

**VIETNAM NATIONAL UNIVERSITY  
UNIVERSITY OF ENGINEERING AND TECHNOLOGY**

-----\*\*\*-----



**MACHINE LEARNING PROJECT REPORT**

**Họ và Tên: Bùi Tùng Lâm**

**Mssv: 22021225**

**Lớp: K67-IT1**

**Hanoi, 2025**

## MỤC LỤC

<b>I. GIỚI THIỆU.....</b>	<b>4</b>
<b>II. XÂY DỰNG, PHÂN TÍCH VÀ MÔ TẢ DỮ LIỆU.....</b>	<b>6</b>
1. Gán nhãn bộ dữ liệu.....	6
2. Khám phá dữ liệu.....	6
3. Trình bày phân bố lớp hoặc giá trị đầu ra.....	7
<b>III.TIỀN XỬ LÝ DỮ LIỆU.....</b>	<b>7</b>
1. Làm sạch dữ liệu.....	7
2. Chuẩn hóa dữ liệu.....	7
<b>IV. XÂY DỰNG VÀ ĐÁNH GIÁ MÔ HÌNH.....</b>	<b>9</b>
1. Các mô hình khác nhau.....	9
2. Tối ưu, cải tiến mô hình.....	10
3. Đánh giá hiệu năng.....	10
<b>V. TRÌNH BÀY KẾT QUẢ.....</b>	<b>11</b>
1. Mô hình tốt nhất.....	11
2. Bảng biểu đồ minh họa.....	13
3. Điểm mạnh, hạn chế, hướng cải tiến.....	14

## **LỜI MỞ ĐẦU**

Bài làm của em còn nhiều thiếu sót và chưa được hoàn thiện một cách đầy đủ và trong bài làm có nhiều chỗ em cũng chưa thực sự hiểu do đây không phải là chuyên ngành của em và thời gian học và làm bài tập còn hơi ngắn. Trong quá trình thực hiện, em đã tham khảo nhiều tài liệu khác nhau và có sử dụng công cụ hỗ trợ (ChatGPT, Grok, ...) để hiểu rõ hơn nội dung và hoàn thành bài. Em rất mong nhận được sự thông cảm cũng như góp ý từ thầy/cô để em có thể học hỏi và cải thiện hơn trong những lần sau.

## I. GIỚI THIỆU

Cây lúa đóng vai trò quan trọng trong đời sống kinh tế và an ninh lương thực của nhiều quốc gia, đặc biệt tại khu vực Đông Nam Á – nơi Việt Nam là một trong những nước xuất khẩu gạo hàng đầu thế giới. Tuy nhiên, năng suất và chất lượng của cây lúa thường xuyên bị ảnh hưởng bởi các loại bệnh, trong đó phổ biến nhất là bệnh trên lá lúa. Lá là bộ phận chính đảm nhiệm chức năng quang hợp, do đó khi lá bị tổn thương, toàn bộ quá trình phát triển của cây sẽ bị ảnh hưởng nghiêm trọng. Việc nhận diện và phân loại chính xác các bệnh lá lúa là một bước quan trọng giúp người nông dân có thể đưa ra biện pháp xử lý phù hợp và kịp thời.

Tập dữ liệu được sử dụng trong nghiên cứu này là **“Rice Leaf Disease Image Dataset”** do Nirmal Sankalana công bố trên nền tảng Kaggle. Đây là một bộ dữ liệu hình ảnh gồm các bức ảnh chụp lá lúa trong điều kiện ánh sáng tự nhiên, với chất lượng rõ nét và được phân loại thủ công thành 4 nhóm bệnh/lớp hình ảnh cụ thể như sau:

- **Bacterial Leaf Blight (Bệnh bạc lá do vi khuẩn)**

Là một trong những bệnh phổ biến và nguy hiểm nhất ở lúa. Triệu chứng điển hình là các vết bạc màu, loang rộng từ đầu lá hoặc mép lá vào bên trong, thường xuất hiện sau mưa hoặc trong điều kiện ẩm ướt.

