ỨNG DỤNG HỌC SÂU TRONG TỰ ĐỘNG HÓA KIỂM ĐỊNH CHẤT LƯỢNG HẠT CÀ PHÊ RANG

Ngô Minh Phát [1], Dương Thị Mộng Thùy [2*], Lê Cung Tưởng [1]

[1]Khoa Công nghệ thông tin, Trường Đại học Công Nghệ TP.HCM
[2] Khoa Công nghệ thông tin, Trường Đại học Công Thương TP.HCM
minhphat2811@gmail.com, thuydtm@huit.edu.vn, lc.tuong@hutech.edu.vn

TÓM TẮT— Nghiên cứu này đề xuất kiến trúc VnCoffeeFr, một mô hình học sâu được phát triển để tự động hóa quá trình phân loại chất lượng hạt cà phê rang dựa trên phân tích đặc trưng hình ảnh. Chúng tôi xây dựng một bộ dữ liệu VnCoffeeSet gồm 4.194 hình ảnh hạt cà phê rang và thực hiện quy trình chuẩn hóa nhằm đảm bảo sự nhất quán về điều kiện ánh sáng, góc chụp và độ phân giải của dữ liệu. Từ bộ dữ liệu đã xây dựng, các kiến trúc CNN phổ biến như ResNet-50, EfficientNet-B3, VGG16 và Inception-V3 được tích hợp vào kiến trúc VnCoffeeFr để huấn luyện và đánh giá hiệu quả phân loại. Kết quả thực nghiệm cho thấy các mô hình đạt độ chính xác cao, trong đó EfficientNet-B3 đạt 99.52%, qua đó chứng minh tính hiệu quả và khả thi của VnCoffeeFr trong việc xây dựng hệ thống kiểm định chất lượng cà phê tự động.

Từ khóa— Deep Learning, Computer Vision, Convolutional Neural Networks, Industrial Automation.

I. GIỚI THIỆU

Trong những năm gần đây, sự phát triển mạnh mẽ của trí tuệ nhân tạo (Artificial Intelligence), đặc biệt các mô hình học sâu (Deep Learning), đã tạo ra những bước tiến vượt bậc trong lĩnh vực thị giác máy tính (Computer Vision). Nhờ khả năng học và trích xuất đặc trưng tự động từ dữ liệu lớn, mô hình học sâu dựa trên mạng nơ-ron tích chập (Convolutional Neural Networks – CNN) đã đạt được hiệu năng vượt trội trong nhiều bài toán nhận dạng và phân loại hình ảnh [1-3]. Sự phát triển của các kiến trúc mạng hiện đại như ResNet, VGG, EfficientNet, Inception đã góp phần quan trọng trong việc nâng cao độ chính xác và khả năng tổng quát hóa của các hệ thống thị giác máy tính, từ đó thúc đẩy sự ra đời của nhiều ứng dụng thực tiễn trong công nghiệp, y tế, giao thông và nông nghiệp. Trong lĩnh vực nông nghiệp và chế biến nông sản, việc ứng dụng các công nghệ trí tuệ nhân tạo và thị giác máy tính vào quá trình tự động hóa sản xuất quy mô lớn đang ngày càng mang lại hiệu quả kinh tế cao. Các hệ thống nhận dạng và phân tích hình ảnh dựa trên học sâu cho phép tự động thực hiện nhiều nhiệm vụ thay thế con người như phát hiện sâu bệnh, phân loại nông sản, kiểm tra chất lượng sản phẩm và giám sát quy trình sản xuất. Việc áp dụng các mô hình học sâu không chỉ giúp nâng cao độ chính xác trong quá trình đánh giá chất lượng mà còn góp phần giảm chi phí nhân công, tăng tốc độ xử lý và đảm bảo tính nhất quán trong các dây chuyền sản xuất hiện đại.

Đối với ngành công nghiệp chế biến cà phê, việc kiểm định chất lượng hạt cà phê rang là một công đoạn quan trọng nhằm đảm bảo tính đồng nhất của sản phẩm trước khi đưa ra thị trường. Chất lượng hạt cà phê rang ảnh hưởng trực tiếp đến hương vị, màu sắc và giá trị thương mại của sản phẩm. Tuy nhiên, trong nhiều cơ sở sản xuất, quá trình kiểm định chất lượng vẫn được thực hiện thủ công dựa trên quan sát cảm quan của chuyên gia, do đó phụ thuộc nhiều vào kinh nghiệm cá nhân, điều kiện chiếu sáng và dễ phát sinh sai lệch trong đánh giá. Bên cạnh đó, với số lượng hạt cà phê lớn trong mỗi mẻ rang, phương pháp này đòi hỏi chi phí nhân công cao và khó đáp ứng yêu cầu kiểm tra nhanh, liên tục trong các dây chuyền sản xuất quy mô lớn. Trong bối cảnh nhu cầu tiêu thụ cà phê chế biến và cà phê đặc sản ngày càng gia tăng, thị trường cũng đặt ra các yêu cầu khắt khe hơn về chất lượng sản phẩm tốt, thời gian kiểm định nhanh [4]. Điều này đòi hỏi các doanh nghiệp không chỉ nâng cao tiêu chuẩn kiểm định mà còn phải áp dụng các giải pháp tự động hóa và công nghệ hiện đại nhằm đảm bảo độ chính xác, nhất quán và hiệu quả nhanh trong quá trình đánh giá chất lượng.

Trong bối cảnh đó, việc nghiên cứu và phát triển các hệ thống kiểm định chất lượng tự động dựa trên công nghệ thị giác máy tính và học sâu được xem là một hướng tiếp cận tiềm năng nhằm nâng cao hiệu quả sản xuất. Trên thực tế, mục tiêu cốt lõi của hệ thống kiểm định sau rang là hỗ trợ sàng lọc nhanh chất lượng hạt trong quy trình sản xuất, trong đó yêu cầu quan trọng nhất là phân biệt kịp thời giữa sản phẩm đạt và không đạt tiêu chuẩn. Một kịch bản điển hình đánh giá chất lượng của một mẻ rang là trải đều hạt cà phê, thu nhận ảnh và tự động xác định số lượng hạt đạt hay không đạt để ước lượng tỷ lệ đạt chuẩn của toàn bộ mẻ. Trong ngữ cảnh này, thông tin quan trọng nhất là tỷ lệ đạt chuẩn, trong khi việc phân biệt chi tiết từng loại lỗi cụ thể không phải là yêu cầu bắt buộc. Do đó, bài toán phân loại nhị phân được lựa chọn nhằm phù hợp với mục tiêu ứng dụng thực tiễn. Đồng thời, việc gộp các dạng lỗi vào một lớp "không đạt" giúp đơn giản hóa bài toán, giảm độ phức tạp trong quá trình huấn luyện và góp phần nâng cao tính ổn định cũng như khả năng tổng quát hóa của mô hình. Tuy nhiên, bài toán phân loại đa lớp vẫn có ý nghĩa trong các kịch bản chuyên sâu hơn để phân tích nguyên nhân lỗi, tối ưu hóa quy trình rang.

