HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Phạm Tuấn Anh

NGHIÊN CỨU GIẢI PHÁP AI TRÊN BIÊN MẠNG SỬ DỤNG CHO BÀI TOÁN CHUẨN ĐOÁN SÂU BỆNH

Chuyên ngành: KỸ THUẬT VIỄN THÔNG
Mã số: 8.52.02.08 (Kỹ thuật Viễn thông)

TÓM TẮT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

HÀ NỘI - NĂM 2024

Đề án tốt nghiệp được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: ……

(Ghi rõ học hàm, học vị)

Phản biện 1: ……

Phản biện 2: ……..

Đề án tốt nghiệp sẽ được bảo vệ trước Hội đồng chấm đề án tốt nghiệp thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng .... năm .....

Có thể tìm hiểu đề án tốt nghiệp tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

MỤC LỤC

I. MỞ ĐẦU

1. Lý do chọn đề tài

Đề tài "Nghiên cứu giải pháp AI trên biên mạng sử dụng cho bài toán chuẩn đoán sâu bệnh" được triển khai vì đề tài có ý nghĩa thực tiễn, có thể ứng dụng để giải quyết vấn đề sâu bệnh trong nông nghiệp ở Việt Nam, góp phần nâng cao năng suất và chất lượng nông sản. Kết quả của đề tài sẽ là một giải pháp AI có khả năng giám sát và dự đoán sâu bệnh từ ảnh lá cây với độ chính xác cao, có thể được ứng dụng trong nông nghiệp, lâm nghiệp, môi trường, v.v.

2. Tổng quan về vấn đề nghiên cứu

Việc giám sát và phát hiện sâu bệnh thường gặp nhiều khó khăn, đặc biệt là ở các vùng nông thôn, miền núi, nơi có điều kiện kinh tế - xã hội khó khăn [2-5]. Các phương pháp giám sát và phát hiện sâu bệnh truyền thống đang gặp phải các hạn chế của riêng họ. Đã có rất nhiều những mô hình, hệ thống, hay các đề tài hay được triển khai [1][4] [6] nhưng đều gặp phải những vấn đề chung, để triển khai thành công đề tài, cần giải quyết các vấn đề về dữ liệu, hệ thống, mô hình, … Kết hợp với những nghiên cứu mới và tân tiến trên thế giới [9-16], việc giải quyết các vấn đề trên sẽ góp phần nâng cao hiệu quả của đề tài.

3. Mục đích nghiên cứu

- Thu thập và xử lý thông tin từ hệ thống IoT cho giải pháp chuẩn đoán sâu bệnh tự động.
- Tìm kiếm giải pháp chuẩn đoán sâu bệnh dựa trên hình ảnh có độ chính xác cao và thuận tiện cho người sử dụng.

II. NỘI DUNG

CHƯƠNG 1: TỔNG QUAN VỀ HỆ THỐNG IOT NÔNG NGHIỆP

1.1 Tổng quan về nông nghiệp thông minh và thực trạng tại Việt Nam.

1.1.1 Tổng quan về nông nghiệp thông minh

Nông nghiệp thông minh là một hệ thống sản xuất nông nghiệp ứng dụng các công nghệ tiên tiến như Internet vạn vật (IoT), trí tuệ nhân tạo (AI), dữ liệu lớn (Big Data), robot, cảm biến,... nhằm tối ưu hóa quy trình sản xuất, tăng hiệu quả và năng suất, đồng thời giảm thiểu tác động đến môi trường. Nền sản xuất nông nghiệp trên thế giới nói chung và tại Việt Nam nói riêng, đang phải chịu áp lực từ hai bài toán lớn là gia tăng dân số và giảm sút diện tích đất nông nghiệp [4]. Trong ngành nông nghiệp, để tận dụng hiệu quả và bền vững các nguồn tài nguyên, việc chuyển đổi số hóa là cần thiết [8, 9].

1.1.2 Thực trạng nông nghiệp thông minh tại Việt Nam

Hiện nay, Việt Nam vẫn là một quốc gia chủ yếu dựa vào nông nghiệp với hơn 66,9% dân số sống ở vùng nông thôn và 42% lao động trong toàn xã hội làm việc trong ngành nông nghiệp [3]. Tuy nhiên, việc ứng dụng nông nghiệp thông minh đáp ứng các công nghệ và tiêu chuẩn trên tại Việt Nam còn gặp nhiều hạn chế về hạ tầng công nghệ thông tin, hệ thống internet và mạng lưới viễn thông chưa được phủ sóng rộng khắp, đặc biệt là ở khu vực nông thôn. Hơn nữa, việc thiếu hụt nguồn nhân lực có trình độ điển hình là nông dân Việt Nam chủ yếu là lao động già, trình độ học vấn và kỹ năng sử dụng công nghệ còn hạn chế. Cùng với đó là chi phí đầu tư cao, việc ứng dụng các công nghệ cao vào sản xuất nông nghiệp cần có vốn đầu tư lớn, trong khi nhiều hộ nông dân còn gặp khó khăn về tài chính [6].

1.2 Khái quát các hệ thống IoT nông nghiệp sử dụng hiện nay và xu hướng.

1.2.1 Tổng quan về hệ thống IoT nông nghiệp

Từ những khó khăn hiện đang gặp phải đã phân tích ở trên, những mô hình IoT trong nông nghiệp hiện tại Việt Nam nhìn chung mới phổ biến bao gồm các mô IoT truyền thống, ví dụ như các hệ thống giám sát thời tiết, đo lường nhiệt độ, độ ẩm, ánh sáng, ... kết hợp với ứng dụng di động và quản lý dữ liệu tập trung trên máy chủ đám mây. Một số ứng dụng của IoT trong nông nghiệp hiện nay đang được sử dụng phổ biến trong các hệ thống lớn và nhỏ.

Cấu trúc của một hệ thống IOT gồm bốn thành phần cơ bản chính gồm: Cảm biến (Things), Trạm kết nối (Gateways), Hạ tầng mạng (Internet) và cuối cùng là lớp dịch vụ (Service).

Giống như trong các ngành công nghiệp khác, ứng dụng IoT trong nông nghiệp hứa hẹn hiệu quả hơn nhiều so với phương pháp thủ công trước đây, giúp giảm tài nguyên và chi phí, tự động hóa dựa trên phân tích dữ liệu, và tối ưu hóa quy trình. Tuy nhiên, riêng đối với ngành nông nghiệp, vai trò của IoT là vô cùng quan trọng. Nó sẽ mang tới các giải pháp bước ngoặt, giải quyết những vấn đề cấp bách liên quan tới sự sinh tồn và phát triển của loài người. Giúp nâng cao chất lượng nông phẩm, cải thiện năng suất canh tác, bảo đảm vệ sinh an toàn thực phẩm, xây dựng một hệ thống trồng trọt bền vững và chính xác trước các biến đổi khó lường của khí hậu hiện nay.