- **Brown Spot (Đốm nâu):**

Gây ra bởi nấm *Bipolaris oryzae*. Xuất hiện dưới dạng các đốm tròn màu nâu, kích thước từ nhỏ đến lớn, thường rải rác khắp bề mặt lá.

- **Leaf Smut (Than lá):**

Do nấm *Entyloma oryzae* gây ra. Biểu hiện là các vết sọc nhỏ, sẫm màu (thường đen hoặc nâu sậm) xuất hiện dọc theo gân lá.

- **Healthy (Lá khỏe mạnh):**

Các lá không bị bệnh, có màu xanh đồng đều, không có vết đốm hoặc sọc lạ. Đây là lớp giúp mô hình học được sự khác biệt giữa lá bị bệnh và lá bình thường.

Tập dữ liệu có tổng cộng hơn 1200 hình ảnh, được phân bố tương đối đồng đều giữa các lớp. Nhờ vào việc phân loại rõ ràng và chất lượng ảnh cao, bộ dữ liệu này rất phù hợp để áp dụng các kỹ thuật học sâu (deep learning), đặc biệt là mạng nơ-ron tích chập (CNN – Convolutional Neural Networks) trong bài toán phân loại hình ảnh.

Mục tiêu của bài báo cáo là xây dựng một mô hình có khả năng tự động nhận diện hình ảnh lá lúa và phân loại chúng vào một trong bốn lớp nêu trên với độ chính xác cao. Quá trình thực hiện bao gồm các bước chính:

- Tiền xử lý dữ liệu hình ảnh (chuẩn hóa kích thước, tăng cường dữ liệu...)
- Thiết kế kiến trúc mô hình CNN hoặc sử dụng các mô hình học chuyển tiếp (Transfer Learning) như ResNet, VGG16, MobileNet...
- Huấn luyện mô hình trên tập huấn luyện, đánh giá trên tập kiểm tra
- Đánh giá độ chính xác bằng các chỉ số như accuracy, precision, recall, F1-score và ma trận nhầm lẫn.

Mô hình sau khi huấn luyện có thể được ứng dụng vào thực tế như tích hợp vào điện thoại di động, máy ảnh nông nghiệp hoặc drone giám sát đồng ruộng, từ đó giúp người nông dân phát hiện bệnh nhanh chóng mà không cần đến chuyên gia. Ngoài ra, hệ thống còn có thể hỗ trợ trong các chương trình nông nghiệp thông minh, giúp tối ưu hóa chi phí và giảm thiểu tác động tiêu cực đến môi trường do lạm dụng thuốc bảo vệ thực vật.

Nghiên cứu này là bước đầu trong việc ứng dụng trí tuệ nhân tạo vào nông nghiệp hiện đại, góp phần hướng đến một nền nông nghiệp chính xác và bền vững hơn trong tương lai.