* Corresponding Author

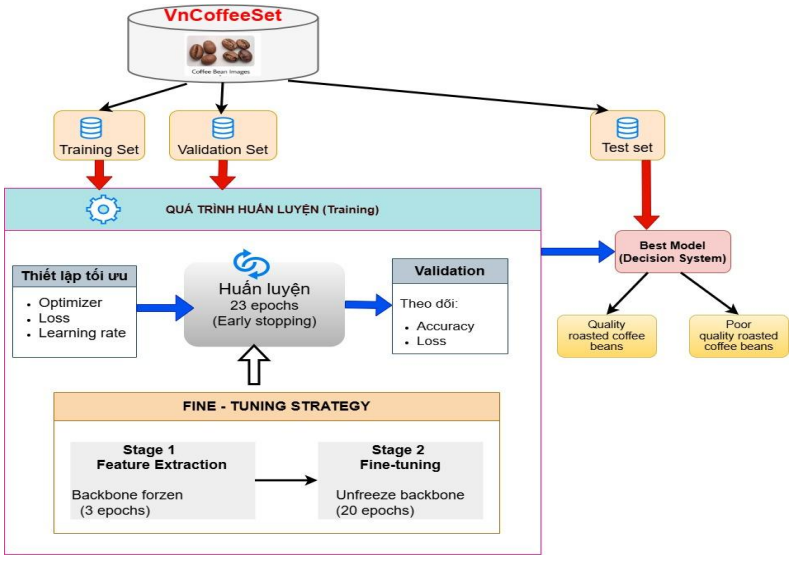
Hiện nay, đã có một số nghiên cứu áp dụng mạng nơ-ron tích chập (CNN) cho các bài toán phân loại hoặc phát hiện lỗi hạt cà phê. Tiêu biểu, bài báo "Robusta Coffee Bean Defect Classification using Convolutional Neural Network" [5] đề xuất một mô hình CNN tùy chỉnh (custom Convolutional Neural Network) nhằm phân loại các lỗi phổ biến của hạt cà phê Robusta từ dữ liệu ảnh. Phương pháp này cho phép mô hình học trực tiếp các đặc trưng quan trọng như màu sắc, kết cấu và hình dạng từ tập dữ liệu ảnh và đạt độ chính xác đến 92%. Bên cạnh đó, bài báo "Coffee Bean Defects Automatic Classification Realtime Application Adopting Deep Learning" [6] tập trung phát triển một hệ thống phân loại lỗi hạt cà phê theo thời gian thực dựa trên các mô hình học sâu. Nghiên cứu không chỉ xây dựng mô hình CNN để nhận diện các loại lỗi đạt độ chính xác khoảng 95% mà còn tích hợp vào một ứng dụng thực tế, cho phép xử lý và dự đoán trực tiếp từ camera. Tuy nhiên, các công trình này thường bị giới hạn bởi quy mô dữ liệu nhỏ, điều kiện thu thập chưa được chuẩn hóa và chưa áp dụng các kiến trúc mô hình huấn luyện sẵn hiện đại (Pre-trained CNN) vào quy trình huấn luyện và đánh giá nên bị giới hạn về hiệu suất và không gian lưu trữ của mô hình. Những hạn chế này đặt ra nhu cầu xây dựng một bộ dữ liệu chất lượng cao, đồng thời phát triển các phương pháp đánh giá toàn diện hơn nhằm nâng cao hiệu quả của bài toán phân loại chất lượng hạt cà phê sau rang.

Xuất phát từ những vấn đề trên, bài báo này nghiên cứu hiệu quả của các phương pháp học sâu nhằm tự động hóa quá trình kiểm định chất lượng hạt cà phê rang từ hình ảnh. Đóng góp chính của nghiên cứu bao gồm: (1) xây dựng bộ dữ liệu ảnh VnCoffeeSet, được thu thập trong điều kiện thực tế và chuẩn hóa về điều kiện ánh sáng, góc chụp và độ phân giải nhằm phục vụ cho bài toán phân loại chất lượng; (2) áp dụng và so sánh hiệu năng của một số kiến trúc mạng nơ-ron tích chập phổ biến gồm ResNet-50, EfficientNet-B3, VGG16, và Inception-V3 trong nhiệm vụ phân loại chất lượng hạt cà phê rang từ hình ảnh; (3) đánh giá khả năng ứng dụng của các mô hình học sâu trong việc xây dựng hệ thống kiểm định chất lượng cà phê tự động cho các môi trường sản xuất quy mô lớn. Kết quả của nghiên cứu góp phần chứng minh hiệu quả của các mô hình học sâu trong bài toán phân loại chất lượng hạt cà phê rang và mở ra tiềm năng ứng dụng rộng rãi của các hệ thống thị giác máy tính dựa trên trí tuệ nhân tạo trong lĩnh vực sản xuất và chế biến nông sản. Kết quả thực nghiệm cho thấy các mô hình học sâu có khả năng nhận dạng và phân loại chất lượng hạt cà phê rang với độ chính xác cao, trong đó mô hình EfficientNet-B3 đạt hiệu năng tốt nhất, cho thấy tiềm năng ứng dụng của phương pháp đề xuất trong việc phát triển các hệ thống kiểm định chất lượng tự động trong ngành công nghiệp chế biến cà phê.

Phần còn lại của bài báo được tổ chức như sau: Phần II trình bày dữ liệu và phương pháp nghiên cứu được sử dụng trong bài toán; Phần III trình bày kết quả thực nghiệm và thảo luận; cuối cùng, Phần IV đưa ra kết luận và định hướng nghiên cứu trong tương lai.

II. DỮ LIỆU VÀ PHƯƠNG PHÁP

Phần này trình bày kiến trúc tổng quan của hệ thống phân loại hạt cà phê – VnCoffeeFr như hình 1 bên dưới.



Hình 1. *Kiến trúc của hệ thống phân loại hạt cà phê rang - VnCoffeeFr.*

Kiến trúc được thiết kế dựa trên ứng dụng học sâu nhằm tự động hóa quá trình nhận dạng và kiểm định chất lượng hạt cà phê rang trên bộ dữ liệu ảnh VnCoffeeSet. Quy trình tổng thể được đề xuất bao gồm các bước chính như thu thập dữ liệu, tiền xử lý dữ liệu, tăng cường dữ liệu, huấn luyện mô hình và đánh giá hiệu năng mô hình.