1.2.2 Khảo sát mô hình nông nghiệp thông minh tại Việt Nam

Một vài nghiên cứu như "Ứng dụng công nghệ IoT và AI giám sát và điều khiển nhà nuôi chim yến thông minh" [1], ứng dụng những công nghệ mới như công trình nghiên cứu [3] với đề tài "Ứng dụng công nghệ xử lý ảnh kết hợp IoT để theo dõi và phân tích tình trạng quả trên cây cà chua" (2022) và một vài nghiên cứu khác là một trong những nghiên cứu mới ứng dụng AI và IoT dẫn tới một xu hướng mới ứng dụng khoa học dữ liệu và internet vạn vật vào canh tác nông nghiệp. Tuy nhiên, nghiên cứu gặp phải một vài hạn chế như thời gian và tốc độ xử lý còn chậm, hệ thống camera chưa xử lí được điều kiện thiếu ánh sáng, và chưa ứng dụng được khoa học

dữ liệu và học sâu, máy học, từ đó mới chỉ dừng lại ở quy mô nghiên cứu mà chưa thể ứng dụng rộng rãi. Một vài công trình nghiên cứu [5,7] như "Hệ thống so màu lá lúa trên thiết bị di động" (2016) đã ứng dụng máy học bằng kỹ thuật so khớp ảnh và kỹ thuật kNN (k-Nearest Neighbors) nhằm mục đích so màu lá lúa tự động từ ảnh chụp trên thiết bị di động nhằm xác định lượng phân đạm cần thiết (tương đối) để bón cho cây lúa dựa trên độ đậm của lá lúa. Từ tập ảnh chụp từ điện thoại, hệ thống tiến hành tiền xử lý, khử nhiễu và sau đó thực hiện so màu lá lúa bằng phương pháp so khớp ảnh và kỹ thuật máy học.

Tuy nhiên, một trong những điểm hạn chế của hệ thống còn tồn tại là kết quả thu được chỉ dừng ở mức độ xử lý và tính toán riêng lẻ, đơn thiết bị trên duy nhất một thiết bị đầu cuối, khác với giải pháp AI tại biên mạng sẽ xử lý và chia sẻ tài nguyên tính toán trên nhiều điểm biên và ra quyết định khi kết nối đa thiết bị, dự án còn chưa có tính kết nối vạn vật và kết nối tập trung trên máy chủ đám mây, chưa có tính chất thu thập dữ liệu và cải tiến thuật toán/mô hình theo thời gian sử dụng.

1.3 Lý thuyết xử lý ảnh và phương pháp xử ý nhằm chuẩn đoán sâu bệnh.

1.3.1 Thông tin biểu diễn dưới dạng ảnh và lý thuyết xử lý ảnh truyền thống

Xử lý ảnh là một lĩnh vực khoa học máy tính tập trung vào việc chuyển đổi và cải thiện hình ảnh thông qua các thuật toán và phương pháp tính toán. Nó có thể được hiểu đơn giản là quá trình biến đổi ảnh đầu vào thành ảnh đầu ra với các đặc điểm mong muốn hoặc trích xuất thông tin từ ảnh.

1.3.2 Phương pháp xử lý ảnh và trích xuất đặc trưng nhằm chuẩn đoán sâu bệnh

Dựa trên các lý thuyết của thông tin biểu diễn dưới dạng ảnh và sự phát triển của khoa học dữ liệu và học sâu, cơ sở lý thuyết và thuật toán tự động trích xuất đặc trưng của dữ liệu ảnh trở nên phổ biến và ngày càng hiệu quả. Học sâu (Deep learning) là một phương pháp tiên tiến trong lĩnh vực trí tuệ nhân tạo, nhằm giúp máy tính học và xử lý dữ liệu theo mô phỏng quá trình tư duy trong não bộ con người. Mô hình học sâu có khả năng nhận diện và hiểu được nhiều mẫu phức tạp trong hình ảnh, văn bản, âm thanh và các dữ liệu khác, từ đó đưa ra thông tin và dự đoán chính xác.

Cấu trúc của một mạng neuron cơ bản:

Một kiến trúc mạng nơ ron được sử dụng phổ biến trong thị giác máy tính đó là mạng nơ ron tích chập (Convolutional Neural Network - CNN). Đây là một kiến trúc đa tầng được sử dụng để xử lý dữ liệu đầu vào và tạo ra một tập hợp số để so sánh với các dữ liệu đã biết, nhằm định nghĩa và phân loại dữ liệu đó.

Lý do là vì ngay cả với hình ảnh đơn giản nhất, các pixel liền kề có sự phụ thuộc lẫn nhau, việc biến đổi thành vector sẽ làm mất đi thông tin phụ thuộc này và làm thay đổi ý nghĩa của bức hình. Ví dụ, biểu tượng của mắt, miệng con người hoặc thậm chí là cạnh của một đối tượng khác được xây dựng từ một số pixel được bố trí theo một cách nhất định. Nếu xử lý hình ảnh thành một vector, những phụ thuộc này bị mất và làm giảm độ chính xác của mô hình.

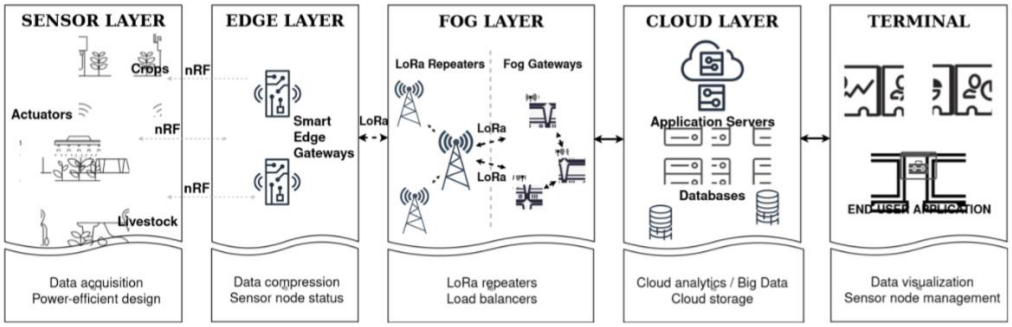
1.2 Kết luận chương.

