HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Vilayvone Phimsipasom

NGHIÊN CỨU, ỨNG DỤNG PHƯƠNG PHÁP HỌC SÂU VÀO PHÁT HIỆN ẢNH DEEPFAKE

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI – NĂM 2025

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Vilayvone Phimsipasom

NGHIÊN CỨU, ỨNG DỤNG PHƯƠNG PHÁP HỌC SÂU VÀO PHÁT HIỆN ẢNH DEEPFAKE

Chuyên ngành : Khoa học máy tính
Mã số: 8.48.01.01

ĐỀ ÁN TỐT NGHIỆP THẠC SĨ KỸ THUẬT
(Theo định hướng ứng dụng)

NGƯỜI HƯỚNG DẪN KHOA HỌC
PGS. TSKH. HOÀNG ĐĂNG HẢI

HÀ NỘI – NĂM 2025

LỜI CAM ĐOAN

Tôi tên là Vilayvone Phimsipasom, là học viên chuyên ngành Khoa học máy tính, khóa 23. Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi dưới sự hướng dẫn của PGS.TSKH. Hoàng Đăng Hải.

Các thông tin được sử dụng tham khảo trong đề án tốt nghiệp được thu thập từ các nguồn đáng tin cậy, đã được kiểm chứng, được công bố rộng rãi và được tôi trích dẫn nguồn gốc rõ ràng ở phần Tài liệu tham khảo. Các số liệu, kết quả nghiên cứu được trình bày trong đề án tốt nghiệp này là do chính tôi thực hiện một cách nghiêm túc, trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tôi xin lấy danh dự và uy tín của bản thân để đảm bảo cho lời cam đoan này.

Hà Nội, ngày 30 tháng 07 năm 2025

Người hướng dẫn	Tác giả thực hiện
(ký tên)	*(ký tên)*
PGS.TSKH. Hoàng Đăng Hải	Vilayvone Phimsipasom

MỤC LỤC

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
AI	Artificial Intelligence	Trí tuệ nhân tạo
ANN	Artificial Neural netmork	Mạng Nơ-ron nhân tạo
CNNs	Convolutional Neural Networks	Mạng nơ-ron tích chập
DL	Deep Learning	Học sâu
DNNs	Deep Neural Networks	Mạng nơ -ron nhân tạo nhiều tầng
GANs	Generative Adversarial Networks	Mạng đối kháng sinh (tạo sinh)
ML	Machine Learning	Học máy
NLP	Natural Language Processing	Xử lý ngôn ngữ tự nhiên
RNNs	Recurrent Neural Networks	Mạng nơ-ron hồi quy

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH

MỞ ĐẦU

Sự phát triển mạnh mẽ của trí tuệ nhân tạo (Artificial Intelligence – AI) và học sâu (Deep Learning – DL) đã mang lại những tiến bộ vượt bậc trong nhiều lĩnh vực khác nhau, đặc biệt là trong thị giác máy tính và xử lý hình ảnh. Một trong những ứng dụng nổi bật của học sâu là khả năng tạo ra các nội dung số chân thực đến mức khó phân biệt bằng mắt thường, điển hình là công nghệ Deepfake. Công nghệ này hỗ trợ việc tạo dựng hình ảnh và video giả mạo với độ chân thực vượt trội, cho phép thay thế khuôn mặt của một đối tượng trong video bằng khuôn mặt của đối tượng khác, đồng thời sản sinh các hình ảnh mang tính hiện thực cao dù không có sự tồn tại ngoài đời thực.

Ban đầu, Deepfake được phát triển nhằm phục vụ các mục đích tích cực như hỗ trợ trong ngành điện ảnh, sản xuất nội dung kỹ thuật số hoặc phục hồi hình ảnh lịch sử. Tuy nhiên, với sự phát triển ngày càng tinh vi của công nghệ này, Deepfake đang trở thành một công cụ có thể bị lợi dụng cho các mục đích xấu, chẳng hạn như lừa đảo, phát tán thông tin sai lệch, xâm phạm quyền riêng tư và gây ảnh hưởng đến danh dự cá nhân. Sự lan truyền nhanh chóng các nội dung Deepfake đặt ra những thách thức lớn đối với tính xác thực của thông tin, an ninh mạng, pháp lý và đạo đức xã hội.

Trước thực trạng này, việc nghiên cứu và phát triển các phương pháp phát hiện ảnh Deepfake trở thành một yêu cầu cấp thiết nhằm hạn chế tác động tiêu cực của công nghệ này đối với xã hội. Các phương pháp truyền thống như phân tích bằng mắt thường, kiểm tra siêu dữ liệu hoặc sử dụng phần mềm chỉnh sửa ảnh hiện có không còn đủ hiệu quả trước sự phát triển không ngừng của các thuật toán tạo ra ảnh Deepfake.

Hiện nay, một số nghiên cứu đã tập trung vào phát hiện Deepfake dựa trên các đặc trưng bất thường trong hình ảnh, chẳng hạn như sai lệch ánh sáng, biến dạng khuôn mặt, thiếu nhất quán trong biểu cảm hoặc dấu vết chỉnh sửa trong các chi tiết nhỏ, ví dụ [1,3, 4-7]. Mặc dù vậy, các phương pháp này vẫn đối mặt với những hạn chế đáng kể liên quan đến độ chính xác và khả năng tổng quát hóa, đặc biệt khi triển khai trên các tập dữ liệu lớn và giàu tính đa dạng. Vì vậy, cần có những phương pháp phát hiện tự động, chính xác và hiệu quả hơn, trong đó các phương pháp học sâu nổi lên như một giải pháp tiềm năng đầy hứa hẹn. Học sâu đã chứng minh được khả năng vượt trội trong việc xử

lý và phân loại hình ảnh thông qua các kiến trúc mạng nơ-ron nhân tạo tiên tiến như Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) và Transformers [11-13, 15-17]. Mô hình học sâu sở hữu khả năng tự động khai thác đặc trưng từ hình ảnh mà không yêu cầu thiết lập thủ công các đặc điểm then chốt, nhờ vậy cải thiện đáng kể độ chính xác trong việc phát hiện hình ảnh Deepfake.

Các nghiên cứu gần đây [7, 19, 62] cho thấy, việc áp dụng mạng CNNs và các phương pháp học sâu khác vào bài toán phát hiện ảnh Deepfake có thể giúp nhận diện các đặc trưng tinh vi mà mắt thường khó phân biệt, chẳng hạn như sự sai lệch trong kết cấu da, bóng đổ, ánh sáng hoặc sự bất thường trong biểu cảm khuôn mặt. Ngoài ra, các phương pháp dựa trên Transformer như Vision Transformers (ViTs) cũng đang được nghiên cứu và thử nghiệm để cải thiện khả năng tổng quát hóa của mô hình, giúp phát hiện ảnh Deepfake hiệu quả hơn trên nhiều loại dữ liệu khác nhau.

Mặc dù đã đạt được những tiến bộ nhất định, việc phát triển các mô hình học sâu nhằm phát hiện ảnh Deepfake vẫn còn nhiều thách thức. Sự phong phú và không ngừng biến đổi của các thuật toán tạo Deepfake đặt ra yêu cầu đối với mô hình phải có năng lực tổng quát hóa để phát hiện các hình ảnh giả mạo chưa từng gặp trong quá trình huấn luyện. Đồng thời, việc tối ưu hóa giữa độ chính xác và hiệu suất tính toán cũng là một yếu tố then chốt cần được quan tâm trong quá trình xây dựng mô hình.