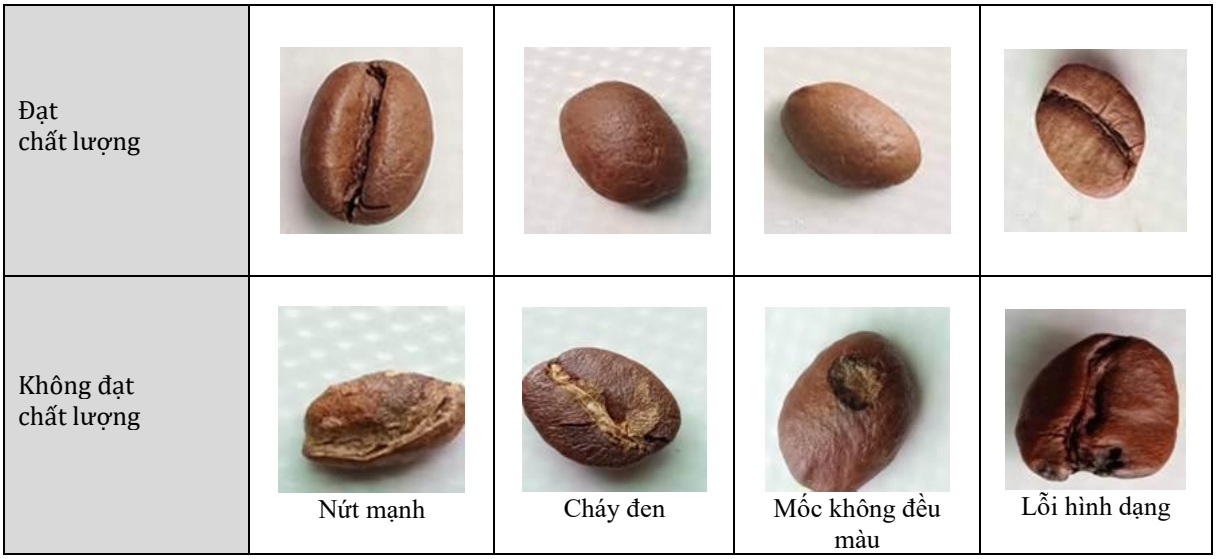
A. BỘ DỮ LIỆU CÀ PHÊ VNCOFFEESET

1. THU THẬP, CHUẨN HÓA VÀ GÁN NHÃN

Trong nghiên cứu này, hai mẫu cà phê Arabica được khảo sát có nguồn gốc từ Cầu Đất và Lạc Dương thuộc tỉnh Lâm Đồng. Ảnh được chụp bằng điện thoại Samsung Galaxy M34 5G và Xiaomi Poco M5 6G trên giá đỡ cố định, dưới hai chế độ ánh sáng là buổng sáng 5500K và ánh sáng tự nhiên khuếch tán. Tất cả ảnh được đưa về kích thước 224×224 và được chuẩn hóa màu. Việc gán nhãn thực hiện bởi hai chuyên gia phân loại, kiểm định cà phê. Hình ảnh hạt cà phê rang được chia thành hai lớp dựa vào các đặc trưng chuyên biệt trên hình ảnh thu thập được minh họa trên Bảng 1. Hình đại diện về các lớp trong bộ dữ liệu phân loại hạt cà phê được thể hiện ở Hình 2.

Bảng 1. *Mô tả đặc trưng của bộ dữ liệu VnCoffeeSet*

Mã hóa lớp	Tên phân loại	Đặc trưng
0	Đạt chất lượng (Quality roasted coffee beans)	hạt đồng đều màu không cháy không nứt mạnh
1	Không đạt chất lượng (Poor quality roasted coffee beans)	cháy đen nứt mạnh lỗi hình dạng mốc hoặc không đồng đều màu



Hình 2. *Hình ảnh về các lớp trong bộ dữ liệu phân loại hạt cà phê.*

2. PHÂN BỐ VÀ TIỀN XỬ LÝ DỮ LIỆU HUẤN LUYỆN

Tập dữ liệu VnCoffeeSet được thu thập rất cẩn thận. Mỗi ảnh trong bộ dữ liệu tương ứng với một hạt cà phê độc lập, và không tồn tại nhiều ảnh của cùng một hạt được phân bố vào các tập khác nhau bao gồm 4.194 ảnh hạt cà phê rang, được phân thành hai lớp dựa trên kết quả đánh giá của chuyên gia kiểm định chất lượng. Trong đó, 1.667 ảnh thuộc lớp 0 (đạt chất lượng) với các hạt có màu rang đồng đều, không cháy và không nứt mạnh; và 2.527 ảnh thuộc lớp 1 (không đạt chất lượng) bao gồm các trường hợp cháy đen, nứt mạnh, biến dạng, hoặc màu sắc không đồng đều. Lớp đạt chất lượng chiếm khoảng 39,8% tổng số dữ liệu, trong khi lớp không đạt chất lượng chiếm khoảng 60,2%, cho thấy bộ dữ liệu có sự mất cân bằng ở mức nhẹ giữa hai lớp nhưng vẫn phản ánh thực tế trong quá trình phân loại chất lượng cà phê rang vì các hạt cà phê đạt chất lượng khá đồng đều về hình dạng, kết cấu nên số lượng thu thập ít hơn.

Trong quá trình thực nghiệm của nghiên cứu này, bộ dữ liệu được phân chia ngẫu nhiên thành ba tập con bao gồm tập huấn luyện (training set – 70%, tương ứng 2.934 ảnh), tập xác thực (validation set – 15%, tương ứng 629 ảnh) và tập kiểm tra (test set – 15%, tương ứng 631 ảnh), đồng thời đảm bảo không có sự chồng lặp giữa các tập này. Điều này giúp đảm bảo rằng kết quả đánh giá phản ánh đúng khả năng tổng quát hóa của mô hình, không bị ảnh hưởng bởi các mẫu trùng lặp hoặc tương tự. Training set được sử dụng để huấn luyện mô hình học các đặc trưng của hạt cà phê, trong khi validation set được sử dụng để quan sát và tối ưu hóa các tham số của mô hình đề xuất. Cuối cùng, test set được sử dụng để đánh giá hiệu suất của các mô hình thử nghiệm. Mức độ phân bố ảnh giữa các lớp và tập dữ liệu huấn luyện được mô tả ở bảng 2. Việc phân chia dữ liệu theo tỷ lệ này giúp đảm bảo tính khách quan trong quá trình đánh giá mô hình, đồng thời duy trì phân bố dữ liệu tương đối ổn định giữa các tập con, góp phần hạn chế sai lệch và nâng cao độ tin cậy của kết quả thực nghiệm.

Bảng 2. *Bảng ma trận phân bố dữ liệu*

Tập dữ liệu	Lớp 0 (Đạt chất lượng)	Lớp 1 (Không đạt chất lượng)	Tổng số ảnh
Training (70%)	1166	1768	2.934
Validation (15%)	250	379	629
Test (15%)	251	380	631
Tổng	1.667	2.527	4.194

Để đảm bảo tính nhất quán của ảnh đầu vào, chúng tôi xử lý ảnh thông qua các bước tiền xử lý chính:

- Chuẩn hóa kích thước: Tất cả ảnh được cắt theo khung hình vuông kích thước 224×224 pixel và hạt cà phê nằm ở trung tâm ảnh.
- Chuẩn hóa màu: Tất cả ảnh được chuẩn hóa sang định dạng RGB để biểu diễn màu sắc đồng nhất.
- Chuẩn hóa pixel: Chuẩn hóa giá trị pixel của tất cả ảnh đầu vào để cải thiện tính ổn định của quá trình huấn luyện và tăng cường sự hội tụ của mô hình.
- Tăng cường dữ liệu: Kỹ thuật được sử dụng để tạo ra các biến thể mới của dữ liệu huấn luyện từ các ảnh gốc thông qua các phép biến đổi như lật, xoay ±10∘, thay đổi độ sáng và độ tương phản ±10%, cắt hoặc phóng to ảnh [0.9-1.0], vung xóa tối đa khoảng [2%-10%]. Mục đích của phương pháp này là tăng kích thước và sự đa dạng của tập dữ liệu, giúp mô hình học sâu giảm hiện tượng quá khớp và cải thiện khả năng tổng quát hóa khi dự đoán trên dữ liệu mới.