CHƯƠNG 2: GIẢI PHÁP AI TRÊN BIÊN MẠNG

2.1 Mô hình triển khai hệ thống AI trên biên mạng.

2.1.1 Tổng quan hạ tầng mạng trong nông nghiệp thông minh

Internet of Things (IoT) có thể được định nghĩa là một nền tảng nơi các đối tượng ảo và vật lý được kết nối với nhau và giao tiếp với nhau. Hệ thống IoT bao gồm các công nghệ khác nhau như mạng cảm biến không dây, điện toán đám mây và trí thông minh nhúng. Các hệ thống này cung cấp các dịch vụ tiên tiến như giám sát từ xa theo thời gian thực, phân tích trực tuyến và quản lý từ xa. Điện toán biên và sương mù có thể được minh họa như một đám mây nhỏ gần biên của mạng hơn. Nói cách khác, điện toán Edge và Fog đại diện cho sự hội tụ của các lớp mạng khác nhau thành các cổng thông minh được kết nối với nhau. Điện toán biên và sương mù có thể giúp khắc phục một số hạn chế của các hệ thống IoT có thể tập trung vào truyền thống.



Hình 2.1. Mô hình cơ bản của hệ thống IoT trong nông nghiệp tiên tiến

Mặc dù điện toán biên và sương mù có thể cung cấp nhiều dịch vụ tiên tiến, các hệ thống dựa trên sương mù vẫn không thể hoạt động bình thường ở các vùng sâu vùng xa, nơi Internet không ổn định hoặc không được phủ sóng, vì chúng thường dựa vào mạng cục bộ tốc độ cao để xử lý thời gian thực và các ứng dụng quan trọng về độ trễ.

Edge Intelligence có tiềm năng cung cấp trí tuệ nhân tạo cho mọi người và mọi tổ chức ở bất kỳ đâu có hạ tầng. Ngày nay, một số lượng lớn các cảm biến và thiết bị thông minh tạo ra một lượng lớn dữ liệu và đòi hỏi sức mạnh tính toán ngày càng tăng, đang thúc đẩy cốt lõi của các nhiệm vụ tính toán và dịch vụ từ đám mây đến biên của mạng. Bằng cách kết hợp IoT với AI, dữ liệu được thu thập bởi các nút có thể được sử dụng bằng cách áp dụng các kỹ thuật AI như học máy và học sâu. Kết quả là, khả năng học máy được di chuyển gần hơn với nguồn dữ liệu. Khái niệm này được gọi là Edge AI, hoặc Edge Intelligence, và nó cho phép khả năng mở rộng lớn hơn, mạnh mẽ và hiệu quả. Do đó, các mô hình học máy trong các hệ thống AI được kết hợp với khả năng kết nối và truyền dữ liệu của Internet vạn vật IoT. Nói cách khác, với sự kết hợp của AI trong các hệ thống IoT, chức năng của chúng không chỉ giới hạn trong việc thu thập và truyền thông tin mà thực sự hiểu và phân tích được dữ liệu.

2.1.2 Tổng quan về trí tuệ nhân tạo và hệ ra quyết định trong hạ tầng mạng

Ban đầu, Machine Learning và Deep Learning bị giới hạn ở Cloud, chủ yếu là do tính sẵn có và khả năng mở rộng của các tài nguyên yêu cầu tính toán cao cần thiết để xử lý các tác vụ ML. Việc kết hợp điện toán đám mây và IoT mang đến những lợi

ích thiết thực, đặc biệt khi sử dụng các cảm biến thông minh. Trước đây, dữ liệu thu thập từ các thiết bị IoT cơ bản (như camera hay micrô) được gửi đến đám mây để phân tích. Quá trình này tốn thời gian và gây tắc nghẽn mạng do lượng dữ liệu lớn.

Chuyển xử lý AI từ Đám mây sang các thiết bị biên được phân tán, kết nối cung cấp một giải pháp để khắc phục các tắc nghẽn, độ trễ và các vấn đề về quyền riêng tư của các ứng dụng AI dựa trên đám mây. So với các thiết bị IoT công suất thấp truyền thống, AIoT yêu cầu các thiết bị biên có đủ tài nguyên để thực hiện các tác vụ học máy trên thiết bị. Tuy nhiên, khả năng tài nguyên và năng lượng của các thiết bị biên tự nhiên bị hạn chế. Do đó, các ứng dụng AIoT dựa trên các thách thức cần tối ưu hóa để cân bằng:

- Chi phí phần cứng và hiệu suất
- Mô hình được tối ưu hóa cho nền tảng thiết bị.

2.1.3 Các mô hình AI đã được tối ưu trên biên

InceptionNets: Biến thể GoogLeNet với các mô-đun Inception đã được giới thiệu vào năm 2016 [7]. Inception-v3 đã đạt được hiệu suất phân loại tốt trong một số ứng dụng y sinh sử dụng học tập chuyển giao. Nó đề xuất một mô hình khởi đầu kết hợp nhiều bộ lọc tích chập có kích thước khác nhau vào một bộ lọc mới. Mục tiêu của mô-đun khởi động là hoạt động như một "trình trích xuất tính năng đa cấp" bằng cách tính toán các kết cấu 1×1, 3×3 và 5×5 trong cùng một mô-đun của mạng. Thiết kế như vậy làm giảm số lượng tham số được đào tạo và do đó làm giảm độ phức tạp tính toán.

ResNets: Các mô hình ResNet, dựa trên các kiến trúc sâu đã cho thấy các hành vi hội tụ tốt và độ chính xác hấp dẫn, được phát triển bởi He et al. [7]. ResNet được xây dựng bởi một số đơn vị còn lại xếp chồng lên nhau và được phát triển với nhiều số lớp khác nhau: 18, 34, 50, 101, 152 và 1202. Các đơn vị còn lại bao gồm tích chập, gộp và lớp. ResNet 50 chứa 49 lớp tích chập và một lớp được kết nối hoàn toàn ở cuối mạng. Để tiết kiệm tài nguyên máy tính và thời gian đào tạo, ResNet 50 đã được chọn để so sánh trong phần sau.

MobileNets: Có ba phiên bản MobileNets, mới nhất, MobileNet V3. Kiến trúc cốt lõi của MobileNetV1 dựa trên một kiến trúc được sắp xếp hợp lý sử dụng các lớp phức tạp có thể tách rời chiều sâu để xây dựng các mạng thần kinh sâu nhẹ. MobileNetV2 đã giới thiệu hai tính năng mới cho kiến trúc: nút cổ chai tuyến tính giữa các lớp và kết nối phím tắt giữa các bottlenecks. MobileNetV3 là phiên bản thứ ba của kiến trúc, cung cấp khả năng phân tích hình ảnh của nhiều ứng dụng di động phổ biến. Đóng góp chính của MobileNetV3 là sử dụng AutoML để tìm kiếm kiến trúc mạng lưới thần kinh tốt nhất có thể cho một vấn đề nhất định. Điều này trái ngược với thiết kế thủ công của các phiên bản kiến trúc trước đó.