## II. XÂY DỰNG, PHÂN TÍCH VÀ MÔ TẢ DỮ LIỆU

### 1. Gán nhãn bộ dữ liệu

Bộ dữ liệu sử dụng bao gồm ba lớp bệnh lá lúa: "bacterial\_leaf\_blight", "brown\_spot" và "leaf\_smut". Mỗi lớp được tổ chức thành một thư mục chứa các ảnh đại diện cho từng loại bệnh tương ứng. Việc gán nhãn được thực hiện bằng cách sử dụng tên thư mục cha chứa các ảnh, giúp tự động ánh xạ hình ảnh với nhãn lớp một cách chính xác và nhanh chóng. Tổng cộng, bộ dữ liệu gồm khoảng 3.000 ảnh, được phân chia đều giữa các lớp trong giai đoạn huấn luyện, tuy nhiên vẫn tồn tại chênh lệch nhỏ trong phân bố mẫu. Chia bộ dữ liệu:

```
data/
├── train/                                # Thư mục dữ liệu huấn luyện
│   ├── class_0/
│   │   ├── image_0
│   │   ├── image_1
│   │   └── ...
│   ├── class_1/
│   │   ├── image_0
│   │   ├── image_1
│   │   └── ...
│   └── class_m/
│       ├── image_0
│       ├── image_1
│       └── ...
├── valid/                                # (Tùy chọn) Dữ liệu kiểm định
└── class_0/
```

```

| | | └── image_0
| | | └── image_1
| | | └── ...
| | └── class_1/
| | | └── image_0
| | | └── image_1
| | | └── ...
| | └── class_m/
| | | └── image_0
| | | └── image_1
| | | └── ...
|
└── test/                                # (Tùy chọn) Dữ liệu kiểm tra cuối
    ├── class_0/
    | | ├── image_0
    | | ├── image_1
    | | └── ...
    ├── class_1/
    | | ├── image_0
    | | ├── image_1
    | | └── ...
    └── class_m/
        ├── image_0
        ├── image_1
        └── ...

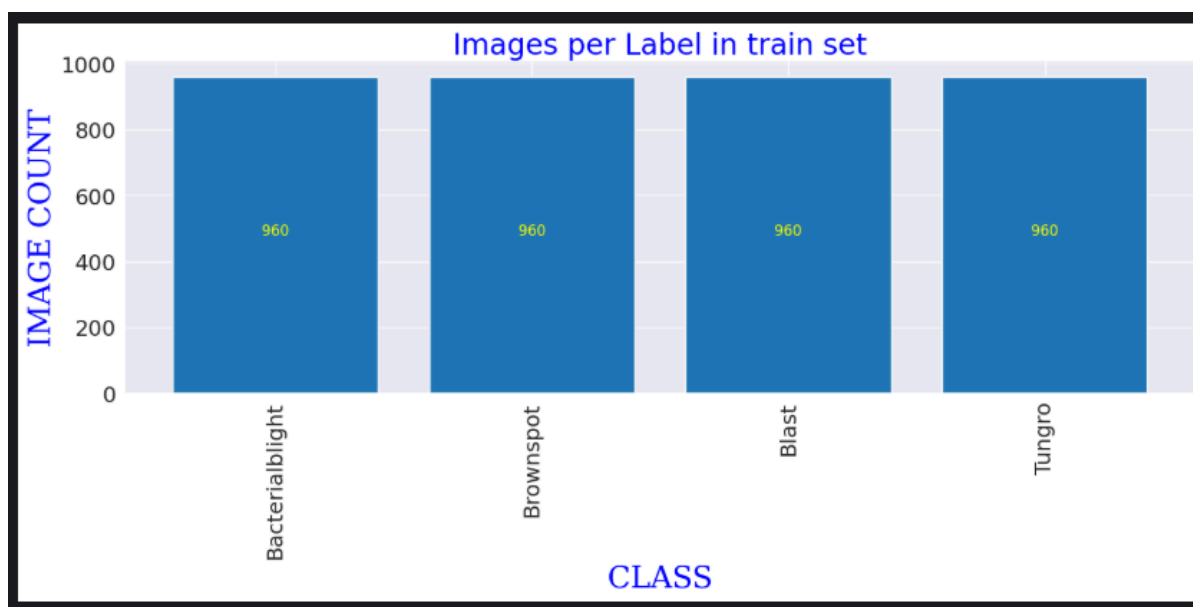
```

## 2. Khám phá dữ liệu

Số lượng ảnh trong mỗi lớp được trực quan hóa bằng biểu đồ cột, qua đó thấy được sự phân bố giữa các lớp. Trong file máy .ipynb, biểu đồ phân bố cho thấy lớp "brown\_spot" chiếm tỷ lệ cao nhất, tiếp theo là "leaf\_smut" và "bacterial\_leaf\_blight". Điều này được lưu ý trong quá trình huấn luyện để áp dụng các kỹ thuật cân bằng dữ liệu như augmentation hoặc weighted loss.