Xuất phát từ thực tiễn trên, đề án tốt nghiệp "**NGHIÊN CỨU, ỨNG DỤNG PHƯƠNG PHÁP HỌC SÂU VÀO PHÁT HIỆN ẢNH DEEPFAKE**" được thực hiện nhằm cung cấp một cái nhìn toàn diện về công nghệ Deepfake và các phương pháp phát hiện ảnh giả mạo, qua đó đánh giá việc áp dụng một số mô hình học sâu tối ưu nhằm nâng cao hiệu quả nhận diện ảnh Deepfake. Kết quả nghiên cứu có thể được ứng dụng trong nhiều lĩnh vực như an ninh mạng, báo chí, truyền thông, pháp lý và bảo vệ danh tính số, góp phần ngăn chặn những tác động tiêu cực của ảnh Deepfake đối với xã hội.

Bố cục đề án tốt nghiệp ngoài phần mở đầu, kết luận gồm 03 chương, như sau:

Chương 1: Tổng quan nghiên cứu về học sâu và ảnh Deepfake: cơ sở lý thuyết về công nghệ AI, học sâu, đặc trưng của ảnh Deepfake do AI tạo ra, các kỹ thuật tạo ảnh Deepfake phổ biến, yêu cầu đối với việc phát hiện ảnh Deepfake, một số nghiên cứu nổi bật liên quan đến phát hiện ảnh Deepfake.

Chương 2: Giải pháp sử dụng học sâu trong phát hiện ảnh Deepfake: đặc điểm của ảnh Deepfake do AI tạo ra; các phương pháp học sâu phổ biến trong phát hiện ảnh Deepfake bao gồm kiến trúc CNNs và Transformers; mô hình học sâu cải tiến cho phát hiện ảnh Deepfake; lựa chọn giải pháp sử dụng học sâu cho cho phát hiện ảnh Deepfake.

Chương 3: Xây dựng mô hình phát hiện ảnh Deepfake: đề xuất mô hình tổng quát; thu thập dữ liệu; mô tả các công cụ, thư viện sử dụng để thực hiện mô hình; triển khai áp dụng các thuật toán học máy cho mô hình; thực hiện thử nghiệm; so sánh giữa các mô hình để đánh giá kết quả phát hiện.

CHƯƠNG 1. TỔNG QUAN NGHIÊN CỨU VỀ HỌC SÂU VÀ ẢNH DEEPFAKE

1.1 Khái quát về học sâu và công nghệ AI

1.1.1 Giới thiệu về học sâu

Học sâu (Deep learning – DL) là một nhánh chuyên biệt của học máy (Machine learning – ML). Ban đầu học sâu được lấy cảm hứng từ các mô hình sinh học về tính toán và nhận thức trong não người. Một trong những điểm mạnh nổi bật của học sâu là khả năng trích xuất các đặc trưng ở cấp độ cao từ dữ liệu đầu vào thô [16].

Sự khác biệt chính giữa học sâu và học máy truyền thống nằm ở cấu trúc của kiến trúc mạng nơ-ron nền tảng. Các mô hình học máy "không sâu" thường sử dụng mạng nơ-ron đơn giản chỉ gồm một hoặc hai lớp tính toán. Trong khi đó, các mô hình học sâu sử dụng ba lớp trở lên, thậm chí thường là hàng trăm hoặc hàng nghìn lớp để huấn luyện mô hình.

Trong khi các mô hình học có giám sát yêu cầu dữ liệu đầu vào có cấu trúc và được gán nhãn rõ ràng để tạo ra kết quả chính xác, thì các mô hình học sâu có thể khai thác phương pháp học không giám sát. Với học không giám sát, mô hình học sâu có khả năng tự động trích xuất các đặc trưng, mối quan hệ và cấu trúc cần thiết từ dữ liệu thô chưa qua xử lý để tạo ra đầu ra chính xác. Ngoài ra, những mô hình này còn có khả năng tự đánh giá và cải thiện kết quả nhằm nâng cao độ chính xác.

Sự phát triển của học sâu bắt đầu từ những nghiên cứu về mạng nơ-ron vào những năm 1950, nhưng chỉ thực sự bùng nổ trong những năm 2010 nhờ vào sự gia tăng của dữ liệu lớn (Big Data), sự tiến bộ của phần cứng như GPU và các thuật toán tối ưu hóa hiệu quả. Ngoài ra, học sâu là một khía cạnh của khoa học dữ liệu đóng vai trò then chốt trong việc thúc đẩy nhiều ứng dụng và dịch vụ tự động hóa, cho phép thực hiện các tác vụ phân tích hoặc vật lý mà không cần sự can thiệp của con người.

Học sâu đã thúc đẩy những bước tiến vượt bậc trong nhiều lĩnh vực, bao gồm xử lý ảnh, nhận dạng giọng nói, dịch máy và điều khiển xe tự hành. Các mô hình như Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) và gần đây là Transformers đã giúp máy tính đạt được khả năng học hỏi và phân tích dữ liệu

với độ chính xác cao, mở ra nhiều cơ hội ứng dụng thực tiễn trong khoa học và công nghệ. Điều này là nền tảng cho nhiều sản phẩm và dịch vụ phổ biến ngày nay như trợ lý ảo, điều khiển từ xa bằng giọng nói, hệ thống phát hiện gian lận thẻ tín dụng, xe tự hành và trí tuệ nhân tạo tự sinh.

1.1.2 Giới thiệu về trí tuệ nhân tạo

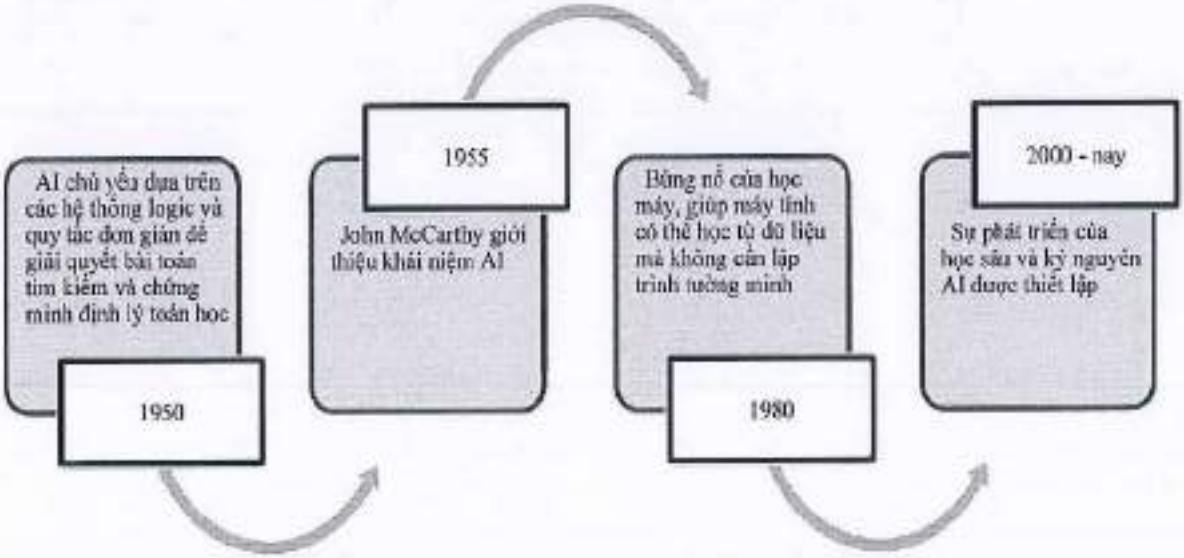
Ra đời vào giữa thế kỷ XX, trí tuệ nhân tạo (Artificial Intelligence - AI) hướng tới việc xây dựng các hệ thống máy tính có khả năng tái hiện các quá trình tư duy và hành vi đặc trưng của con người. Khái niệm AI lần đầu tiên được đề cập vào năm 1956 tại Hội nghị Dartmouth, nơi các nhà khoa học như John McCarthy, Marvin Minsky, Allen Newell và Herbert Simon đặt nền móng cho ngành nghiên cứu này. McCarthy đã đưa ra định nghĩa vào năm 1955 như sau: "*Artificial intelligence is the science and engineering of making intelligent machines, especially intelligent computer programs*", được tạm dịch "Trí tuệ nhân tạo là khoa học và kỹ thuật của việc tạo ra các cỗ máy thông minh, đặc biệt là các chương trình máy tính thông minh" [45].