MobileNetV3 là một mô hình DNN nhẹ được điều chỉnh cho phù hợp với các CPU của điện thoại di động hay các thiết bị biên thông qua sự kết hợp của tìm kiếm kiến trúc mạng nhận biết phần cứng (NAS) được bổ sung bởi thuật toán NetAdapt và sau đó được cải thiện thông qua các tiến bộ kiến trúc mới. MobileNetV3 được định nghĩa có hai mô hình: MobileNetV3- Large và MobileNetV3-Small. Các mô hình này được nhắm mục tiêu vào các trường hợp sử dụng tài nguyên cao và thấp tương ứng. MobileNets là một loạt các mạng nơ-ron sâu có trọng lượng nhẹ dựa trên các Depthwise Separable Convolutions. Tiếp sau đó phiên bản cải tiến từ Version 1 là MobileNetV2. MobileNetV2 tiếp tục sử dụng Depthwise Separable Convolutions, ngoài ra còn đề xuất thêm: Linear bottlenecks và Inverted Residual Block (shortcut connections giữa các bottlenecks). MobileNetV3 đạt được hiệu suất tốt hơn với ít FLOP hơn với các cải tiến mới so với các mô hình tiền nhiệm với khối kiến trúc mới như Hình 27 dưới đây. Trái ngược với phiên bản MobileNet trước đó được thiết kế thủ công, MobileNetV3 có thể tự tìm kiếm kiến trúc tốt nhất có thể trong không gian tìm kiếm phù hợp với các tác vụ thị giác máy tính di động.

Để khai thác hiệu quả nhất không gian tìm kiếm, hai kỹ thuật được triển khai theo trình tự là MnasNet và NetAdapt. Đầu tiên, tìm kiếm một kiến trúc thô bằng MnasNet, sử dụng tính năng học tăng cường để chọn cấu hình tối ưu từ một tập hợp các lựa chọn rời rạc. Sau đó, tinh chỉnh kiến trúc bằng cách sử dụng NetAdapt, một kỹ thuật bổ sung giúp cắt bỏ các kênh kích hoạt chưa được sử dụng theo mức độ nhỏ.

Để cung cấp hiệu suất tốt nhất có thể trong các điều kiện khác nhau, ta có thể tạo các mô hình lớn hoặc nhỏ. Ngoài ra cải tiến mạng bằng cách thiết kế lại các lớp tốn nhiều chi phí tính toán và sửa đổi hàm phi tuyến tính thành hard-swish (h-swish) dựa trên hàm phi tuyến tính của Swish để có thể khắc phục hạn chế lớn nhất của hàm Swish là nó rất kém hiệu quả khi tính toán trên phần cứng di động.

- MnasNet: Khối xây dựng chính của MnasNet là một khối còn lại đảo ngược (từ MobileNet V2 đề cập ở trên). Lấy cảm hứng từ sự tiến bộ trong tìm kiếm kiến trúc thần kinh AutoML, cách tiếp cận tìm kiếm kiến trúc MnasNet để thiết kế các mô hình di động bằng cách sử dụng học tăng cường. Tổng thể của phương pháp này bao gồm chủ yếu là ba thành phần: bộ điều khiển dựa trên RNN để học hỏi và lấy mẫu kiến trúc mô hình, một huấn luyện viên xây dựng và đào tạo các mô hình để có được độ chính xác và một động cơ suy luận để đo tốc độ mô hình trên điện thoại di động thực. MnasNet thực hiện một vấn đề tối ưu hóa đa ngôn từ nhằm mục đích đạt được cả accuracy cao và tốc độ cao.

- EfficientNets Lite: EfficientNet-Lite mang lại sức mạnh của EfficientNet cho các thiết bị biên và có năm biến thể, cho phép người dùng chọn từ tùy chọn độ trễ / kích thước mô hình thấp (EfficientNet-Lite0) đến độ chính xác cao(EfficientNet-Lite4). Một số hoạt động trong EffcientNet không được hỗ trợ tốt bởi một số máy gia tốc nhất định. Để giải quyết vấn đề không đồng nhất, EfficientNets ban đầu được điều chỉnh với các sửa đổi đơn giản sau:

• Loại bỏ một vài các layer mạng vì chúng không được hỗ trợ tốt.

• Thay thế tất cả các kích hoạt swish bằng RELU6, điều này cải thiện đáng kể chất lượng định lượng sau đào tạo.

• Cố định tham số và scale-down mô hình để giảm kích thước và tính toán của các mô hình thu nhỏ.

2.2 Các phương pháp xử lý dữ liệu và cải tiến hệ thống.

2.2.1 Ý tưởng chính trong việc cải tiến và tối ưu mô hình

Đầu tiên, có thể nói các thuật toán Neural Network Compression (nén mạng) là một nhánh nhỏ trong tập thuật toán tối ưu hóa mô hình (model optimization), nó

được sinh ra với mục đích giúp giải quyết bài toán khi triển khai (deploy) các model Deep Learning trên các thiết bị phần cứng không được mạnh mẽ như (mobile devices, edge devices ...).

2.2.2 Pruning (Cắt tỉa)

Cắt tỉa mạng được lấy cảm hứng bắt nguồn từ sự cắt tỉa liên kết nơ ron trong não người, nơi các liên kết thần kinh giữa các nơron(axon) bị phân giã hoàn toàn và chết đi xảy ra giữa thời thơ ấu và sự khởi đầu của dậy thì.

Bộ não con người lưu trữ thông tin bằng cách tạo ra các liên kết thần kinh. Khi một liên kết không được sử dụng trong một thời gian dài, nó sẽ bị cắt tỉa đi. Việc cắt tỉa này giúp não bộ tiết kiệm năng lượng và tăng hiệu quả hoạt động. Kỹ thuật Pruning trong học máy cũng dựa trên nguyên tắc tương tự. Kỹ thuật này loại bỏ các thành phần dư thừa trong mô hình, giúp mô hình nhỏ gọn và hiệu quả hơn.

2.2.3 Quantization (Lượng tử hóa)

Quantization là kỹ thuật tối ưu hóa việc lưu trữ trọng số trong mạng nơ-ron. Thay vì tập trung vào việc tối ưu hóa giá trị của trọng số, Quantization hướng đến việc giảm số lượng bit cần thiết để biểu diễn chúng mà vẫn đảm bảo độ chính xác của mô hình.