## 3. Trình bày các phân bố lớp hoặc giá trị đầu ra

Số lượng ảnh trong mỗi lớp được trực quan hóa bằng biểu đồ cột, qua đó thấy được sự phân bố giữa các lớp. Trong file .ipynb, biểu đồ phân bố cho thấy lớp "brown\_spot" chiếm tỷ lệ cao nhất, tiếp theo là "leaf\_smut" và "bacterial\_leaf\_blight". Điều này được lưu ý trong quá trình huấn luyện để áp dụng các kỹ thuật cân bằng dữ liệu như augmentation hoặc weighted loss.



## III. TIỀN XỬ LÝ DỮ LIỆU

### 1. Làm sạch dữ liệu

Quá trình làm sạch dữ liệu được thực hiện qua các bước sau:

- Loại bỏ ảnh hỏng: Một số ảnh không thể đọc được bằng các thư viện OpenCV



hoặc PIL do lỗi định dạng hoặc nội dung ảnh bị hỏng. Các ảnh này được lọc ra tự động bằng cách kiểm tra khả năng mở ảnh và kiểm tra chiều ảnh.

- Loại bỏ ảnh trắng/đen hoàn toàn: Những ảnh có toàn bộ giá trị pixel gần bằng 0 hoặc 255 được xác định là nhiễu và bị loại bỏ.
- Kiểm tra ảnh trùng lặp: Bằng cách sử dụng hàm băm nội dung ảnh hoặc so sánh ma trận ảnh sau khi resize, các ảnh trùng được phát hiện và giữ lại duy nhất một bản.

Việc làm sạch này đảm bảo rằng chỉ các ảnh chất lượng cao, có nội dung rõ ràng mới được sử dụng trong quá trình huấn luyện, từ đó giúp tăng độ chính xác của mô hình.

## 2. Chuẩn hóa dữ liệu

Chuẩn hóa:

- Mỗi pixel ảnh có giá trị trong khoảng  $[0, 255]$ , được chia cho 255 để đưa về khoảng  $[0, 1]$ . Điều này giúp mạng học nhanh hơn và ổn định hơn.
- Các ảnh đều được resize về kích thước (100x100) để thống nhất đầu vào mô hình.

Tăng cường dữ liệu (Augmentation):