Theo McCarthy, AI không đơn thuần là nghiên cứu về việc mô phỏng trí thông minh con người bằng máy tính, mà còn bao hàm quá trình phát triển các thuật toán và hệ thống có khả năng học tập, suy luận và thích nghi với môi trường. AI theo quan điểm của McCarthy không giới hạn ở việc mô phỏng con người mà còn bao gồm các cách tiếp cận khác nhau để đạt được hành vi thông minh, có thể không nhất thiết phải hoạt động giống như con người nhưng vẫn có thể giải quyết vấn đề hiệu quả.

Trí tuệ nhân tạo là một trong những lĩnh vực công nghệ có tốc độ phát triển nhanh chóng nhất hiện nay, tạo ra ảnh hưởng sâu rộng đối với nhiều ngành công nghiệp và lĩnh vực nghiên cứu khoa học. Học sâu đóng vai trò trung tâm trong việc mở rộng năng lực của AI, cho phép hệ thống máy tính đảm nhiệm các nhiệm vụ phức tạp như xử lý ngôn ngữ tự nhiên (Natural Language Processing – NLP), nhận diện hình ảnh và phân tích dữ liệu quy mô lớn. Sự phát triển mạnh mẽ của AI không chỉ đem lại những đổi mới đột phá về mặt công nghệ mà còn thúc đẩy ứng dụng vào thực tiễn, góp phần nâng cao chất lượng đời sống và tối ưu hóa quy trình vận hành trong nhiều lĩnh vực.

1.1.3 Sự phát triển của công nghệ trí tuệ nhân tạo

Trong giai đoạn sơ khai (1950 - 1970), AI chủ yếu dựa trên các hệ thống logic và quy tắc đơn giản để giải quyết bài toán tìm kiếm và chứng minh định lý toán học. Trong thập niên 1980, AI ghi nhận sự phát triển mạnh mẽ của ML, cho phép máy tính tự động học từ dữ liệu mà không đòi hỏi phải lập trình một cách tường minh các quy tắc hoạt động. Cột mốc quan trọng tiếp theo là sự phát triển của DL vào những năm 2000, nhờ vào tiến bộ của phần cứng và lượng dữ liệu khổng lồ. Các mô hình học sâu, đặc biệt là mạng nơ-ron nhân tạo nhiều tầng (Deep Neural Networks - DNNs), đã đưa AI lên một tầm cao mới với khả năng nhận diện hình ảnh, NLP và điều khiển xe tự hành. Ngày nay, AI đã trở thành công nghệ cốt lõi trong nhiều lĩnh vực như y tế, tài chính, thương mại điện tử và giáo dục, góp phần tạo ra những thay đổi mang tính đột phá trong đời sống con người.



Hình 1.1: Sự phát triển của công nghệ AI

1.2 Ứng dụng của AI, học sâu trong phát hiện ảnh giả mạo

1.2.1 Ứng dụng của AI và học sâu trong các lĩnh vực hiện nay

AI và học sâu là những lĩnh vực trọng yếu của cách mạng công nghiệp 4.0, đã và đang tác động mạnh mẽ đến hầu hết các ngành công nghiệp và đời sống xã hội. Bảng 1.1 trình bày các ứng dụng phổ biến của AI và học sâu trong đời sống và nghiên cứu hiện nay.

Bảng 1.1: Các ứng dụng của AI và học sâu

Lĩnh vực	Ứng dụng cụ thể	Công nghệ nổi bật	Nghiên cứu liên quan
Thị giác máy tính	Phân loại hình ảnh, phát hiện đối tượng, chẩn đoán hình ảnh y tế	CNN, YOLO, ResNet	[23] [14]
Xử lý ngôn ngữ tự nhiên (NLP)	Chatbot, dịch máy, phân tích cảm xúc, tóm tắt văn bản	Transformer, BERT, GPT	[20] [71]
Hệ thống đề xuất	Đề xuất phim, sản phẩm, âm nhạc cá nhân	Machine learning, collaborative	[50] [36]
Tài chính & Phát hiện gian lận	Phát hiện giao dịch bất thường, phân tích rủi ro tài chính	Anomaly detection, học sâu giám	[13] [50]
Y tế & Chăm sóc sức khỏe	Chẩn đoán bệnh, phân tích hồ sơ y tế, liệu pháp cá nhân hóa	Deep CNN, học sâu kết hợp	[46] [14]
Ô tô tự hành & Robot	Xử lý hình ảnh, định vị, ra quyết định thời gian thực	Reinforcement learning, sensor fusion	[58] [34]
Công nghiệp thông minh	Bảo trì dự đoán, kiểm tra chất lượng, tối ưu hóa sản xuất	AI công nghiệp, học sâu thời gian thực	[41] [66]
Trí tuệ nhân tạo sinh tạo (Generative AI)	Tạo văn bản, hình ảnh, video và âm thanh mới	GAN, diffusion models, GPT	[26] [55]

Lĩnh vực	Ứng dụng cụ thể	Công nghệ nổi bật	Nghiên cứu liên quan
An ninh thông tin & Pháp y số	Phát hiện ảnh/video giả mạo, Deepfake, chỉnh sửa khuôn mặt	CNN, RNN, EfficientNet, Xception, Vision Transformer	[2] [57] [21] [72]

Trong các ứng dụng được liệt kê, ứng dụng liên quan đến phát hiện nội dung đa phương tiện (ảnh/video) bị giả mạo, đặc biệt là ảnh Deepfake, là một trong những ứng dụng nổi bật và mang tính cấp thiết hiện nay.

Với sự phát triển mạnh mẽ của các mô hình AI tự sinh như GANs (Generative Adversarial Networks), việc tạo ra các hình ảnh giả mạo có độ chân thực cao ngày càng trở nên dễ dàng và phổ biến. Điều này không chỉ đặt ra thách thức nghiêm trọng đối với an ninh thông tin, đạo đức truyền thông và quyền riêng tư cá nhân, mà còn làm gia tăng nguy cơ về tin giả, thao túng dư luận và tội phạm số.

Trong bối cảnh đó, việc nghiên cứu và phát triển các hệ thống phát hiện Deepfake trở thành một yêu cầu cấp thiết nhằm bảo vệ tính xác thực và độ tin cậy của thông tin số. Đây là một trong những hướng đi mới nhưng đầy tiềm năng của thị giác máy tính ứng dụng học sâu và đang thu hút sự quan tâm đặc biệt của cộng đồng khoa học cũng như các tổ chức truyền thông, pháp lý và chính phủ.