2.3 Kết luận chương

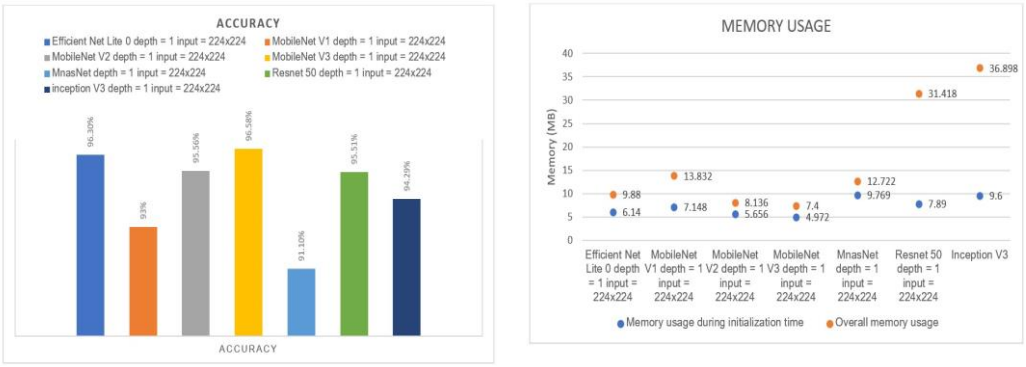
CHƯƠNG 3: ĐỀ XUẤT GIẢI PHÁP CHUẨN ĐOÁN SÂU BỆNH SỬ DỤNG MÔ HÌNH AI TẠI BIÊN

3.1 Đề xuất giải pháp cải tiến mô hình và dữ liệu chuẩn đoán sâu bệnh thông qua mô hình AI tại biên.

3.1.1 Đánh giá hiệu năng các mô hình AI được cải tiến để thực thi trên biên mạng

Với mục tiêu tập trung vào các thiết bị biên công suất thấp, đề án khảo sát và đánh giá một số mẫu máy hiện đại và nhẹ nhất hiện nay. Trong đề án, hiệu suất của

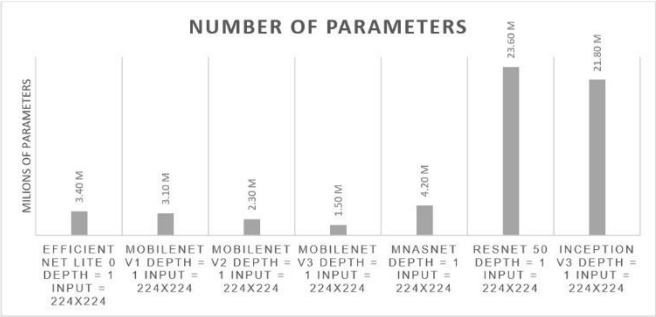
các mô hình DL đã được đánh giá trên Raspberry Pi 3 Model B. Mặc dù nguồn lực tính toán hạn chế, nền tảng nhúng chi phí thấp này có đủ sức mạnh tính toán để suy luận DNN theo thời gian thực. Với CPU ARM Cortex-A53 1.2GHz 64-bit lõi tứ có thể hoạt động ở tần số từ 700 MHz đến 1.2 GHz. Hệ thống tích hợp RAM 1GB LPDDR2 ở tốc độ 900MHz.



Hình 3.1. So sánh độ chính xác và tài nguyên sử dụng của các mô hình [13]

Độ chính xác: Kết quả được thể hiện trong hình 3.1. Kết quả cho thấy MobileNetV3 đạt độ chính xác cao nhất là 96,58%, cao hơn 0,28% so với vị trí thứ hai, EfficientNet Lite 0 (96,3%).

Sử dụng bộ nhớ: Hình 3.1 cho thấy bộ nhớ sử dụng các mô hình được đánh giá. Tất cả các mô hình trong bài báo được đánh giá, cụ thể với MobileNetV3 vượt trội hơn các kiến trúc khác khi chỉ sử dụng 4.972MB trong quá trình khởi tạo model và tổng cộng 7,4MB trong RAM 1GB của RPi 3B. MobileNetV2 và EfficientNet Lite 0 cũng cho kết quả đáng tin cậy khi chúng chỉ chiếm lần lượt 8.136MB và 9,88MB tổng lượng bộ nhớ sử dụng.

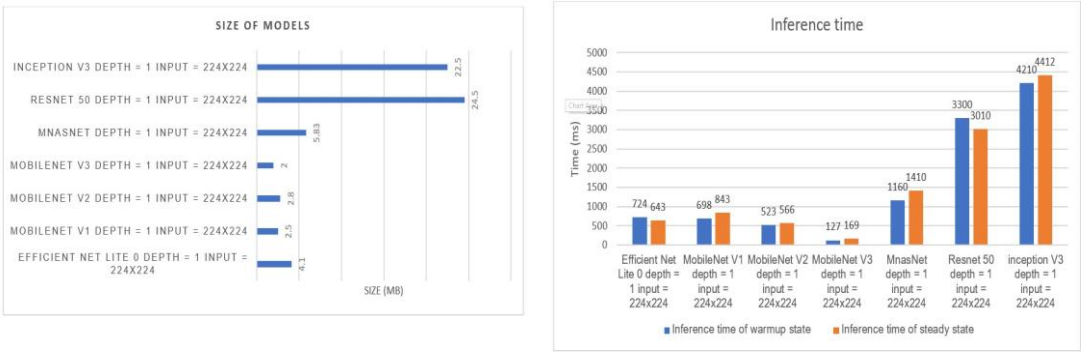


Hình 3.2 Đánh giá số lượng tham số của các mô hình [13]

Số lượng tham số: Hình 3.2 cho thấy kết quả số lượng tham số của mỗi mạng. Số lượng tham số của MobileNetV3 là nhỏ nhất so với các kiến trúc mạng khác (1,5 triệu), vượt trội so với các kiến trúc mạng phổ biến không được tối ưu hóa cho Edge/ Mobile như ResNet 50 (23,6 triệu) và InceptionV3 (21,8 triệu).

Kích thước của mô hình: Lượng tử hóa tham số có thể được sử dụng để giảm kích thước của mô hình. Các mô hình nhỏ hơn có những lợi ích sau:

- Kích thước lưu trữ / tải xuống nhỏ hơn: Các mô hình nhỏ hơn chiếm ít dung lượng lưu trữ hơn trên thiết bị của người dùng do đó yêu cầu ít thời gian và băng thông hơn để tải xuống.
- Sử dụng ít bộ nhớ hơn: Các mô hình nhỏ hơn sử dụng ít RAM hơn khi chúng chạy, giúp giải phóng bộ nhớ để các phần khác trong ứng dụng của bạn sử dụng và có thể chuyển thành hiệu suất và độ ổn định xác định tốt hơn.



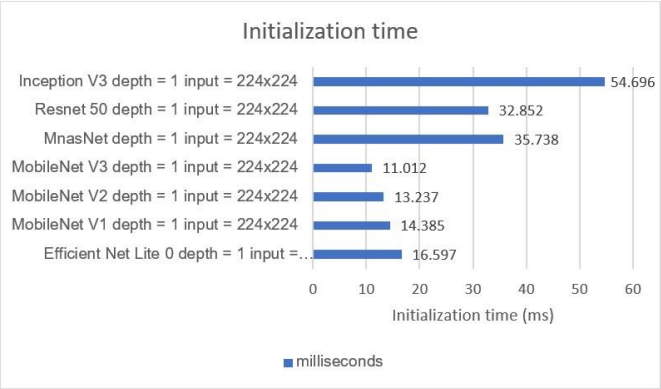
Hình 3.3: Đánh giá kích thước mô hình và tốc độ suy luận của mô hình [13]

Kết quả của lượng tử hóa là kích thước của các mô hình DNN như thể hiện trong hình 3.3. Các mô hình được thiết kế đặc biệt cho Edge / Mobile là MobileNets,

MnasNets, EfficientNets Lite được tham số hóa và tối ưu hóa cho bộ nhớ và kích thước mô hình tốt hơn so với các DNN thường được sử dụng.