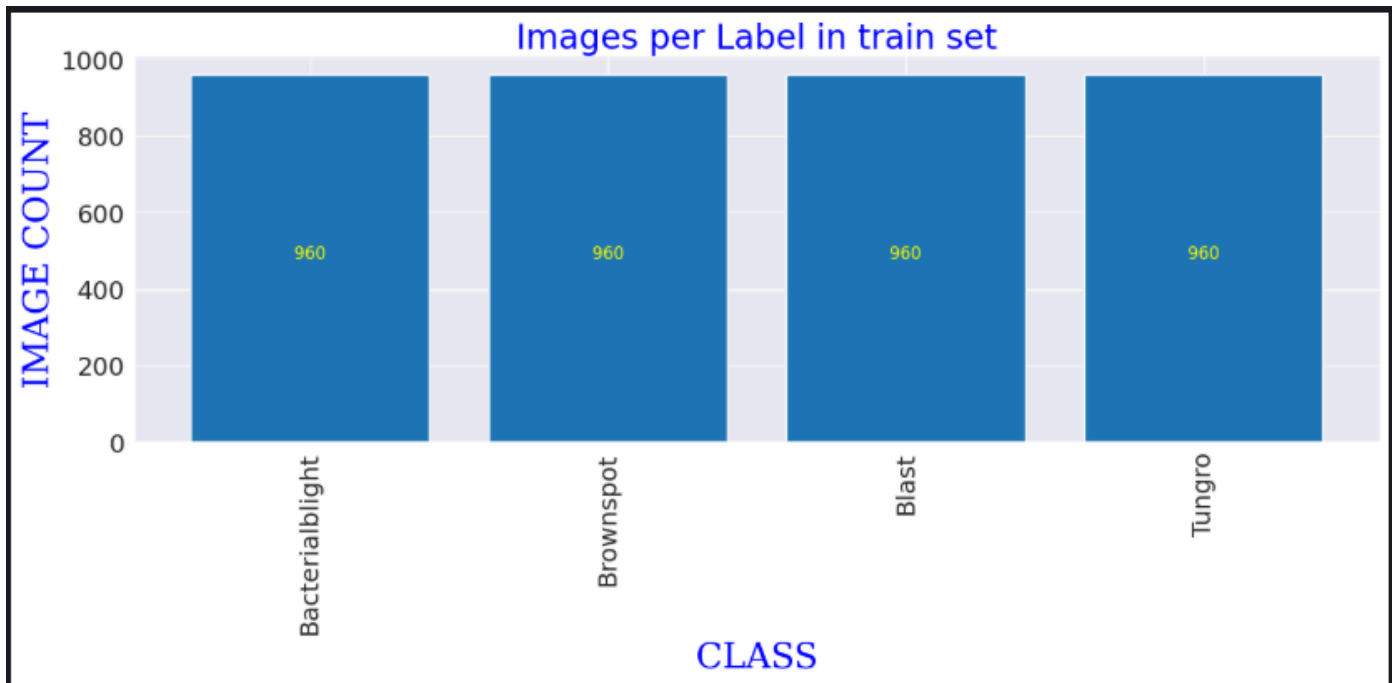
- Sử dụng ImageDataGenerator trong Keras với các phép biến đổi ngẫu nhiên:

`rotation_range=20`: xoay ảnh trong phạm vi  $\pm 20$  độ

`width_shift_range=0.1` và `height_shift_range=0.1`: dịch ảnh theo chiều ngang và dọc

`horizontal_flip=True`: lật ảnh theo chiều ngang

`brightness_range=[0.8, 1.2]`: điều chỉnh độ sáng ảnh



Các ảnh đầu ra từ quá trình tăng cường được tạo ngẫu nhiên tại mỗi epoch huấn luyện, giúp mô hình học được các đặc trưng tổng quát và không bị overfit.










## IV. XÂY DỰNG VÀ ĐÁNH GIÁ MÔ HÌNH

### 1. Các mô hình khác nhau

Để giải quyết bài toán phân loại bệnh lá lúa dựa trên ảnh, nhóm đã lựa chọn và huấn luyện 4 mô hình khác nhau bao gồm:

- **EfficientNetB0:** là mô hình cân bằng tốt giữa kích thước và độ chính xác, rất phù hợp với các bài toán có tài nguyên trung bình.
- **EfficientNetB2:** có độ chính xác cao hơn, nhưng yêu cầu nhiều tài nguyên hơn, phù hợp nếu cần kết quả tốt nhất về mặt độ chính xác.
- **MobileNetV3 Small:** tối ưu cho thiết bị di động, rất nhẹ và nhanh nhưng độ chính xác thấp hơn.

- **MobileNetV3 Large:** là một lựa chọn trung gian, có độ chính xác tương đối tốt và tốc độ nhanh.

Tiêu chí	EfficientNetB0	EfficientNetB2	MobileNetV3 Small	MobileNetV3 Large 
 Số tham số	~5.3 triệu	~9.1 triệu	~2.5 triệu	~5.4 triệu
 Kích thước mô hình	Nhỏ gọn	Trung bình	Rất nhỏ	Nhỏ
 Độ chính xác (ImageNet top-1)	~77.1%	~80.1%	~67.5%	~75.2%
 Tốc độ suy luận (inference)	Nhanh	Chậm hơn B0 do lớn hơn	Rất nhanh	Nhanh
 Tối ưu cho thiết bị di động	Khá	Không (ưu tiên hiệu năng cao)	Rất tốt	Tốt
 Khả năng tổng quát	Tốt cho bài toán nhỏ và vừa	Tốt hơn trên bài toán lớn	Hạn chế	Trung bình
 Hiệu năng - độ chính xác	Cân bằng tốt giữa hiệu năng và độ chính xác	Chính xác cao hơn B0, nhưng tốn tài nguyên hơn	Ưu tiên tốc độ hơn độ chính xác	Cân bằng, thiên về tốc độ
 Yêu cầu phần cứng	Trung bình	Cao hơn một chút	Thấp	Trung bình