1.2.2 AI và học sâu trong nhận diện hình ảnh và phát hiện Deepfake

Nhận diện hình ảnh được xem là một trong những lĩnh vực nghiên cứu và ứng dụng nổi bật nhất của AI và DL. Quá trình này nhằm trang bị cho máy tính khả năng diễn giải và phân loại các đối tượng xuất hiện trong hình ảnh kỹ thuật số. Với khả năng trích xuất đặc trưng không tuyến tính, mô hình hóa quan hệ không gian và học biểu diễn dữ liệu ở nhiều cấp độ trừu tượng, các mạng nơ-ron tích chập (Convolutional Neural Networks – CNNs) và các biến thể nâng cao như ResNet, EfficientNet hoặc các kiến trúc Transformer cho thị giác máy tính (Vision Transformer – ViT) đã đạt được những bước tiến vượt bậc trong nhiều bài toán thị giác máy tính. Sự kết hợp giữa AI và DL đã giúp

nâng cao đáng kể độ chính xác và hiệu suất của các hệ thống nhận diện hình ảnh, góp phần vào sự phát triển của nhiều ngành nghề khác nhau.

Các ứng dụng cụ thể bao gồm:

- Nhận diện khuôn mặt trong hệ thống giám sát an ninh và kiểm soát truy cập [19], [24].
- Phân loại bệnh lý trong ảnh chụp y học (X-quang, MRI, CT-scan) [17], [4].
- Nhận dạng vật thể trong xe tự hành và thiết bị IoT [33], [59].
- Phân loại hàng hóa trong chuỗi cung ứng và bán lẻ [69], [37].
- Nhận diện ký tự và chữ viết trong OCR (Optical Character Recognition) [28], [47].

Bên cạnh những ứng dụng trên, một trong những ứng dụng nổi bật và mang tính chất thời sự nhất của AI và DL trong nhận diện hình ảnh là *nhận diện nội dung hình ảnh và video giả mạo – thường được gọi là phát hiện Deepfake.* Đây là một vấn đề không chỉ mang tính kỹ thuật mà còn liên quan trực tiếp đến an ninh thông tin, truyền thông đại chúng và pháp y kỹ thuật số.

Deepfake là sản phẩm của các mô hình tự sinh – đặc biệt là GANs, trong đó một mô hình tự sinh học cách tạo dữ liệu giả (hình ảnh hoặc video) sao cho không thể phân biệt được với dữ liệu thật, trong khi mô hình phân biệt cố gắng phát hiện ra sự khác biệt. Quá trình này tạo ra nội dung giả mạo có độ chân thực cao, đặc biệt là khuôn mặt, biểu cảm, giọng nói [68].

Nhiều nghiên cứu đã đề xuất các phương pháp phát hiện Deepfake dựa trên học sâu như:

- **CNNs**: Khai thác đặc trưng vi mô và hiện tượng nhiễu [3].
- **Xception, EfficientNet**: Mô hình nhẹ, hiệu quả trong nhận diện giả mạo ảnh/video [15, 65].
- **Vision Transformer (ViT)**: Học đặc trưng toàn cục và không gian ảnh tốt [77].
- **Kết hợp CNN–RNN hoặc CNN–LSTM**: Phân tích chuỗi thời gian trong video Deepfake [2].
- **Học tập tổ hợp**: Kết hợp nhiều mô hình và nguồn dữ liệu để nâng cao độ chính xác [68].

Mặc dù công nghệ phát hiện Deepfake đã có những bước tiến, vẫn còn nhiều thách thức như:

- Deepfake ngày càng tinh vi, khó bị phát hiện bởi mô hình truyền thống [70, 76].
- Thiếu dữ liệu thật–giả đa dạng, đặc biệt là trong môi trường thực tế [64].
- Vấn đề tổng quát hóa giữa các tập dữ liệu khác nhau và sự thay đổi liên tục của kỹ thuật làm giả [1, 7, 57].
- Chi phí tính toán cao, đòi hỏi hạ tầng GPU mạnh để huấn luyện mô hình phức tạp [5].

1.3 Các kỹ thuật tạo ảnh Deepfake phổ biến bằng công nghệ AI

1.3.1 Giới thiệu khái quát về các kỹ thuật tạo ảnh AI

Các kỹ thuật tạo ảnh bằng AI đã trải qua một hành trình phát triển đáng kể, từ những phương pháp cơ bản đến các kỹ thuật tiên tiến như GANs. Những tiến bộ này không chỉ mở ra khả năng tạo ra hình ảnh chất lượng cao mà còn thúc đẩy sự đổi mới trong nhiều lĩnh vực khác nhau.

a) Mạng đối kháng tư sinh (GANs)

Được Ian Goodfellow và các cộng sự giới thiệu vào năm 2014 [27], GANs đã nhanh chóng trở thành một trong những công cụ mạnh mẽ nhất trong lĩnh vực tạo sinh hình ảnh bằng trí tuệ nhân tạo. Cấu trúc của GANs bao gồm hai mạng nơ-ron: mạng sinh và mạng phân biệt, hoạt động theo cơ chế đối kháng nhằm nâng cao chất lượng hình ảnh được tạo ra [27]. Cụ thể, mạng sinh có nhiệm vụ tạo ra hình ảnh giả, trong khi mạng phân biệt cố gắng phân biệt giữa hình ảnh thật và hình ảnh giả. Quá trình huấn luyện tiếp diễn cho đến khi mạng phân biệt không còn khả năng phân biệt chính xác hai loại hình ảnh này, dẫn đến việc tạo ra các hình ảnh có độ chân thực cao.

b) Mạng nơ-ron tích chập (Convolutional Neural Networks - CNNs)

CNNs đã đóng vai trò quan trọng trong việc phân tích và tạo ra hình ảnh [39]. CNNs ban đầu được thiết kế để giải quyết các bài toán nhận dạng và phân loại hình ảnh. Với khả năng trích xuất đặc trưng hiệu quả từ dữ liệu hình ảnh quy mô lớn, CNNs đã được mở rộng ứng dụng trong việc tạo sinh hình ảnh mới. Các mô hình như DeepDream của Google sử dụng CNNs để tạo ra hình ảnh mới bằng cách khuếch đại các mẫu nhận

dạng được trong quá trình huấn luyện, tạo ra những hình ảnh trừu tượng và nghệ thuật [49].

c) *Mạng tự mã hóa (Autoencoders)*

Mạng tự mã hóa là một dạng mạng nơ-ron chuyên biệt, nhằm mục tiêu học các biểu diễn mã hóa của dữ liệu đầu vào. Một trong những ứng dụng phổ biến nhất của Mạng tự mã hóa là giảm chiều dữ liệu, qua đó hỗ trợ tối ưu hóa quá trình xử lý và phân tích dữ liệu [35]. Trong ngữ cảnh tạo ảnh, mạng tự mã hóa có thể học cách tái tạo lại hình ảnh đầu vào và sau đó được sử dụng để tạo ra hình ảnh mới bằng cách lấy mẫu từ không gian mã hóa. Các biến thể như Variational Autoencoders đã được phát triển để tạo ra hình ảnh mới bằng cách học phân phối xác suất của dữ liệu đầu vào, cho phép tạo ra hình ảnh mới bằng cách lấy mẫu từ phân phối này.