B. MÔ HÌNH HUẤN LUYỆN SẴN (PRE-TRAINED MODEL)

Sự phát triển của các mạng nơ-ron tích chập (CNN) đã mang lại bước tiến lớn trong các bài toán nhận dạng hình ảnh nhờ khả năng tự học đặc trưng từ dữ liệu. Trong nghiên cứu này, các mô hình CNN huấn luyện sẵn (pre-trained) được sử dụng như một nền tảng, cho phép tận dụng tri thức đã học từ các tập dữ liệu lớn để cải thiện hiệu quả phân loại và rút ngắn thời gian huấn luyện. Phương pháp này đã được ứng dụng rộng rãi trong nhiều lĩnh vực như nhận dạng khuôn mặt, phát hiện đối tượng và phân tích hình ảnh.

1. VGG16

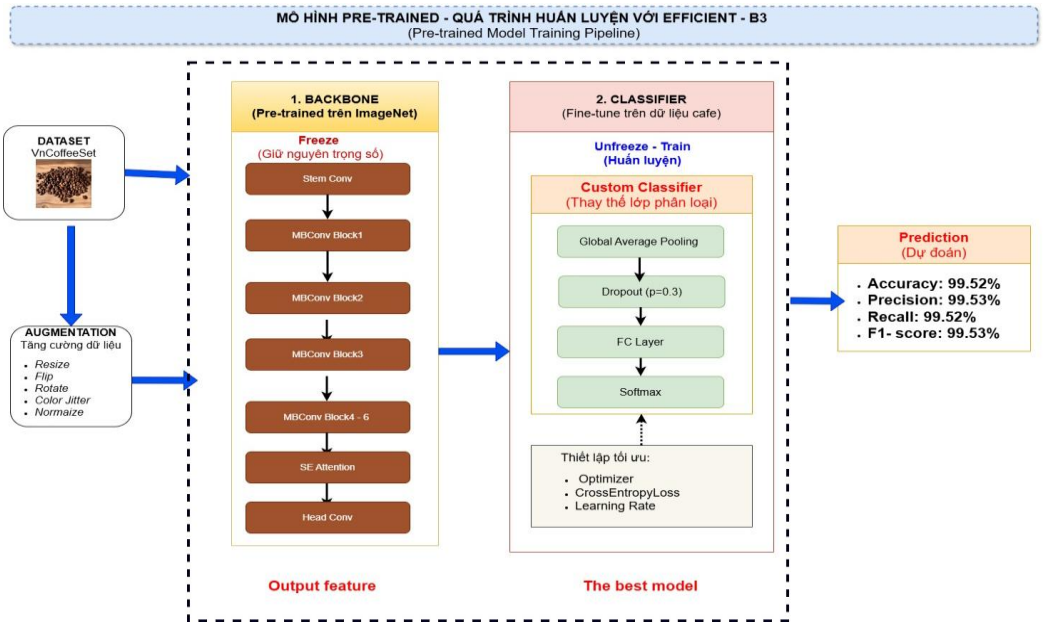
Mô hình VGG được giới thiệu lần đầu trong bài báo Very Deep Convolutional Networks for Large-Scale Image Recognition của Karen Simonyan và Andrew Zisserman thuộc Visual Geometry Group tại University of Oxford [7]. Các tác giả đề xuất kiến trúc mạng tích chập sâu sử dụng các bộ lọc nhỏ 3×3 xếp chồng nhau, cho phép xây dựng các mạng có 16–19 lớp có trọng số giúp tăng khả năng học đặc trưng phi tuyến mà vẫn giữ số tham số ở mức hợp lý. Kiến trúc VGG-16 gồm 16 lớp có tham số học, bao gồm 13 lớp convolution và 3 lớp fully connected, được tổ chức thành 5 khối tích chập; mỗi khối sử dụng các convolution 3×3 với hàm kích hoạt ReLU và được theo sau bởi lớp max-pooling 2×2 để giảm kích thước đặc trưng. Mạng nhận đầu vào kích thước 224×224×3 và lớp phân loại cuối cùng, sử dụng softmax cho 1000 lớp trong bộ dữ liệu ImageNet. Tổng số tham số của VGG16 khoảng 138 triệu, trong đó phần lớn nằm ở các lớp fully connected.

2. RESNET-50

Mô hình ResNet-50 được giới thiệu trong bài báo Deep Residual Learning for Image Recognition của Kaiming He, Xiangyu Zhang, Shaoqing Ren và Jian Sun thuộc Microsoft Research [8]. Mô hình sử dụng cơ chế học phần dư (residual learning) thông qua các kết nối tắt (skip connections), cho phép tín hiệu đầu vào được cộng trực tiếp với đầu ra của các lớp tích chập giúp giải quyết vấn đề suy giảm độ chính xác khi mạng trở nên quá sâu (degradation problem) và hỗ trợ quá trình lan truyền gradient hiệu quả hơn trong quá trình huấn luyện. Kiến trúc ResNet-50 gồm 50 lớp có tham số, được xây dựng từ các residual blocks dạng bottleneck, mỗi block bao gồm ba lớp convolution với kích thước 1×1, 3×3 và 1×1. Các block này được tổ chức thành bốn giai đoạn chính với số lượng block lần lượt 3, 4, 6 và 3, xen kẽ với các lớp downsampling để giảm kích thước đặc trưng. Mạng nhận đầu vào kích thước 224×224×3, sử dụng Batch Normalization và ReLU sau mỗi lớp tích chập, và kết thúc bằng global average pooling và lớp fully connected với hàm softmax cho phân loại. Tổng số tham số của ResNet-50 khoảng 25 triệu, ResNet-50 nhẹ hơn VGG16 nhưng vẫn đạt hiệu năng cao.

3. INCEPTION-V3

Mô hình Inception-V3 được giới thiệu trong bài báo Rethinking the Inception Architecture for Computer Vision của Christian Szegedy và cộng sự tại Google Research [9]. Mô hình Inception-V3 sử dụng phương pháp phân tách các phép tích chập lớn thành nhiều phép tích chập nhỏ hơn đồng thời sử dụng Inception modules để trích xuất đặc trưng ở nhiều kích thước khác nhau trong cùng một tầng mạng. Kiến trúc của mô hình bao gồm nhiều Inception blocks được tổ chức theo các giai đoạn, kết hợp với batch normalization, auxiliary classifiers và global average pooling để cải thiện quá trình huấn luyện. Mạng sử dụng kích thước đầu vào 299×299×3 và có khoảng 23 triệu tham số. Nhờ thiết kế mô-đun hiệu quả và khả năng khai thác đặc trưng đa tỷ lệ, Inception-V3 đạt hiệu suất cao trong các nhiệm vụ phân loại ảnh quy mô lớn.