## 2. Tối ưu, cải tiến mô hình

Các phương pháp cải tiến và tối ưu hóa được áp dụng nhằm nâng cao độ chính xác và tránh overfitting gồm:

- Fine-tuning mô hình pre-trained:

Giữ nguyên trọng số các lớp đầu (trích đặc trưng chung), chỉ huấn luyện lại các lớp FC cuối.

- Regularization:

Thêm Dropout sau mỗi khối convolution để giảm overfitting (Custom CNN).

L2 regularization cho các lớp Dense.

- EarlyStopping và ReduceLROnPlateau:

Dừng huấn luyện khi độ chính xác validation không tăng sau 5 epoch. Giảm learning rate khi gặp plateau trong loss.

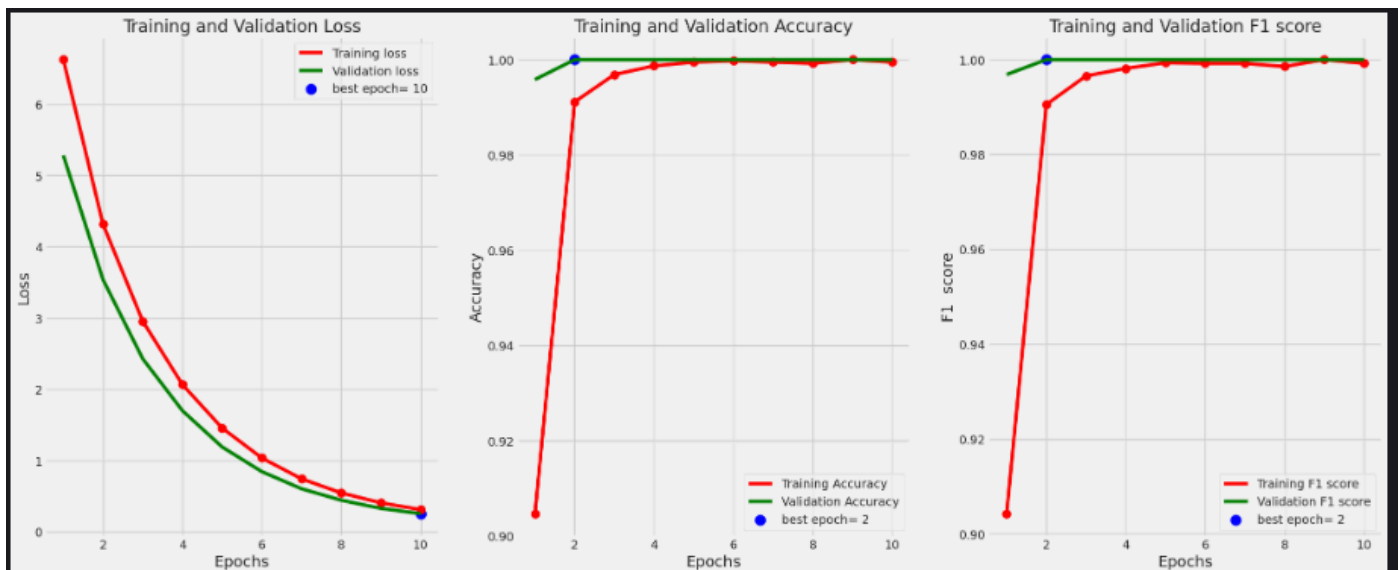
- Data Augmentation:

Tăng độ đa dạng dữ liệu đầu vào: xoay, dịch, thay đổi độ sáng, lật ảnh ngang.