d) *Mạng chuyển dịch phong cách*

Kỹ thuật chuyển dịch phong cách sử dụng mạng nơ-ron để áp dụng phong cách của một hình ảnh (ví dụ như tranh vẽ của một nghệ sĩ nổi tiếng) lên một hình ảnh khác [25]. Phương pháp này vận hành bằng cách tách riêng nội dung và phong cách của hình ảnh, sau đó tái tổ hợp chúng nhằm tạo ra hình ảnh mới. Kỹ thuật này đã mở ra những cơ hội mới trong việc sáng tạo các tác phẩm nghệ thuật số bằng cách kết hợp đa dạng các phong cách nghệ thuật.

e) *Mạng tạo hình ảnh từ văn bản (Text-to-Image Generation Networks)*

Một đột phá trong lĩnh vực này chính là khả năng tạo ra hình ảnh từ mô tả văn bản. Các mô hình như DALL·E của OpenAI sử dụng mạng nơ-ron để tạo ra hình ảnh dựa trên mô tả văn bản [56], mở ra khả năng tạo ra hình ảnh theo yêu cầu mà không cần dữ liệu hình ảnh đầu vào cụ thể.

Sự phát triển nhanh chóng của các kỹ thuật tạo sinh hình ảnh bằng AI đã mở ra nhiều cơ hội tiềm năng, song cũng đặt ra không ít thách thức. Do đó, việc nắm vững và ứng dụng hợp lý các kỹ thuật này là yếu tố then chốt để tận dụng hiệu quả các lợi ích, đồng thời giảm thiểu các nguy cơ phát sinh.

1.3.2 Giới thiệu một số công cụ tạo ảnh AI phổ biến hiện nay

Hiện nay, có nhiều công cụ tạo ảnh bằng AI được phát triển dựa trên các kỹ thuật sinh ảnh tiên tiến như GANs, mô hình Diffusion và các mô hình dựa trên Transformer

[51, 48, 63, 80, 25]. Các công cụ này giúp đơn giản hóa quá trình sáng tạo, cung cấp hình ảnh chất lượng cao và đã được áp dụng rộng rãi trong nhiều lĩnh vực. Dưới đây là một số công cụ tạo ảnh AI phổ biến và các ứng dụng thực tế của chúng.

DALL-E (OpenAI): DALL-E, do OpenAI phát triển, là một mô hình sinh ảnh dựa trên kiến trúc Transformer, cho phép tạo ra hình ảnh chi tiết với phong cách đa dạng từ các mô tả ngôn ngữ tự nhiên. Công cụ này đã chứng tỏ hiệu quả trong việc hỗ trợ sáng tạo nội dung nghệ thuật và truyền thông [51].

Midjourney: Midjourney, một nền tảng tạo sinh hình ảnh bằng AI khác, nổi bật nhờ khả năng tạo ra các tác phẩm với phong cách nghệ thuật đặc sắc và sáng tạo. Midjourney được ứng dụng rộng rãi trong các lĩnh vực như thiết kế đồ họa, minh họa và phát triển ý tưởng truyền thông [48].

Stable Diffusion (Stability AI): Là một trong những công cụ sinh ảnh tiên tiến dựa trên mô hình Diffusion, cho phép người dùng tạo ảnh từ văn bản một cách linh hoạt. Stable Diffusion có thể chạy trên các hệ thống cá nhân mà không cần kết nối đám mây, giúp người dùng linh hoạt hơn trong việc sáng tạo nội dung [63].

Artbreeder: Sử dụng GANs để cho phép người dùng pha trộn các đặc điểm của các hình ảnh khác nhau nhằm tạo ra hình ảnh mới. Công cụ này chủ yếu được sử dụng cho sáng tạo nhân vật, phong cảnh, và các tác phẩm nghệ thuật mang tính cá nhân hóa cao [80].

DeepArt: DeepArt khai thác kỹ thuật chuyển đổi phong cách nghệ thuật để tạo ra các tác phẩm hội họa mô phỏng phong cách của các danh họa nổi tiếng như Van Gogh, Monet, cùng nhiều nghệ sĩ khác. DeepArt chủ yếu phục vụ các nghệ sĩ, nhà thiết kế muốn mang lại phong cách nghệ thuật độc đáo cho ảnh [85].

1.4 Ảnh Deepfake và những vấn đề đặt ra

1.4.1 Giới thiệu về ảnh Deepfake do AI tạo ra

Ảnh Deepfake là sản phẩm của AI, đặc biệt sử dụng các kỹ thuật DL, nhằm tạo ra hoặc chỉnh sửa hình ảnh và video để chúng trông giống như thật, khiến việc phân biệt giữa thật và giả trở nên khó khăn. Thuật ngữ "Deepfake" được ghép từ hai từ *Deep Learning* và *Fake*, dùng để mô tả các nội dung video hoặc hình ảnh có độ chân thực cao được tạo ra nhờ hỗ trợ của các phương pháp DL. Thuật ngữ này bắt nguồn từ một người

dùng ẩn danh trên Reddit vào cuối năm 2017, người đã áp dụng các phương pháp học sâu để thay thế khuôn mặt của một người trong các video phản cảm bằng khuôn mặt của người khác, tạo ra các video giả có độ chân thực cao.

Deepfake chủ yếu dựa trên CNNs và mạng đối nghịch tạo sinh GANs. GANs bao gồm hai thành phần chính: một mạng tạo sinh tạo ra hình ảnh giả và một mạng phân biệt cố gắng phân biệt giữa hình ảnh thật và giả. Quá trình huấn luyện liên tục giữa hai mạng này dẫn đến việc tạo ra hình ảnh giả ngày càng chân thực. Theo nghiên cứu [70, 76], Deepfake được phân loại thành ba nhóm chính: tạo khuôn mặt, hoán đổi khuôn mặt và chỉnh sửa thuộc tính khuôn mặt.

Mặc dù Deepfake có thể được sử dụng trong các lĩnh vực như điện ảnh, truyền thông và giải trí để tạo hiệu ứng đặc biệt hoặc tái hiện hình ảnh lịch sử, nhưng nó cũng đặt ra nhiều thách thức nghiêm trọng. Sự lan rộng của Deepfake đã dẫn đến việc tạo ra nội dung xấu không đồng thuận, lừa đảo tài chính và lan truyền thông tin sai lệch, gây ảnh hưởng tiêu cực đến danh dự và quyền riêng tư của cá nhân. Một bài báo trên *Scientific Reports* đã đề cập rằng Deepfake có thể được sử dụng để tạo ra nội dung giả mạo nhằm thao túng dư luận và gây rối loạn xã hội [54].

1.4.2 Mặt trái của việc tạo ảnh bằng công nghệ AI và tác động xã hội

Bên cạnh những tiềm năng sáng tạo mạnh mẽ, công nghệ tạo ảnh bằng AI cũng kéo theo nhiều hệ quả tiêu cực, đặc biệt là trong việc tạo và lan truyền thông tin sai lệch. Với sự phát triển nhanh chóng của các mô hình sinh tạo như GANs, Diffusion và các công cụ như DALL·E, Midjourney, Stable Diffusion hay Artbreeder, việc tạo ra hình ảnh giả mạo ngày càng trở nên dễ dàng, với độ chân thực cao đến mức con người khó có thể phân biệt bằng mắt thường [68].