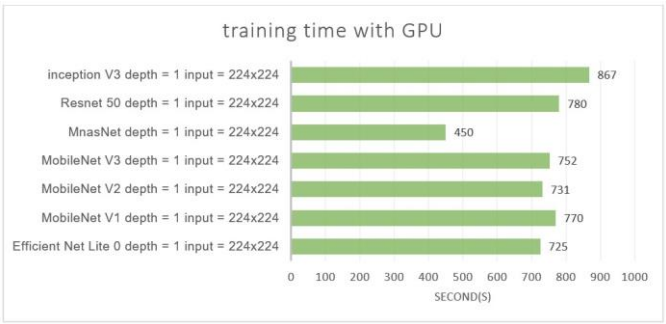
Thời gian suy luận: Kết quả trong hình 3.3, MobileNetV3 vẫn cho kết quả đáng kể nhất.

Thời gian khởi tạo: Thời gian khởi tạo mô hình của mobbileNet V3 là nhanh nhất với 11.012 ms, kết quả so sánh được thể hiện bằng hình 3.4



Hình 3.4 So sánh thời gian khởi tạo mô hình trên thiết bị [13]

Thời gian đào tạo: Hình 3.5 so sánh thời gian đào tạo của các mô hình



Hình 3.5: Thời gian đào tạo của các mô hình [13]

Thu nhỏ/tối ưu mô hình: Như vậy sau quá trình đánh giá, đề án đã chứng minh mô hình MobileNetV3 đạt độ hiệu quả cao trong công việc. Để triển khai hiệu quả MobileNetV3 trên các thiết bị cạnh / di động, đề án tiếp tụ giảm độ sâu và độ phân giải ảnh đầu vào. So sánh giữa mô hình thu nhỏ và mô hình bình thường được thể hiện trong Bảng 3.1. Mô hình thu nhỏ trong khi vẫn có độ chính xác thỏa đáng là 94,4%, chỉ bằng một nửa kích thước và mất ít thời gian hơn đáng kể để đào tạo.

Bảng 3.1. Mô hình thu nhỏ

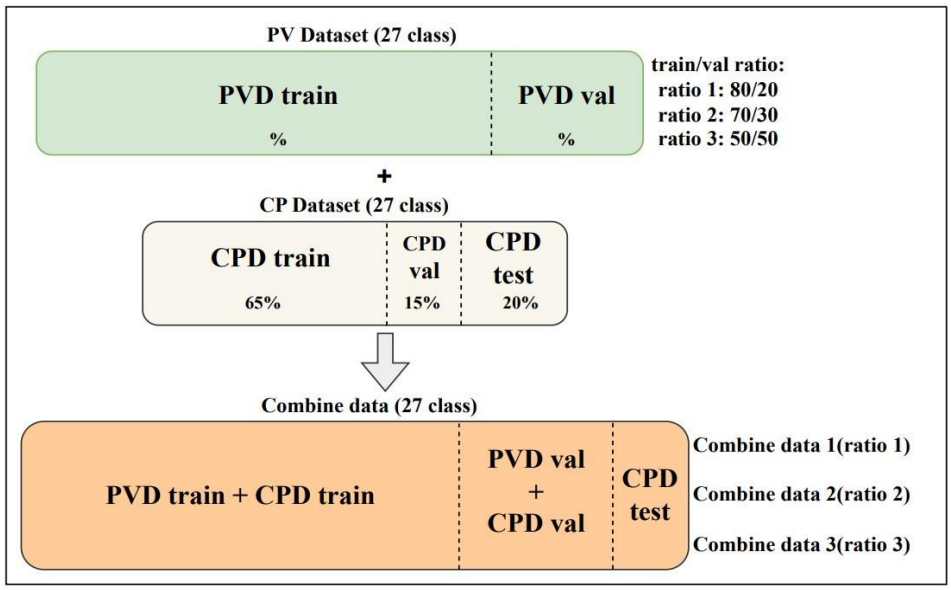
MobileNet V3	Độ sâu = 1,0 Đầu vào = 224x224	Độ sâu = 0,35 Đầu vào = 96x96
Độ chính xác	96.58%	94.4%
Sử dụng bộ nhớ (init/overall)	4.972 MB / 7.4 MB	4.464 MB / 5.644 MB
Thời gian suy luận (khởi tạo/tổng thể)	0,127 giây / 0,169 giây	0, 045 giây / 0, 043 giây
Thời gian khởi tạo	11.012 mili giây	0.815 mili giây
Thời gian đào tạo mô hình	752 giây	166 giây
Kích thước mô hình	2MB	797 KB
Số lượng tham số	1,5 triệu	0,4 triệu

3.1.2 Giải pháp cải tiến mô hình và dữ liệu chuẩn đoán sâu bệnh

Kế thừa từ kết quả đánh giá các mô hình của phần trước, đề án đã chọn ra mô hình tốt nhất để phát triển giải pháp là MobileNetV3. Nội dung phần tiếp theo tập trung vào việc phát triển một phương pháp để tối ưu hóa mô hình phân loại bệnh thực vật bằng DCNN (Mạng nơ-ron tích chập sâu) có tên là MobileNetV3 đã được chứng minh ở phần trước.

Tiền xử lý dữ liệu:

Đề án kết hợp hai bộ dữ liệu trong bài báo này: bộ dữ liệu phòng thí nghiệm (PVD) và bộ dữ liệu được thu thập tự nhiên (CPD). Tuy nhiên, có một vấn đề với số lượng lớp trong hai bộ dữ liệu; tức là PVD có 38 lớp, nhưng CPD chỉ có 27 lớp. Vì 27 lớp trong CPD đều được bao gồm trong PVD và nghiên cứu này nhằm mục đích làm cho mô hình có thể triển khai trong thực tế, đề án sẽ kiểm tra hiệu suất mô hình trên CPD kết hợp với 27 trong số 38 lớp của PVD. Sau khi kết hợp, tập dữ liệu mới sẽ có 27 lớp tương tự như CPD.