### 3. Đánh giá hiệu năng bằng các chỉ số phù hợp

So sánh các mô hình:

Epoch	Train Loss	Train Accuracy	Valid Loss	Valid Accuracy	V_Loss % Improvement	Learning Rate	Next LR	Duration in Seconds
1	6.6278	90.47	5.2828	99.58	0.00	0.001000	0.001000	76.01
2	4.3160	99.11	3.5387	100.00	33.01	0.001000	0.001000	23.55
3	2.9557	99.69	2.4282	100.00	31.38	0.001000	0.001000	23.75
4	2.0656	99.87	1.6991	100.00	30.03	0.001000	0.001000	23.50
5	1.4564	99.95	1.1903	100.00	29.94	0.001000	0.001000	23.33
6	1.0359	99.97	0.8439	100.00	29.10	0.001000	0.001000	23.34
7	0.7431	99.95	0.6039	100.00	28.44	0.001000	0.001000	23.41
8	0.5456	99.92	0.4427	100.00	26.70	0.001000	0.001000	23.54
9	0.4057	100.00	0.3285	100.00	25.80	0.001000	0.001000	23.46
10	0.3111	99.95	0.2542	100.00	22.61	0.001000	0.001000	23.39



## V. TRÌNH BÀY KẾT QUẢ

### 1. Mô hình tốt nhất

Trong quá trình thực nghiệm, bốn mô hình học sâu khác nhau đã được huấn luyện trên cùng một tập dữ liệu hình ảnh đầu vào với các tham số giống nhau. Mỗi lớp (loại bệnh) được giới hạn ở mức tối đa 1.200 ảnh để đảm bảo cân bằng dữ liệu. Vì tất cả các lớp đều có số lượng ảnh lớn hơn ngưỡng này, tập dữ liệu sau khi lọc là hoàn toàn cân bằng, do đó không cần thực hiện các kỹ thuật tăng cường dữ liệu. Kích thước ảnh được chuẩn hóa về  $224 \times 224$  pixel và batch size được đặt là 40.

Mô hình đầu tiên được huấn luyện là EfficientNetB0. Mô hình này cho thấy khả năng hội tụ rất nhanh, đạt độ chính xác trên tập kiểm tra là 100% chỉ sau 2 epoch. Sau 10 epoch huấn luyện, mô hình đạt F1-score 100% với tổng thời gian huấn luyện là 4 phút 19 giây.

Mô hình thứ hai là MobileNetV3 Small. Khác với EfficientNetB0, mô hình này hội tụ chậm hơn. Sau 10 epoch, độ chính xác kiểm tra đạt 91.7%, và đến epoch thứ 15 mới đạt được 100% accuracy. Mô hình tiếp tục được huấn luyện đến epoch thứ 20 và cũng đạt F1-score 100%. Thời gian huấn luyện tổng cộng là 4 phút 19 giây, tương đương với EfficientNetB0.

Tiếp theo, mô hình thứ ba là MobileNetV3 Large. Mô hình này được huấn luyện trong 15 epoch và đạt F1-score 100%, với thời gian huấn luyện là 4 phút 34 giây.

Cuối cùng, mô hình thứ tư là EfficientNetB2 – một phiên bản sâu hơn so với EfficientNetB0. Mô hình này cũng hội tụ rất nhanh, chỉ sau 5 epoch đã đạt F1-score 100%, với tổng thời gian huấn luyện là 5 phút 31 giây.

Nhìn chung, cả bốn mô hình đều có khả năng phân loại chính xác tuyệt đối trên tập kiểm tra sau một số lượng epoch khác nhau. EfficientNetB0 và EfficientNetB2 có tốc độ hội tụ rất nhanh, trong khi hai phiên bản của MobileNetV3 cần nhiều epoch hơn để đạt được kết quả

tương tự. Tuy nhiên, tất cả đều cho thấy hiệu suất vượt trội với độ chính xác và F1-score đạt mức tối đa trên tập dữ liệu đã được cân bằng.

