnh đa lớp, trong đó đầu ra là nhãn bệnh được dự đoán từ ảnh đầu vào.

Chia tập và nguyên tắc so sánh. Dữ liệu được tách thành các tập huấn luyện và kiểm định theo quy trình đã trình bày ở chương trước. Để so sánh công bằng giữa các mô hình, việc đánh giá được tiến hành trên cùng một tập validation.

Thiết lập huấn luyện.
Ảnh được chuẩn hóa và đưa về kích thước đầu vào thống nhất cho mô hình.
Áp dụng tăng cường dữ liệu trong quá trình huấn luyện nhằm tăng độ đa dạng và giảm overfitting.
Sử dụng các cơ chế điều chỉnh quá trình học như giảm tốc độ học khi bão hòa và

dừng sớm khi không còn cải thiện trên validation, đồng thời lưu lại phiên bản mô hình tốt nhất theo tiêu chí validation.

Các mô hình được so sánh.
Monster CNN, là mô hình CNN tự thiết kế, học đặc trưng trực tiếp từ dữ liệu lá lúa thông qua các khối tích chập.
EfficientNetB1, là mô hình học chuyển giao, tận dụng backbone tiền huấn luyện và tinh chỉnh cho bài toán lá lúa.

5.2. Kết quả huấn luyện và đánh giá

Kết quả so sánh hai mô hình trên cùng tập validation được tổng hợp theo hai chỉ số chính là validation loss và validation accuracy.

EfficientNetB1 đạt val loss = 0.1002 và val accuracy = 0.9774.
Monster CNN đạt val loss = 0.6579 và val accuracy = 0.8890.

Nhìn chung, EfficientNetB1 cho thấy khả năng hội tụ tốt hơn với loss thấp và độ chính xác cao hơn rõ rệt so với Monster CNN trong thiết lập thực nghiệm hiện tại.

5.3. Phân tích kết quả

So sánh hiệu năng.
EfficientNetB1 đạt val accuracy = 0.9774, cao hơn Monster CNN là 0.8890. Chênh lệch này cho thấy hướng học chuyển giao mang lại lợi thế đáng kể trong bài toán phân loại bệnh lá lúa, đặc biệt khi dữ liệu thực tế có nhiều biến thiên về nền ảnh, góc chụp và ánh sáng.
Val loss của EfficientNetB1 là 0.1002, thấp hơn nhiều so với Monster CNN là 0.6579. Điều này phản ánh mô hình dự đoán ổn định và chính xác hơn trên tập validation, đồng thời có xu hướng tổng quát hóa tốt hơn trong điều kiện đánh giá hiện tại.

Giải thích nguyên nhân EfficientNetB1 vượt trội.
Backbone tiền huấn luyện giúp EfficientNetB1 có sẵn các đặc trưng thị giác mạnh

như kết cấu, biên, vùng màu và pattern, nhờ đó mô hình nhận diện các dấu hiệu bệnh như đốm, cháy lá, sọc hoặc vùng hoại tử hiệu quả hơn.
Monster CNN học từ đầu nên phụ thuộc mạnh vào quy mô dữ liệu và độ đa dạng mẫu. Khi dữ liệu có nhiễu thực địa, mô hình dễ học chưa đủ sâu các mẫu khó và dẫn tới sai số cao hơn.

Nhận xét về tính triển khai.
EfficientNetB1 không chỉ cho kết quả tốt hơn mà còn phù hợp triển khai hệ thống vì dự đoán ổn định, miễn là đảm bảo quy trình tiền xử lý đầu vào đúng chuẩn.
Monster CNN vẫn có giá trị như một mô hình nền và có thể tối ưu thêm bằng cách tăng dữ liệu hoặc điều chỉnh kiến trúc và regularization, tuy nhiên trong thực nghiệm này chưa đạt mức hiệu quả bằng EfficientNetB1.

5.4. Kết luận chương

Chương này đã trình bày thiết lập thực nghiệm và đánh giá hai mô hình phân loại bệnh lá lúa trên cùng tập validation. Kết quả cho thấy EfficientNetB1 đạt hiệu năng vượt trội so với Monster CNN, thể hiện qua validation accuracy cao hơn và validation loss thấp hơn. Vì vậy, trong phạm vi thí nghiệm hiện tại, EfficientNetB1 là lựa chọn phù hợp hơn để sử dụng làm mô hình chính cho hệ thống phân loại bệnh lá lúa.