✓ ***Mặt trái của việc tạo ảnh bằng AI***

+ **Xâm phạm bản quyền và đạo đức sáng tạo**: Các mô hình AI có thể tái tạo phong cách nghệ thuật từ dữ liệu huấn luyện mà không có sự cho phép của tác giả, dẫn đến tranh cãi về quyền sở hữu trí tuệ và đạo đức sử dụng dữ liệu [64].

+ **Suy giảm tính xác thực của hình ảnh số**: Khi hình ảnh có thể được tạo ra mà không cần sự tồn tại vật lý, niềm tin của công chúng vào ảnh chụp và phương tiện truyền thông giảm sút nghiêm trọng [5].

+ **Tạo điều kiện cho tin giả và thao túng thông tin**: Các hình ảnh AI có thể được dùng trong mục đích tuyên truyền, thao túng chính trị, hoặc dựng bằng chứng giả trong truyền thông đại chúng và mạng xã hội [67].

✓ ***Ảnh hưởng xã hội của Deepfake***

Công nghệ Deepfake, dựa trên mô hình học sâu như GAN và autoencoder, là một trong những biểu hiện rõ rệt nhất của mặt trái trong ứng dụng AI sinh tạo. Ảnh hưởng xã hội của ảnh Deepfake bao gồm:

+ **Đe dọa an ninh cá nhân và danh tiếng**: Nhiều trường hợp Deepfake được sử dụng để giả mạo lời nói, hành vi hoặc khuôn mặt của một cá nhân nổi tiếng nhằm phá hoại uy tín hoặc tống tiền [77].

+ **Khó khăn trong pháp lý và kiểm chứng sự thật**: Deepfake làm phức tạp quá trình điều tra tội phạm số và làm yếu đi hiệu lực của bằng chứng kỹ thuật số trong pháp lý.

+ **Tác động tiêu cực đến xã hội và chính trị**: Deepfake có thể gây bất ổn xã hội bằng cách tạo ra video giả mạo các nhà lãnh đạo chính trị, kích động bạo lực, hoặc lan truyền thông tin sai lệch trong thời kỳ bầu cử.

+ **Khủng hoảng niềm tin cộng đồng**: Khi cộng đồng không thể tin vào tính xác thực của hình ảnh, điều này dẫn đến sự suy giảm niềm tin vào báo chí, khoa học và cả cơ chế dân chủ [73].

1.4.3 Yêu cầu nhận diện, phát hiện ảnh Deepfake do AI tạo ra

Ảnh Deepfake có độ chân thực cao, ranh giới giữa thật và giả rất khó phát hiện bằng mắt thường, gây ra nhiều hậu quả nghiêm trọng đối với an ninh thông tin và niềm tin xã hội. Do đó, việc nhận dạng và phát hiện ảnh Deepfake do AI tạo ra là một thách thức ngày càng lớn trong lĩnh vực công nghệ thông tin, đòi hỏi sự kết hợp giữa các phương pháp kỹ thuật tiên tiến, các mô hình học sâu và cơ chế pháp lý phù hợp.

Việc phát hiện Deepfake ngày càng khó khăn hơn còn do sự cải tiến liên tục của công nghệ tạo ảnh giả mạo, đòi hỏi các hệ thống nhận diện phải có khả năng phát hiện những bất thường tinh vi trong kết cấu hình ảnh, ánh sáng và chuyển động [30]. Một trong những phương pháp phổ biến nhất hiện nay là sử dụng CNNs để trích xuất các đặc trưng của hình ảnh và phân loại chúng dựa trên mức độ chân thực. Tuy nhiên, các mô

hình CNNs truyền thống gặp khó khăn trong việc tổng quát hóa đối với các kỹ thuật Deepfake mới, khiến chúng dễ bị qua mặt bởi các thuật toán giả mạo tiên tiến hơn.

Để tăng cường hiệu quả khả năng nhận dạng, phát hiện ảnh Deepfake, các hệ thống nhận diện và phát hiện ảnh Deepfake cần tập trung nghiên cứu vào các vấn đề:

✓ *Độ chính xác và độ nhạy cao*

Hệ thống cần có khả năng phát hiện chính xác các dấu hiệu giả mạo tinh vi, kể cả trong trường hợp hình ảnh có độ phân giải cao, ánh sáng ổn định và không có hiện tượng méo đặc trưng. Đặc biệt, thuật toán phải đảm bảo độ nhạy cao để tránh bỏ sót các nội dung nguy hiểm.

✓ *Khả năng tổng quát hóa*

Mô hình phát hiện phải tổng quát hóa tốt trên các nguồn dữ liệu khác nhau và các kỹ thuật Deepfake mới mà nó chưa được huấn luyện trước. Yêu cầu này đặc biệt quan trọng do các kỹ thuật Deepfake liên tục được cải tiến và biến hóa phức tạp theo thời gian.

✓ *Tính thời gian thực và khả năng triển khai*

Trong bối cảnh lan truyền thông tin nhanh chóng trên mạng xã hội, hệ thống phát hiện cần hoạt động trong thời gian thực hoặc gần thời gian thực để kịp thời ngăn chặn ảnh hưởng tiêu cực. Mô hình cần triển khai hiệu quả trên thiết bị biên hoặc hệ thống phân tán với tài nguyên tính toán hạn chế [77].

✓ *Khả năng giải thích và minh bạch*

Để phục vụ các mục đích pháp lý, báo chí và xã hội, mô hình nhận diện, phát hiện ảnh Deepfake cần cung cấp các chỉ báo rõ ràng và khả năng giải thích tại sao ảnh bị gán nhãn giả mạo. Điều này giúp tăng độ tin cậy và tính chấp nhận trong thực tiễn [73].

✓ *Đảm bảo quyền riêng tư và đạo đức*

Hệ thống phát hiện cần đảm bảo không xâm phạm quyền riêng tư, đặc biệt trong các ứng dụng giám sát hoặc truyền thông đại chúng. Ngoài ra, các công cụ AI cần được phát triển và sử dụng theo hướng có trách nhiệm và minh bạch, tuân thủ quy định pháp luật liên quan đến nội dung số [64].

Tuy nhiên, một trong những thách thức lớn nhất trong việc phát hiện ảnh Deepfake là sự thiếu hụt dữ liệu huấn luyện chất lượng cao, do các mẫu Deepfake liên tục thay đổi và trở nên ngày càng tinh vi hơn. Do đó, việc xây dựng các tập dữ liệu lớn và đa dạng

về Deepfake để huấn luyện mô hình phát hiện là một yêu cầu cấp thiết, giúp tăng cường khả năng nhận diện và phân biệt giữa nội dung thật và giả.

1.5 Một số nghiên cứu nổi bật liên quan đến phát hiện ảnh Deepfake

Công nghệ Deepfake, sử dụng AI để tạo ra hình ảnh giả mạo với độ chân thực cao, đã đặt ra những thách thức lớn về an ninh và đạo đức. Để giải quyết vấn đề này, các nghiên cứu gần đây đã chú trọng phát triển các phương pháp phát hiện Deepfake hiệu quả. Trong đó, một hướng tiếp cận phổ biến là khai thác học sâu nhằm phân tích và nhận diện các đặc trưng bất thường xuất hiện trong hình ảnh. Theo nghiên cứu của Sm Zobaed và cộng sự (2021), các phương pháp phát hiện dựa trên học sâu cho thấy hiệu suất cao trong việc phân biệt hình ảnh giả mạo bằng cách học từ dữ liệu lớn và sử dụng các mạng CNNs để nhận diện các dấu vết giả tạo [81].