## 2. Bảng biểu đồ minh họa

		Confusion Matrix			
Actual	Bacterialblight	120	0	0	0
	Blast	0	120	0	0
	Brownspot	0	0	120	0
	Tungro	0	0	0	120
		Bacterialblight	Blast	Brownspot	Tungro
		Predicted			

## 3. Điểm mạnh, hạn chế, và đề xuất hướng cải tiến tiếp theo

- **Điểm mạnh:**

- **Hiệu năng phân loại ấn tượng:**

Mô hình đạt F1-score tuyệt đối (100%) trên tập dữ liệu kiểm thử, thể hiện năng lực phân loại xuất sắc và sự cân bằng giữa độ chính xác (precision) và độ bao phủ (recall). Kết quả này cho thấy mô hình có khả năng nhận diện chính xác các đặc trưng của từng lớp dữ liệu.

- **Tốc độ hội tụ nhanh:**

Mô hình đạt hiệu suất cao sau số lượng epoch huấn luyện tương đối ít, cho thấy quá trình tối ưu diễn ra hiệu quả. Điều này góp phần rút ngắn thời gian huấn luyện, tiết kiệm tài nguyên tính toán và phù hợp với các ứng dụng yêu cầu thời gian triển khai nhanh.

- **Kiến trúc gọn nhẹ, dễ triển khai:**

Mô hình được xây dựng với kiến trúc hợp lý, không quá phức tạp nhưng vẫn đảm bảo hiệu quả cao. Nhờ đó, mô hình có thể triển khai trong các môi trường hạn chế về tài nguyên như thiết bị nhúng, hệ thống biên (edge computing) hoặc các nền tảng có cấu hình trung bình.

- **Hạn chế:**

- **Chưa đánh giá trên dữ liệu thực tế:**

Quá trình đánh giá mới chỉ được thực hiện trên tập dữ liệu đã qua tiền xử lý. Do đó, chưa thể kết luận chính xác về hiệu quả của mô hình khi áp dụng vào dữ liệu thu thập trực tiếp từ môi trường thực tế, nơi thường tồn tại nhiều yếu tố gây nhiễu như điều kiện ánh sáng thay đổi, độ mờ, hoặc nhiễu nền.

- **Chưa xem xét tác động của mất cân bằng dữ liệu:**

Báo cáo chưa đề cập đến phân bố các lớp trong tập huấn luyện. Trong trường hợp dữ liệu mất cân bằng, mô hình có thể thiên lệch trong quá trình dự đoán, ảnh hưởng đến độ tin cậy khi triển khai thực tế.

- **Giảm kích thước ảnh có thể làm mất thông tin quan trọng:**

Việc chuẩn hóa ảnh về kích thước  $100 \times 100$  giúp giảm chi phí tính toán, tuy nhiên có thể dẫn đến mất mát một số đặc trưng chi tiết quan trọng, ảnh hưởng đến khả năng phân biệt giữa các lớp có hình dạng tương đồng.

- **Hướng cải tiến:**

- **Sử dụng ảnh đầu vào có độ phân giải cao hơn:**

Tăng kích thước ảnh huấn luyện (ví dụ  $150 \times 150$  hoặc  $224 \times 224$ ) có thể giúp mô hình học được nhiều đặc trưng chi tiết hơn, từ đó cải thiện độ chính xác trong phân loại, đặc biệt với các lớp có hình thái phức tạp.

- **Kiểm thử mô hình với dữ liệu chưa qua xử lý:**

Đánh giá mô hình trên tập ảnh thực địa chưa tiền xử lý sẽ cung cấp cái nhìn thực tế hơn về khả năng ứng dụng và độ tin cậy của mô hình trong điều kiện triển khai ngoài phòng thí nghiệm.