Hình 3. *Mô hình huấn luyện EfficientNet-B3.*

4. EFFICIENTNET – B3

Mô hình EfficientNet-B3 thuộc họ mạng EfficientNet được đề xuất trong bài báo EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks của Mingxing Tan và Quoc V. Le tại Google Research [10]. Mô hình EfficientNet sử dụng compound scaling, tức là mở rộng mô hình một cách đồng thời và cân bằng theo ba chiều: độ sâu (depth), độ rộng (width) và độ phân giải ảnh đầu vào (resolution) nên khai thác đặc trưng hiệu quả, cân bằng tốt giữa độ chính xác và chi phí tính toán. Kiến trúc EfficientNet được xây dựng dựa trên MBConv (Mobile Inverted Bottleneck Convolution) kết hợp với Squeeze-and-Excitation (SE) blocks để tăng cường khả năng học đặc trưng. EfficientNet-B3 sử dụng kích thước đầu vào 300×300×3 và gồm nhiều khối MBConv với các hệ số mở rộng khác nhau, được tối ưu hóa thông qua quá trình Neural Architecture Search (NAS). Mô hình có khoảng 12 triệu tham số nhưng vẫn đạt hiệu năng rất cao. Ngoài ra, với số lượng tham số tương đối thấp, mô hình giúp giảm nguy cơ overfitting và phù hợp với các tập dữ liệu nhỏ, đồng thời vẫn đạt hiệu năng cao trong các bài toán phân loại ảnh. Hình 3 mô tả chi tiết quá trình huấn luyện theo mô hình Efficient-B3. Vì vậy, EfficientNet-B3 là lựa chọn phù hợp cho các hệ thống cần hiệu quả cao nhưng bị hạn chế về tài nguyên tính toán.

III. KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN

A. THIẾT LẬP THỰC NGHIỆM

Toàn bộ quá trình huấn luyện và đánh giá mô hình được thực hiện trên môi trường Kaggle Notebook sử dụng GPU NVIDIA Tesla T4 với 16GB RAM. Để đảm bảo tính khách quan và công bằng trong quá trình so sánh, tất cả các mô hình đều được huấn luyện trên cùng bộ dữ liệu VnCoffeeSet, sử dụng cùng các kỹ thuật tăng cường dữ liệu, cùng bộ dữ liệu ảnh đã được tiền xử lý. Các mô hình được huấn luyện theo chiến lược hai giai đoạn (two-stage training) [11]. Thuật toán có áp dụng cơ chế dừng sớm (early stopping) trong quá trình huấn luyện nhằm hạn chế hiện tượng quá khớp và lựa chọn mô hình có hiệu năng tốt nhất trên tập dữ liệu xác thực. Bên cạnh đó, để cải thiện quá trình hội tụ của mô hình, nghiên cứu áp dụng cơ chế điều chỉnh tốc độ học Cosine Annealing Scheduler, cho phép tốc độ học giảm dần theo hàm cosine trong suốt quá trình huấn luyện. Chiến lược này giúp mô hình ổn định hơn ở các giai đoạn cuối của quá trình huấn luyện và đạt được nghiệm tối ưu tốt hơn. Việc kết hợp thuật toán tối ưu AdamW với cơ chế điều chỉnh learning rate theo Cosine Annealing góp phần nâng cao hiệu quả huấn luyện và cải thiện độ chính xác của mô hình trong bài toán phân loại chất lượng hạt cà phê rang.

Trong nghiên cứu này, bên cạnh kỹ thuật tăng cường dữ liệu, mô hình áp dụng hàm mất mát có trọng số (weighted loss), trong đó trọng số của từng lớp được xác định dựa trên tần suất xuất hiện trong tập huấn luyện. Cụ thể, giả sử tập dữ liệu có K lớp, với n_i là số lượng mẫu của lớp thứ i, khi đó:

Tổng số mẫu: $$N = \sum_{i=1}^{K} n_i$$

Trọng số của lớp thứ i: $$w_i = \frac{N}{K \cdot n_i}$$

Với phương pháp này, các lớp có số lượng mẫu ít sẽ nhận được trọng số lớn hơn, từ đó đóng góp nhiều hơn vào hàm mất mát trong quá trình huấn luyện. Ngoài ra, để tránh hiện tượng trọng số quá lớn gây mất ổn định, các giá trị w_i có thể được chuẩn hóa hoặc giới hạn trong một khoảng nhất định. Trong quá trình huấn luyện, các trọng số này được tích hợp trực tiếp vào hàm Cross-Entropy Loss có trọng số. Ở mỗi bước cập nhật, mô hình tính toán giá trị mất mát cho từng mẫu, sau đó nhân với trọng số tương ứng của lớp trước khi thực hiện lan truyền ngược (backpropagation).

Thời gian huấn luyện trung bình cho mỗi mô hình dao động từ 14 đến 22 phút, tùy thuộc vào độ phức tạp của kiến trúc mạng và số lượng tham số cần tối ưu. Trong giai đoạn thứ nhất, các lớp trích xuất đặc trưng của mô hình được giữ nguyên trọng số từ mô hình tiền huấn luyện trong 3 epochs rút trích đặc trưng ảnh. Trong giai đoạn thứ hai, một số lớp ở tầng sâu của mạng được mở khóa và tiếp tục huấn luyện trong 20 epochs nhằm tinh chỉnh các đặc trưng đã học để phù hợp hơn với bộ dữ liệu hạt cà phê rang. Bảng siêu tham số được thiết lập trong các mô hình EfficientNet-B3, Inception-V3 và ResNet-50, VGG16 được mô tả trong bảng 3 và bảng 4.

Bảng 3. *Bảng tham số huấn luyện trong giai đoạn 1*

Stage 1: Feature Extraction Stage (Giai đoạn trích xuất đặc trưng)				
STT	Mô hình	Learning Rate	Weight_decay	Input size
1	EfficientNet-B3	3e-5	1e-4	300x300
2	Inception-V3	1e-4	2e-4	299x299
3	ResNet-50	3e-5	1e-4	224x224
4	VGG16	1e-4	2e-4	224x224

Bảng 4. *Bảng tham số huấn luyện trong giai đoạn 2*

Stage 2: Fine-tuning Stage (Giai đoạn tinh chỉnh mô hình)						
STT	Mô hình	Unfreeze	Learning Rate	Learning Rate của classifier	Weight_decay	Input size
1	EfficientNet-B3	block3-6	8e-6, 2e-5, 5e-5, 6e-5	2e-4	2e-4	300x300
2	Inception-V3	Mixed_6e, 7e	2e-5, 3e-5	1e-3	2e-4	299x299
3	ResNet-50	layer2-4	5e-6, 5e-6, 2e-5	8e-5	2e-4	224x224
4	VGG16	features [17:]	5e-6	1e-4	3e-4	224x224

B. TỐI ƯU HÓA THAM SỐ CHO MÔ HÌNH

Để đảm bảo độ tin cậy của kết quả thực nghiệm, mỗi mô hình được huấn luyện 5 lần với các giá trị seed ngẫu nhiên khác nhau. Kết quả được trình bày trong bảng 5 dưới dạng giá trị trung bình và độ lệch chuẩn (mean ± standard deviation). Trong đó, mô hình EfficientNet-B3 đạt hiệu suất cao nhất với 99.48 ± 0.08, vượt trội so với ResNet50 (99.02 ± 0.34), VGG16 (99.11 ± 0.18) và Inception-V3 (98.64 ± 0.47).