Luca Guarnera và cộng sự (2020) đã đề xuất một phương pháp dựa trên việc khai thác dấu vết chập để phát hiện Deepfake. Phương pháp này sử dụng thuật toán Expectation-Maximization để trích xuất các dấu vết do mạng GANs để lại trong quá trình tạo hình ảnh, đạt độ chính xác trên 98% trong việc phân loại hình ảnh thật và giả [31].

Một nghiên cứu của nhóm tác giả năm 2023 [52] đề xuất một kiến trúc mạng nơ-ron tích chập sâu (deep-CNN hay D-CNN) mới và cải tiến nhằm phục vụ cho nhiệm vụ phát hiện ảnh deepfake với độ chính xác hợp lý và khả năng tổng quát hóa cao. Trong đó có sử dụng hàm mất mát nhị phân cùng với thuật toán tối ưu hóa Adam được sử dụng nhằm cải thiện tốc độ học của mô hình D-CNN. Ngoài ra, nghiên cứu thử nghiệm bảy bộ dữ liệu khác nhau thuộc thách thức tái tạo, với 5.000 ảnh Deepfake và 10.000 ảnh thật. Mô hình được đề xuất đạt độ chính xác 98,33% trên bộ dữ liệu AttGAN (*Facial Attribute Editing by Only Changing What You Want*), 99,33% trên GDWCT (*Group-wise Deep Whitening-and-Coloring Transformation*), 95,33% trên StyleGAN, 94,67% trên StyleGAN2 và 99,17% trên StarGAN (*A GAN capable of learning mappings among multiple domains*), cho cả ảnh thật và ảnh Deepfake, cho thấy tính khả thi cao trong các thiết lập thực nghiệm [52].

Cũng trong năm 2023, các tác giả [7] đã đề xuất sử dụng các mô hình CNN như ResNeXt và Xception trong một kiến trúc lai, kết hợp với các kỹ thuật tiền xử lý dữ liệu

và điều chỉnh siêu tham số, nhằm cải thiện độ chính xác và khả năng tổng quát hóa của mô hình phát hiện [7].

Zhongjie Ba và cộng sự (2024) [8] đã đề xuất một khung làm việc mới nhằm phát hiện Deepfake bằng cách trích xuất nhiều đại diện cục bộ không chồng chéo và kết hợp chúng thành một đặc trưng toàn cục giàu ngữ nghĩa. Cách tiếp cận này giúp cải thiện đáng kể độ chính xác và khả năng tổng quát hóa trong các tình huống thực tế [8].

Nghiên cứu [40] đã phân loại 16 mô hình phát hiện Deepfake theo 4 nhóm chính và 13 nhóm con, đánh giá khả năng tổng quát hóa của chúng trong các kịch bản tấn công khác nhau. Nghiên cứu nhấn mạnh rằng nhiều mô hình hiện tại phụ thuộc quá nhiều vào dữ liệu huấn luyện trong phòng thí nghiệm, dẫn đến hiệu suất kém khi đối mặt với các Deepfake thực tế [40].

Gần đây, nhóm nghiên cứu do Nuria Alina Chandra dẫn đầu đã giới thiệu *Deepfake-Eval-2024*, một bộ dữ liệu benchmark đa phương tiện (video, audio, hình ảnh) thu thập từ mạng xã hội và nền tảng phát hiện Deepfake TrueMedia.org. Bộ dữ liệu này phản ánh các kỹ thuật Deepfake mới nhất và cho thấy hiệu suất của các mô hình phát hiện hiện tại giảm đáng kể khi áp dụng vào dữ liệu thực tế, với AUC giảm 45–50% so với các bộ dữ liệu học thuật trước đó [12].

Tại Việt Nam, việc nghiên cứu và phát hiện ảnh Deepfake đang trở thành một lĩnh vực quan trọng trong bối cảnh công nghệ AI ngày càng phát triển và được sử dụng trong các hành vi lừa đảo tinh vi.

Faceless – Công cụ phát hiện Deepfake "Make in Vietnam" được nghiên cứu và phát triển bởi hai tác giả là Dương Tiểu Đồng (sinh năm 2005) và Phạm Tiến Mạnh (sinh năm 1996) [82]. Đây là một giải pháp nội địa, hoạt động dựa trên cơ chế tiếp nhận hình ảnh đầu vào từ người dùng, sau đó tiến hành nhận diện khuôn mặt và phân tích thông qua CNNs nhằm xác định xem khuôn mặt đó có phải là sản phẩm của công nghệ Deepfake hay không. Kết quả thử nghiệm cho thấy Faceless có thể phát hiện Deepfake với độ chính xác đạt 94,5% chỉ trong thời gian dưới 2 giây – một thành tích nổi bật so với nhiều công cụ hiện có cùng mục tiêu. Hiện tại, công cụ này đã được phát hành dưới dạng mã nguồn mở, với định hướng phát triển tiếp theo dựa vào sự đóng góp từ cộng đồng và không nhằm mục đích thương mại hóa [82].

Một nghiên cứu khác tại Việt Nam [38] đã áp dụng các mô hình học sâu như XceptionNet và ResNet101 để phát hiện Deepfake. Phương pháp tiếp cận này sử dụng học chuyển giao để tận dụng các mô hình đã được huấn luyện trên những tập dữ liệu quy mô lớn, góp phần cải thiện hiệu quả trong phát hiện hình ảnh và video giả mạo.

Tại Lào, các hoạt động nghiên cứu và phát triển công nghệ phát hiện Deepfake vẫn đang trong giai đoạn khởi đầu. Vào tháng 4 năm 2025, công ty X-PHY Inc đã ra mắt công cụ phát hiện Deepfake theo thời gian thực, cho phép người dùng xác minh tính xác thực của video, âm thanh và hình ảnh trực tiếp trên thiết bị mà không cần kết nối Internet. Công cụ này sử dụng trí tuệ nhân tạo để phân tích các biểu cảm khuôn mặt, dấu vân tay giọng nói và các dấu hiệu do mạng GANs tạo ra, đạt độ chính xác lên đến 90% trong việc phát hiện nội dung giả mạo. Công nghệ này đã được giới thiệu tại Hội nghị RSA 2025 [83].

Có thể thấy rằng, các nghiên cứu này không chỉ cung cấp các giải pháp kỹ thuật hiệu quả trong việc phát hiện nội dung giả mạo mà còn giúp nâng cao nhận thức về những nguy cơ tiềm ẩn của Deepfake, đồng thời thúc đẩy sự phát triển của các công cụ kiểm chứng thông tin nhằm bảo vệ tính toàn vẹn của dữ liệu trong môi trường số hóa hiện nay.

1.6 Kết luận chương

Trong chương 1, đề án tốt nghiệp đã trình bày tổng quan nền tảng về AI, học máy và đặc biệt là học sâu – lĩnh vực đóng vai trò cốt lõi trong các hệ thống xử lý và phân tích hình ảnh hiện đại. Học sâu, với các kiến trúc mạng nơ-ron sâu như CNNs, Transformer hay Diffusion, đã tạo nên bước ngoặt trong khả năng sinh và phân tích dữ liệu hình ảnh.