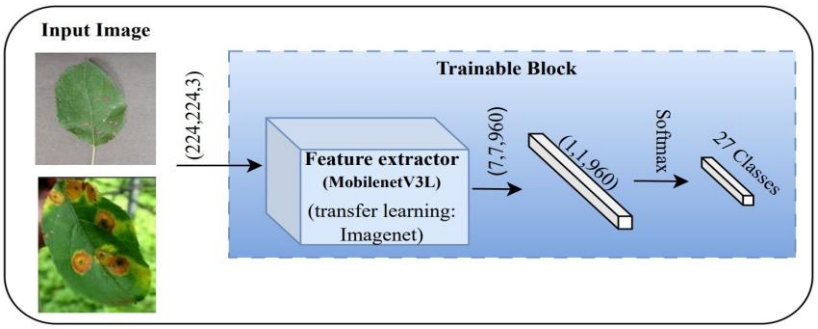


Hình 3.8. Bộ dữ liệu kết hợp của CPD và PVD

Trong nghiên cứu này, đề án kết hợp hai bộ dữ liệu bao gồm PVD và CPD theo cách minh họa trong Hình 3.8 Đầu tiên, PVD được chia ngẫu nhiên thành hai tập con: PVD train và PVD val. Việc phân chia này được thực hiện trong ba trường hợp khác nhau với tỷ lệ phân chia lần lượt là 80:20, 70:30 và 50:50. Tương tự, tập dữ liệu CPD được chia ngẫu nhiên một lần để tạo thành ba tập con: CPD train, CPD val và CPD test với tỷ lệ 65:15:20.

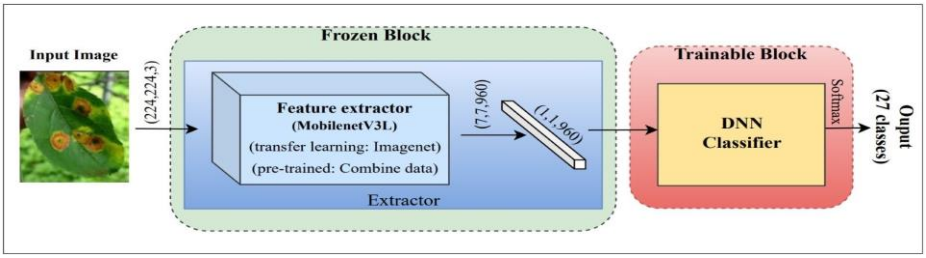
Quá trình hợp nhất dữ liệu chỉ được áp dụng trong các giai đoạn đào tạo và xác nhận. Khi tiến hành thử nghiệm mô hình cuối cùng, đề án tập trung vào tập dữ liệu phức tạp hơn, tức là CPD. húng tôi thực hiện kết hợp theo cặp: tập train CPD được kết hợp với tập train PVD, tập CPD val được kết hợp với PVD val và tập test CPD được giữ nguyên. Kết quả của sự kết hợp này tạo ra một tập dữ liệu có tên là "Combine Dataset", bao gồm ba trường hợp tương ứng với ba tỷ lệ phân chia PVD: Kết hợp dữ liệu 1, Kết hợp dữ liệu 2 và Kết hợp dữ liệu 3. Mỗi dữ liệu Kết hợp này bao gồm ba tập hợp con dữ liệu, được sử dụng để đào tạo, xác thực và thử nghiệm tương ứng. Bằng cách so sánh và phân tích kết quả của mô hình trên mỗi trong ba tỷ lệ phân chia tập dữ liệu PVD, đề án muốn xác định trường hợp phân vùng tối ưu nhất của mô hình của đề án.

Ngoài ra, việc tăng cường dữ liệu cũng được áp dụng trước khi cho huấn luyện mô hình đào tạo. Các kỹ thuật tăng cường được sử dụng bao gồm lật ngẫu nhiên hình ảnh theo chiều ngang, xoay ngẫu nhiên hình ảnh với góc tối đa 30 độ, phóng to ngẫu nhiên hình ảnh với hệ số tối đa 30% và Thay đổi độ tương phản của hình ảnh với hệ số tối đa là 30%. Đào tạo trước của đề án sử dụng trình trích xuất MobileNetV3 được rút gọn để trích xuất các tính năng từ dữ liệu hình ảnh đầu vào (Hình 3.9). Trọng số ban đầu của bộ trích xuất được học từ tập dữ liệu ImageNet. Cuối cùng, bộ trích xuất của MobileNetV3 được đào tạo lại trên tập dữ liệu kết hợp để cập nhật trọng số.



Hình 3.9. Khối trích xuất tính năng của lá

Khi kết thúc quá trình đào tạo trước, mô hình sẽ được tinh chỉnh. Đầu ra của model sẽ được trích xuất và tách hàm softmax ở khối "Trainable Block" trên hình. Điều này giúp kết quả có được tính năng của hình ảnh được trích xuất sau khi đi qua trình trích xuất để xử lý phân loại thêm. Sau khi được đào tạo trước, bộ phân loại DNN được đặt ở đầu ra của mô hình. Cụ thể, 4 lớp Dense với số lượng nút 512, 512, 128, 128 và 27 tương ứng với 27 lớp trong tập dữ liệu, sử dụng chức năng kích hoạt "RELU" được tăng cường nhằm nâng cao khả năng phân loại.

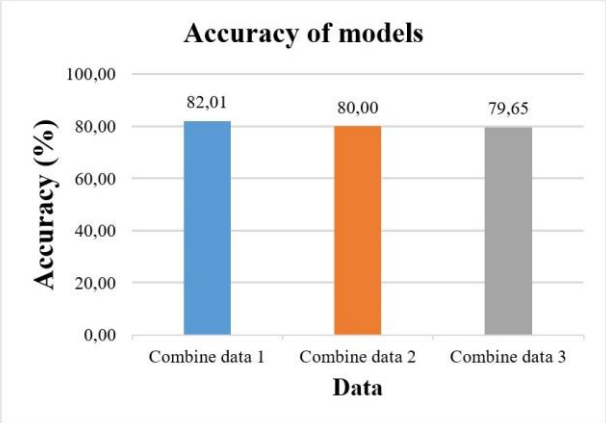


Hình 3.10. Đề xuất mô hình đào tạo với máy trích xuất được đào tạo trước và phân loại DNN