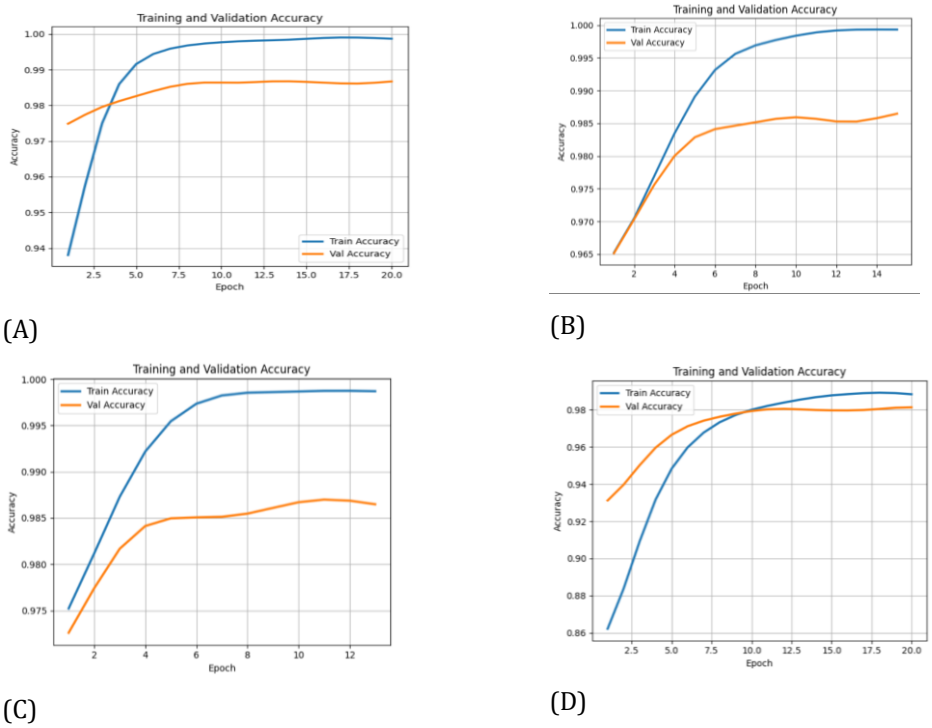
Bảng 5. *Bảng kết quả Accuracy của mô hình tương ứng 5 lần thực nghiệm*

Mô hình	**Lần 1**	**Lần 2**	**Lần 3**	**Lần 4**	**Lần 5**	**Mean ± Std**
EfficientNet-B3	99.53	99.52	99.55	99.37	99.42	99.48 ± 0.08
ResNet50	99.05	99.05	99.52	98.89	98.57	99.02 ± 0.34
VGG16	99.05	99.21	99.37	98.89	99.05	99.11 ± 0.18
Inception-V3	98.57	98.89	99.21	98.57	97.94	98.64 ± 0.47

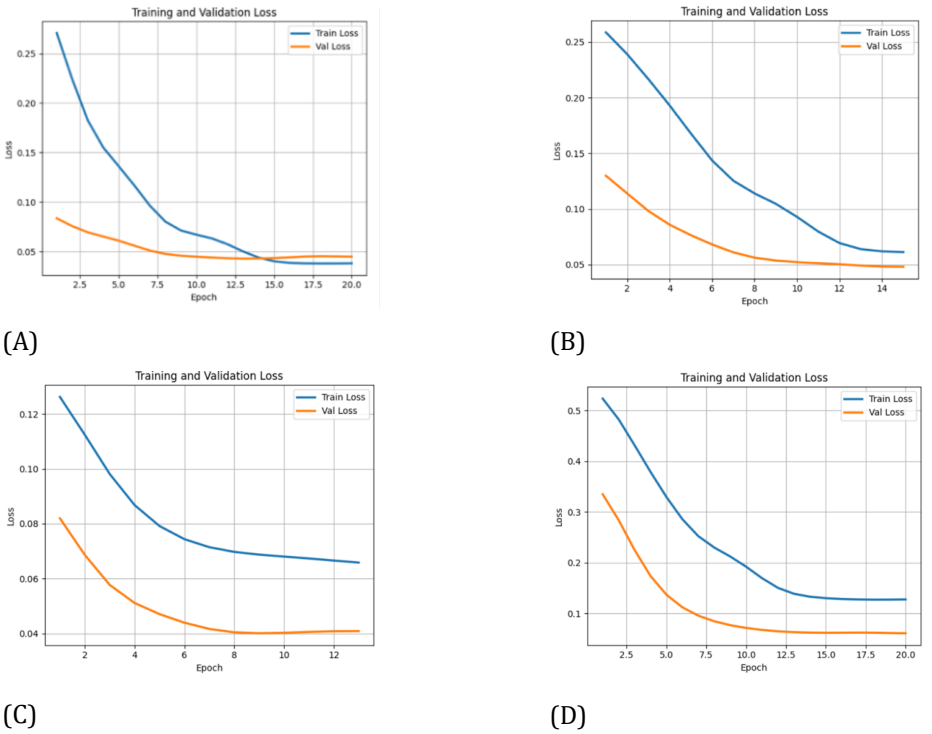
Nhằm đánh giá sự khác biệt có ý nghĩa thống kê về hiệu năng giữa các mô hình, chúng tôi tiến hành kiểm định t-test cặp (paired t-test) với mức ý nghĩa 95%. Giả thuyết không (H_0) được đặt ra là không có sự khác biệt đáng kể về hiệu năng giữa các mô hình. Kết quả cho thấy mô hình EfficientNet-B3 vượt trội so với ResNet-50 ($p = 0.0237$), VGG16 ($p = 0.0030$) và Inception-V3 ($p = 0.0113$). Do tất cả các giá trị p đều nhỏ hơn 0.05, sự khác biệt này có ý nghĩa thống kê, chứng tỏ hiệu suất của EfficientNet-B3 là thực và không phải do yếu tố ngẫu nhiên.

Bên cạnh bảng kết quả thực nghiệm trên, Hình 4 trình bày các biểu đồ biểu diễn độ chính xác (Accuracy) và Hình 5 trình bày các biểu đồ biểu diễn độ mất mát (Loss) của bốn mô hình tương ứng với lần thực nghiệm khi cố định giá trị seed = 42. Các biểu đồ cho thấy xu hướng hội tụ và mức độ ổn định của từng mô hình trong quá trình huấn luyện.