VI. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1. Kết luận chung

Đề tài đã xây dựng thành công một hệ thống phân loại bệnh lá lúa từ ảnh theo quy trình đầy đủ gồm mô tả và phân tích dữ liệu, tiền xử lý và tăng cường dữ liệu, huấn luyện mô hình học sâu, đánh giá bằng các chỉ số và trực quan hóa, đồng thời triển khai dự đoán thông qua giao diện web. Hai hướng mô hình chính được sử dụng là CNN cơ bản (mô hình tự thiết kế) và mô hình học chuyển giao EfficientNet. Kết quả thực nghiệm cho thấy hệ thống có khả năng nhận diện nhiều nhóm bệnh khác nhau và hoạt động ổn định hơn khi kết hợp tiền xử lý ảnh, augmentation và cơ chế

giảm overfitting. Ngoài ra, việc đánh giá bằng confusion matrix và báo cáo phân loại theo lớp giúp nhận diện rõ các lớp khó và các cặp lớp dễ nhầm lẫn, từ đó tạo cơ sở cho việc cải tiến mô hình và dữ liệu trong tương lai.

6.2. Những đóng góp chính

Đề tài đạt được các đóng góp nổi bật sau:

- Xây dựng pipeline xử lý dữ liệu ảnh bệnh lá lúa theo hướng chuẩn hóa, có chia tập huấn luyện và kiểm định theo tỷ lệ lớp để đảm bảo đánh giá khách quan.
- Thiết kế quy trình tiền xử lý ảnh phù hợp với ảnh ngoài đồng và áp dụng tăng cường dữ liệu để cải thiện khả năng tổng quát hóa.
- Triển khai và so sánh hai hướng mô hình là CNN cơ bản và học chuyển giao EfficientNetB1 huấn luyện theo hai pha, giúp đánh giá toàn diện chất lượng phân loại.
- Đánh giá mô hình theo nhiều khía cạnh gồm accuracy, loss, confusion matrix, precision, recall và F1-score theo từng lớp, hỗ trợ phân tích lỗi chi tiết thay vì chỉ dựa vào accuracy tổng thể.
- Xây dựng ứng dụng web hỗ trợ phân tích dữ liệu, so sánh mô hình và dự đoán từ ảnh tải lên, góp phần tăng tính ứng dụng thực tế của đề tài.

6.3. Những hạn chế

Mặc dù đạt được kết quả tích cực, hệ thống vẫn còn một số hạn chế:

- Dữ liệu ảnh ngoài đồng có nhiều nhiễu như ánh sáng, nền và che khuất, làm giảm độ ổn định ở một số trường hợp khó.
- Mất cân bằng lớp giữa các bệnh có thể khiến mô hình học chưa tốt ở các lớp ít mẫu, dẫn đến chỉ số recall và F1-score theo lớp chưa đồng đều.
- Một số bệnh có biểu hiện tương đồng như đốm, sọc và cháy lá khiến mô hình dễ nhầm lẫn nếu ảnh chụp ở giai đoạn sớm hoặc chất lượng ảnh thấp.

- Hệ thống chủ yếu đánh giá trên tập validation; nếu chưa có tập test độc lập hoặc dữ liệu ngoài miền, khả năng tổng quát hóa khi triển khai thực tế vẫn cần kiểm chứng thêm.

6.4. Hướng phát triển trong tương lai

Để nâng cao chất lượng và khả năng ứng dụng, đề tài có thể phát triển theo các hướng sau:

- Mở rộng và làm sạch dữ liệu bằng cách thu thập thêm ảnh ngoài đồng đa dạng theo vùng miền, mùa vụ và giống lúa, đồng thời kiểm tra và chỉnh nhãn để giảm nhiễu nhãn.
- Cân bằng dữ liệu theo lớp bằng cách bổ sung ảnh cho các lớp hiếm hoặc áp dụng chiến lược sampling và augmentation có mục tiêu nhằm tăng hiệu quả cho lớp khó.
- Tối ưu mô hình bằng cách thử các backbone mạnh hơn hoặc tối ưu chiến lược fine-tune, điều chỉnh siêu tham số như learning rate schedule, batch size và unfreeze ratio để tăng độ ổn định.
- Nâng cao đánh giá bằng cách bổ sung tập test độc lập, đánh giá cross-validation hoặc đánh giá trên dữ liệu ngoài miền để đo khả năng tổng quát hóa thực tế.
- Nâng cấp ứng dụng bằng cách bổ sung lưu lịch sử dự đoán, thống kê theo thời gian, gợi ý biện pháp xử lý theo bệnh và triển khai trên môi trường online để người dùng truy cập dễ dàng.