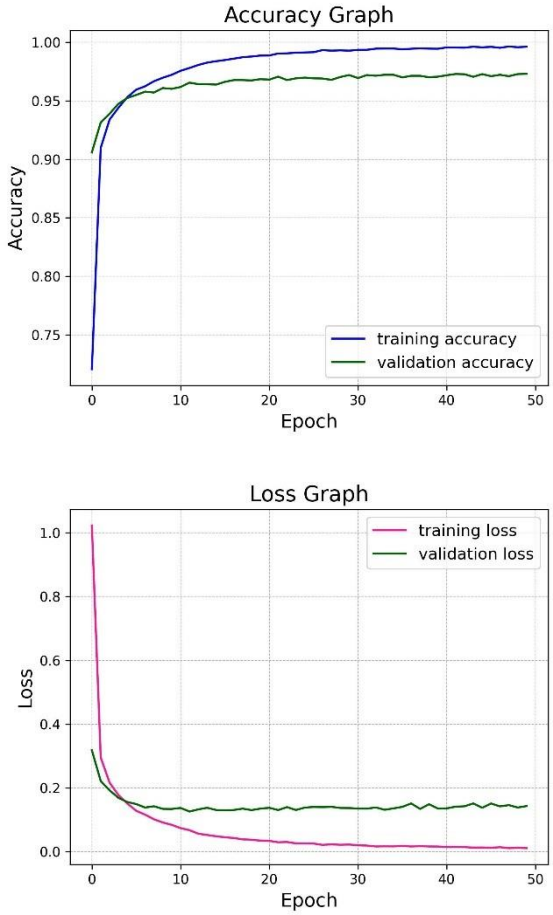
Sau khi thiết lập xong mô hình, đề án đóng băng khối trích xuất đặc trưng đã được tiền huấn luyện phần trước với tập dữ liệu được tăng cường, làm giàu, nhằm giữ nguyên những đặc trưng tốt nhất, sau đó bắt đầu quá trình tinh chỉnh khối phân loại DNN. Trong quá trình tinh chỉnh, đề án tinh chỉnh mô hình trên tập dữ liệu ảnh lá cây trên đồng, ruộng thực tế, chỉ với dữ liệu CPD. Các layer trước đó trong bộ trích xuất đặc trưng sẽ bị đóng băng. Quá trình xác nhận và thử nghiệm mô hình cuối cùng cũng sẽ được thực hiện trên tập dữ liệu ảnh lá cây ngoài thực tế cánh đồng, ruộng, tập test CPD.

3.2 Các tham số và chỉ tiêu đánh giá & kết quả thu được từ thực nghiệm

Thử nghiệm của đề án được triển khai bằng Python 3.7 với nền tảng TensorFlow và thư viện Keras cho các tác vụ Deep Learning. Các thí nghiệm được thực hiện trên một máy tính có Intel ® Core i9 10900K, GPU Nvidia® RTX A4000 và RAM 48 GB. Sau khi đào tạo, kết quả mô hình trên 3 trường hợp của dữ liệu Kết hợp được trình bày trong Hình 3.11.



Hình 3.11. Độ chính xác của mô hình trên 3 bộ Dữ liệu kết hợp



Hình 3.12. Độ chính xác và hàm loss của quá trình đào tạo và xác nhận

Model	Year	Parameters	Accuracy
EfficientNet	2021	664K	64.39%
kEffNet-B0 32ch	2022	1.08M	65.74%
InceptionResNetV2	2020	“	70.53%
Color-Aware Two-Branch	2022	5M	76.91%
MobilenetV3-Large	2022	5M	77.71%

Bảng 3.2 Kết quả đánh giá mô hình base của mobileNet V3 so với các mô hình khác dựa trên tập dữ liệu kết hợp

Kết quả ở Bảng 3.2 đã chỉ ra rằng mô hình mô hình cơ bản mobileNetV3 khi không được tinh chỉnh và làm giàu dữ liệu đạt được kết quả chính xác là 77,71%, cao

hơn các mô hình trước đó. Tuy nhiên, mức độ chính xác và hiệu suất của mô hình cơ bản MobileNetV3 vẫn chưa thực sự đáp ứng được dữ liệu thực tế. Do đó cần có một giải pháp huấn luyện hai giai đoạn, với kết quả đạt được trong bảng so sánh sau:

Model	Parameters	Accuracy	F1-Score
MobilenetV3large	5M	77.71%	0.7723
Proposed Model	3.8M	82.01%	0.8194

Bảng 3.3 Kết quả đánh giá mô hình base của mobileNet V3 so với các mô hình được cải tiến dựa trên tập dữ liệu kết hợp

Kết quả so sánh giữa mô hình được đề xuất và mô hình MobileNetV3 được thiết kế cơ bản chưa tinh chỉnh và làm giàu dữ liệu được trình bày ở bảng 3.3. Kết quả thí nghiệm cho thấy độ chính xác của mô hình là 82%, cao hơn so với các nghiên cứu trước đây. Mô hình được đề xuất trong bài viết này đã đạt được các thông số, độ chính xác và điểm F1 tốt hơn so với mô hình cơ bản. Cụ thể, độ chính xác của mô hình được đề xuất tốt hơn 5%, điểm F1-Score của mô hình này cũng cao hơn so với mô hình cơ bản.

3.3 Kết luận chương

III. KẾT LUẬN

Tổng kết lại, đề án tốt nghiệp hướng đến và đã đạt được các kết quả về việc chi tiết hóa và thực hiện xử lý dữ liệu ảnh lá cây bao gồm lá cây cà chua, nghiên cứu và ứng dụng mô hình AI thích hợp trên thiết bị IoT tại biên mạng, từ đó triển khai hệ thống thực nghiệm mẫu để đưa ra các đối sánh. Nghiên cứu này sử dụng kỹ thuật học chuyển hai bước để giảm chi phí tính toán kết hợp với phương pháp làm giàu dữ liệu bằng cách trộn tập dữ liệu, một cách tiếp cận để tăng tính đa dạng của các bộ dữ liệu và tăng cường khái quát hóa mô hình, trước khi đưa dữ liệu vào mô hình DCNN tối ưu. Đáng chú ý, kết quả đạt được với ít thông số hơn trong khi vẫn duy trì hiệu suất ổn định so với nghiên cứu trước đó. Điều này chứng tỏ rằng mô hình này sử dụng hiệu quả các nguồn lực tính toán hạn chế. Do đó, mô hình được đề xuất có thể được

triển khai trên các thiết bị biên để tối ưu hóa tính khả dụng và hiệu quả trong môi trường thực tế, đồng thời góp phần triển khai các dịch vụ nông nghiệp và điện toán biên mới.

Để tiếp tục phát triển, đề án có thể được tiếp tục thực hiện các đề xuất và hướng nghiên cứu tương lai như nghiên cứu và phát triển các giải pháp phần cứng và phần mềm có khả năng tự động hóa các quá trình thu thập dữ liệu, phân tích dữ liệu, v.v., nghiên cứu và phát triển các giải pháp IoT nhúng AI cho các loại cây trồng khác nhau, các điều kiện môi trường khác nhau. Phương pháp luận của đề án có tiềm năng cho các ứng dụng thực tế, chẳng hạn như hỗ trợ nông dân phát hiện và kiểm soát dịch bệnh, hứa hẹn mở ra những hướng nghiên cứu mới trong tương lai về thiết bị biên và cảm biến thông minh.