Quá trình huấn luyện theo mô hình ResNet-50 khá ổn định. Giá trị train loss và validation loss giảm dần theo số epoch, trong khi train accuracy và validation accuracy tăng tương đối đều. Khoảng cách giữa các đường train và validation tương đối nhỏ cho thấy mô hình có khả năng tổng quát hóa tốt và ít xảy ra hiện tượng overfitting.



Hình 4. *Biểu đồ accuraccy của (A) ResNet-50, (B) VGG16, (C) EfficientNet-B3 và (D) Inception-V3.*



Hình 5. *Biểu đồ loss của (A) ResNet-50, (B) VGG16, (C) EfficientNet-B3 và (D) Inception-V3.*

Các kết quả thực nghiệm cho thấy hiệu năng của các mô hình CNN. Mô hình VGG16 có khả năng học đặc trưng tốt nhưng tốc độ hội tụ chậm và dễ xuất hiện hiện tượng chênh lệch giữa train accuracy và validation accuracy do số lượng tham số lớn. EfficientNet-B3 cho thấy hiệu suất vượt trội với quá trình hội tụ nhanh, ổn định và khả năng tổng quát hóa tốt, thể hiện qua khoảng cách nhỏ giữa train và validation accuracy. Bên cạnh đó, Inception-V3 cũng đạt kết quả khả quan nhờ khả năng khai thác đặc trưng đa tỷ lệ, tuy nhiên kiến trúc phức tạp hơn và đòi hỏi chi phí tính toán cao hơn so với các mô hình khác.

Sau khi huấn luyện, tất cả các mô hình được đánh giá theo các chỉ số chuẩn như accuracy, precision, recall và F1-score và được ghi nhận, phân tích tại lần thực nghiệm thứ 2 (seed 42). Kết quả này được mô tả trong Bảng 6.

Bảng 6. *Tổng hợp kết quả kiểm tra của các mô hình*

Pre-trained Model	Accuracy	Precision	Recall	F1-score
ResNet-50	99.05	99.05	99.05	99.05
VGG16	99.21	99.21	99.21	99.21
EfficientNet-B3	99.52	99.53	99.52	99.53
Inception-V3	98.89	98.89	98.89	98.89

Kết quả cho thấy tất cả các mô hình học sâu đều đạt hiệu suất cao trong bài toán phân loại hạt cà phê rang. Trong đó, EfficientNet-B3 đạt kết quả tốt nhất với độ chính xác khoảng 99.52% cùng các giá trị Precision, Recall và F1-score gần như tương đương, cho thấy khả năng phân loại ổn định và cân bằng giữa các lớp. VGG16 và ResNet-50 cũng đạt độ chính xác cao trên 99%, chứng tỏ các kiến trúc CNN truyền thống vẫn có khả năng trích xuất đặc trưng hiệu quả từ dữ liệu ảnh. Trong khi đó, Inception-V3 có kết quả thấp hơn một chút so với các mô hình còn lại nhưng vẫn đạt độ chính xác gần 99%. Tổng thể, các kết quả thực nghiệm cho thấy các mô hình học sâu có khả năng nhận dạng tốt các đặc trưng của hạt cà phê rang, trong đó EfficientNet-B3 thể hiện ưu thế hơn về hiệu quả phân loại.

Với dữ liệu mất cân bằng giữa hai lớp (39.8%; 60.2%), việc Accuracy và Recall cùng đạt 99.52%, cũng như Precision và F1-score cùng đạt 99.53%, cho thấy mô hình không bị thiên lệch đáng kể về phía lớp chiếm đa số phản ánh đúng hiệu suất trên lớp thiểu số. Thêm vào đó, việc Recall vẫn duy trì mức tương đương Accuracy cho thấy mô hình vẫn nhận diện tốt các mẫu của lớp thiểu số, cho thấy hiệu suất của hai nhóm đều được kiểm soát tốt. Kết quả này cho thấy các kỹ thuật như weighted loss và tăng cường dữ liệu đã phát huy hiệu quả trong việc giảm tác động của mất cân bằng, giúp mô hình học được ranh giới phân lớp rõ ràng và tổng quát tốt trên cả hai lớp.

Bên cạnh độ chính xác phân loại, thời gian suy luận là một yếu tố quan trọng đối với các hệ thống triển khai trong môi trường thực tế yêu cầu xử lý theo thời gian thực. Do đó, nghiên cứu đã đánh giá thêm các chỉ số độ trễ (latency, ms/ảnh) và tốc độ xử lý (FPS) của từng mô hình. Thực nghiệm được đo trên GPU-T4x2 với kích thước batch 32. Kích thước ảnh đầu vào được điều chỉnh phù hợp với yêu cầu của từng kiến trúc mô hình. Thời gian suy luận (inference time) bao gồm độ trễ (latency) và số khung hình trên giây (FPS) được trình bày trong bảng 7.

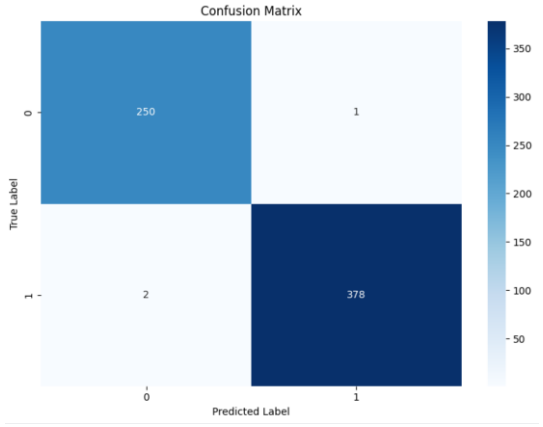
Bảng 7. *Tổng hợp về thời gian suy luận của các mô hình*

Model	Accuracy	Latency (ms/image)	FPS
ResNet-50	99.05	3.00	333.68
VGG16	99.21	4.76	210.09
EfficientNet-B3	99.52	4.39	227.81
Inception-V3	98.89	3.75	266.89

Mặc dù mô hình EfficientNet-B3 có thời gian suy luận cao hơn so với ResNet-50 và InceptionV3, mô hình này vẫn được lựa chọn cho hệ thống kiểm định chất lượng cà phê tự động. Lý do chính là EfficientNet-B3 không những đạt độ chính xác cao nhất (99.52%) mà còn đạt chi phí tính toán. Bên cạnh đó, mặc dù độ trễ của EfficientNet-B3 là 4.39 ms/ảnh (227.81 FPS), tốc độ này vẫn đáp ứng tốt yêu cầu của các hệ thống xử lý thời gian thực trong thực tế phù hợp cho các hệ thống kiểm định chất lượng cà phê tự động. Sự cân bằng giữa hiệu năng và tốc độ xử lý cho thấy EfficientNet-B3 là lựa chọn phù hợp cho các ứng dụng thực tế, nơi cả độ chính xác và thời gian phản hồi đều đóng vai trò quan trọng.

C. KẾT LUẬN THỰC NGHIỆM

Ma trận nhầm lẫn của mô hình EfficientNet-B3 trong Hình 6 cho thấy phần lớn các mẫu được phân loại đúng, tuy nhiên vẫn còn một số trường hợp nhầm lẫn giữa các lớp có đặc điểm hình ảnh tương đối giống nhau. Hình 6 cho thấy mô hình đạt hiệu suất phân loại rất cao với số lượng mẫu dự đoán chính xác chiếm ưu thế tuyệt đối. Cụ thể, có 250 mẫu lớp 0 và 378 mẫu lớp 1 được phân loại đúng, trong khi số lượng mẫu bị phân loại sai rất thấp (chỉ 1 mẫu lớp 0 bị nhầm sang lớp 1 và 2 mẫu lớp 1 bị nhầm sang lớp 0). Điều này cho thấy mô hình có khả năng nhận diện tốt cả hai lớp dữ liệu. Độ chính xác tổng thể đạt khoảng 99.52%, cùng với precision (99.53%) và recall (99.52%) ở mức rất cao, chứng tỏ mô hình không chỉ chính xác mà còn có khả năng tổng quát hóa tốt và hạn chế hiệu quả các lỗi phân loại.



Hình 6. *Ma trận nhầm lẫn của mô hình EfficientNet-B3*

Hiện tượng nhầm lẫn giữa hai lớp chủ yếu bắt nguồn từ sự tương đồng cao về đặc trưng và sự chồng lấn trong không gian đặc trưng, khiến ranh giới phân tách trở nên kém rõ ràng. Bên cạnh đó, biến thiên nội lớp lớn kết hợp với dữ liệu chưa đủ đa dạng làm suy giảm khả năng học các đặc trưng phân biệt. Ngoài ra, các phép tăng cường dữ liệu và điều kiện thu thập ảnh (ánh sáng, góc chụp) cũng có thể làm mờ các tín hiệu đặc trưng tinh vi, từ đó góp phần gia tăng tỷ lệ nhầm lẫn.

IV. KẾT LUẬN

Nghiên cứu này đề xuất một hệ thống phân loại chất lượng hạt cà phê rang dựa trên học sâu từ dữ liệu ảnh. Bộ dữ liệu ảnh đã được xây dựng và gán nhãn để huấn luyện và đánh giá mô hình. Bốn kiến trúc CNN gồm ResNet-50, VGG16, EfficientNet-B3 và Inception-V3 được áp dụng thông qua học chuyển giao. Kết quả thực nghiệm cho thấy các mô hình đạt hiệu suất cao, trong đó EfficientNet-B3 đạt độ chính xác khoảng 99.52% cùng các chỉ số precision, recall và F1-score cao.

Tuy nhiên, bộ dữ liệu chủ yếu được thu thập trong điều kiện chụp tương đối đồng nhất, có thể ảnh hưởng đến khả năng tổng quát hóa của mô hình trong môi trường thực tế. Trong các nghiên cứu tiếp theo, bộ dữ liệu sẽ được mở rộng với nhiều điều kiện chụp và mức độ lỗi khác nhau, đồng thời mô hình sẽ được thử nghiệm trong các hệ thống thực phân loại cà phê rang tự động.

V. TÀI LIỆU THAM KHẢO

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems (NIPS 2012)*, pp. 1097–1105

[2] Cynthia, E. P., Ismanto, E., Arifandy, M. I., Sarbaini, S., Nazaruddin, N., Manuhutu, M. A., & Akbar, M. A. (2022), Convolutional neural network and deep learning approach for image detection and identification, *Journal of Physics: Conference Series*, 2394(1)

[3] Sarvamangala, D. R., & Kulkarni, R. V. (2022). Convolutional neural networks in medical image understanding: A survey, *Evolutionary Intelligence*, 15(1), 1–22.

[4] International Coffee Organization-ICO (2023): Coffee market report. *International Coffee Organization*, London.

[5] Ranzel V. Dimaculangan; Marife A. Rosales (2024). Robusta Coffee Bean Defect Classification using Convolutional Neural Network, *https://ieeexplore.ieee.org/document/10636865*

[6] Hong-Danh Thai; Han-Jong Ko; Jun-Ho Huh (2024). Coffee Bean Defects Automatic Classification Realtime Application Adopting Deep Learning, *https://ieeexplore.ieee.org/document/10659872*

[7] Karen Simonyan; Andrew Zisserman (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition, *ImageNet Large Scale Visual Recognition Challenge 2014*

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun (2016), Deep Residual Learning for Image Recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778

[9] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, Zbigniew Wojna (2016), Rethinking the Inception Architecture for Computer Vision, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[10] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning (ICML), PMLR 97: 6105–6114.

[11] Sinno Jialin Pan and Qiang Yang (2010), A Survey on Transfer Learning, *IEEE Transactions on Knowledge and Data Engineering*

DEEP LEARNING-BASED AUTOMATED QUALITY INSPECTION OF ROASTED COFFEE BEANS

Ngo Minh Phat, Duong Thi Mong Thuy, Le Cung Tuong

ABSTRACT— In this paper, we propose the VnCoffeeFr architecture, a deep learning model developed to automate the process of classifying the quality of roasted coffee beans based on image feature analysis. We constructed the VnCoffeeSet dataset, consisting of 4,194 images of roasted coffee beans, and performed a normalization process to ensure consistency in lighting conditions, shooting angles, and image resolution. Based on this dataset, popular CNN architectures such as ResNet-50, EfficientNet-B3, VGG16, and Inception-V3 were integrated into the VnCoffeeFr architecture for model training and performance evaluation. Experimental results show that the models achieve high accuracy, with EfficientNet-B3 reaching 99.52%, thereby demonstrating the effectiveness and feasibility of VnCoffeeFr in building an automated coffee quality inspection system.

Keywords — Deep Learning, Computer Vision, Convolutional Neural Networks, Industrial Automation.



Ngô Minh Phát là sinh viên tại Khoa Công nghệ thông tin, Trường Đại học Công nghệ TP.HCM (HUTECH), Việt Nam.



Dương Thi Mộng Thùy nhận bằng thạc sĩ ngành khoa học máy tính tại Trường Đại học Khoa học tự nhiên (Việt Nam) vào năm 2013. Hiện tại, bà là giảng viên Khoa Công nghệ thông tin, Trường Đại học Công thương Thành phố Hồ Chí Minh (HUIT), Việt Nam.



Lê Cung Tưởng nhận bằng tiến sĩ ngành Khoa học máy tính tại Trường Đại học Sejong, Hàn Quốc vào năm 2020. Hiện tại, ông là giảng viên kiêm nhà nghiên cứu tại Khoa Công nghệ thông tin, Trường Đại học Công nghệ TP.HCM (HUTECH), Việt Nam.

Các lĩnh vực nghiên cứu của ông bao gồm học máy, học trên tập dữ liệu mất cân bằng, học sâu, trí tuệ doanh nghiệp, phân tích dữ liệu, khai phá dữ liệu và khai phá mẫu.

Email: lc.tuong@hutech.edu.vn.