Bên cạnh những thành tựu vượt trội, học sâu cũng mở ra nhiều thách thức mới, điển hình là sự xuất hiện của ảnh Deepfake – sản phẩm của các mô hình sinh tạo có khả năng giả mạo khuôn mặt, biểu cảm và danh tính một cách tinh vi. Các ứng dụng Deepfake đang đặt ra những nguy cơ nghiêm trọng về đạo đức, an ninh thông tin, và độ tin cậy của truyền thông số.

Trong bối cảnh đó, việc phát triển các hệ thống có khả năng phát hiện ảnh Deepfake một cách chính xác và đáng tin cậy đang trở thành một hướng nghiên cứu cấp thiết.

Những nội dung trình bày trong chương này là cơ sở lý luận quan trọng để từ đó đi sâu vào phân tích các phương pháp nhận diện ảnh Deepfake và đề xuất mô hình phù hợp trong bài toán nghiên cứu đặt ra ban đầu.

CHƯƠNG 2. GIẢI PHÁP SỬ DỤNG HỌC SÂU TRONG PHÁT HIỆN ẢNH DEEPFAKE

2.1 Đặc điểm của ảnh Deepfake

2.1.1 Dấu hiệu đặc trưng trong ảnh Deepfake

Mặc dù công nghệ Deepfake ngày càng trở nên tinh vi, các hình ảnh giả mạo do mô hình sinh tạo như GANs, StyleGAN hoặc Diffusion vẫn thường để lại một số dấu hiệu đặc trưng mà các hệ thống phát hiện có thể khai thác. Những dấu hiệu này thường xuất hiện ở cấp độ hình học, thống kê hoặc tần số.

✓ ***Bất thường về chi tiết khuôn mặt***

Các mô hình sinh ảnh thường gặp khó khăn trong việc tái tạo chính xác các chi tiết vi mô trên khuôn mặt người thật. Các biểu hiện phổ biến bao gồm đồng tử không đối xứng, ánh sáng trong mắt không phù hợp với nguồn sáng, biến dạng các bộ phận như tai hoặc răng [2] [73].

✓ ***Biến dạng về hình học và tỷ lệ***

Tỷ lệ giữa các bộ phận khuôn mặt như mắt – mũi – miệng có thể bị sai lệch, đặc biệt khi khuôn mặt quay nghiêng hoặc biểu cảm mạnh. Một số hình ảnh deepfake cho thấy cấu trúc xương mặt bị bóp méo không tự nhiên [73].

✓ ***Vết nhiễu trong miền tần số hoặc khi nén***

Ảnh Deepfake thường chứa vết nhiễu đặc trưng trong miền tần số, phát hiện được thông qua các kỹ thuật Fourier hoặc DCT. Các mô hình sinh ảnh chưa tái hiện hoàn hảo phân bố nhiễu tự nhiên như ảnh chụp từ cảm biến thật [75].

✓ ***Mất đồng nhất giữa khuôn mặt và các vùng xung quanh***

Do quá trình tổng hợp thường tập trung vào vùng mặt, các vùng như cổ, tóc hoặc nền phía sau có thể mờ, thiếu chi tiết, hoặc khác biệt về ánh sáng và màu sắc. Viền khuôn mặt đôi khi bị nhòe, gãy khúc hoặc không hòa trộn tốt với nền [2].

✓ ***Thiếu chuyển động vi mô tự nhiên***

Ảnh Deepfake – đặc biệt là ảnh được cắt từ video – thường thiếu các yếu tố phi ngôn ngữ như phản xạ đồng tử, vi chuyển động mắt hoặc căng cơ mặt tự nhiên. Trong ảnh tĩnh, điều này thể hiện qua ánh nhìn trống rỗng hoặc biểu cảm cứng nhắc [73].

2.1.2 Nhận biết các đặc trưng của ảnh Deepfake

Việc phát hiện các đặc trưng của ảnh Deepfake đòi hỏi áp dụng các kỹ thuật xử lý ảnh, học sâu và phân tích thống kê ở nhiều cấp độ.

✓ ***Phân tích hình học và tỷ lệ khuôn mặt***

Kỹ thuật này nhằm phát hiện khuôn mặt và điểm đặc trưng. Một số công cụ được sử dụng như: MTCNN, dlib, Mediapipe,... Ngoài ra, có thể sử dụng kỹ thuật so sánh tỷ lệ khoảng cách giữa các điểm đặc trưng (mắt – mũi – miệng – cằm), đối xứng khuôn mặt, từ đó nhận diện được sự biến dạng hoặc méo mó khuôn mặt không phù hợp với giải phẫu học tự nhiên.

✓ ***Phân tích nhiễu trong miền tần số***

Việc phân tích nhiễn trong miền tần số có thể được phát hiện bằng việc sử dụng biến đổi Fourier (FFT), DCT, phân tích gradient hoặc ảnh tín hiệu dư. Ngoài ra, cũng có thể trích xuất thông tin miền tần số cao, phát hiện mẫu nhận dạng gây nhiễu không tự nhiên từ mô hình GANs [75].

✓ ***Trích xuất đặc trưng bằng học sâu***

Một số mô hình phổ biến có thể trích xuất đặc trưng bao gồm XceptionNet, EfficientNet, MesoNet, Vision Transformer,...trên bộ dữ liệu FaceForensics++, Celeb-DF, DFFD. Qua đó có thể trích xuất đặc trưng không gian/tần số phức tạp mà phương pháp truyền thống khó phát hiện [2].

✓ ***Phân tích sự đồng bộ giữa khuôn mặt và vùng nền***

Sử dụng các công cụ như phân đoạn ngữ nghĩa, kiểm tra tính nhất quán giữa khuôn mặt và nền ảnh bằng việc so sánh ánh sáng, độ sắc nét, màu sắc giữa vùng mặt và cổ/tóc/nền. Từ đó có thể phát hiện vùng chắp vá, viền bị mờ hoặc ánh sáng lệch pha do không khớp giữa vùng sinh tổng hợp và vùng gốc.

✓ ***Phân tích biểu cảm và chi tiết vi mô***

Phát hiện vi chuyển động, phân tích chuyển động mắt và vùng cơ mặt. Sử dụng các kỹ thuật chuyển động quang học, mô hình hóa không gian – thời gian (khi xử lý chuỗi ảnh từ video), LBP (Local Binary Patterns) cho ảnh tĩnh. Ứng dụng cho phát hiện ảnh Deepfake tĩnh có ánh nhìn "vô hồn", thiếu tương tác thần kinh tinh vi.

2.2 Các phương pháp học sâu ứng dụng trong phát hiện ảnh Deepfake

2.2.1 Một số phương pháp ML truyền thống trong nhận dạng ảnh Deepfake

Các thuật toán ML truyền thống cũng đóng vai trò quan trọng trong việc phát hiện ảnh sinh bởi AI, đặc biệt khi cần tiết kiệm tài nguyên tính toán:

✓ **Hồi quy logistic (Logistic Regression)**: Đây là phương pháp cơ bản để phân loại ảnh thật và giả dựa trên một số đặc điểm đã được trích xuất trước, chẳng hạn như độ sắc nét, độ sáng hoặc các đặc điểm phân biệt rõ ràng khác [1].

✓ **Máy vector hỗ trợ (Support Vector Machines - SVM)**: SVM giúp phân loại ảnh với độ chính xác cao, đặc biệt khi áp dụng cho các đặc điểm đã qua tiền xử lý và trích xuất. SVM phù hợp với các tập dữ liệu nhỏ, khi mà sử dụng CNNs không khả thi [62].

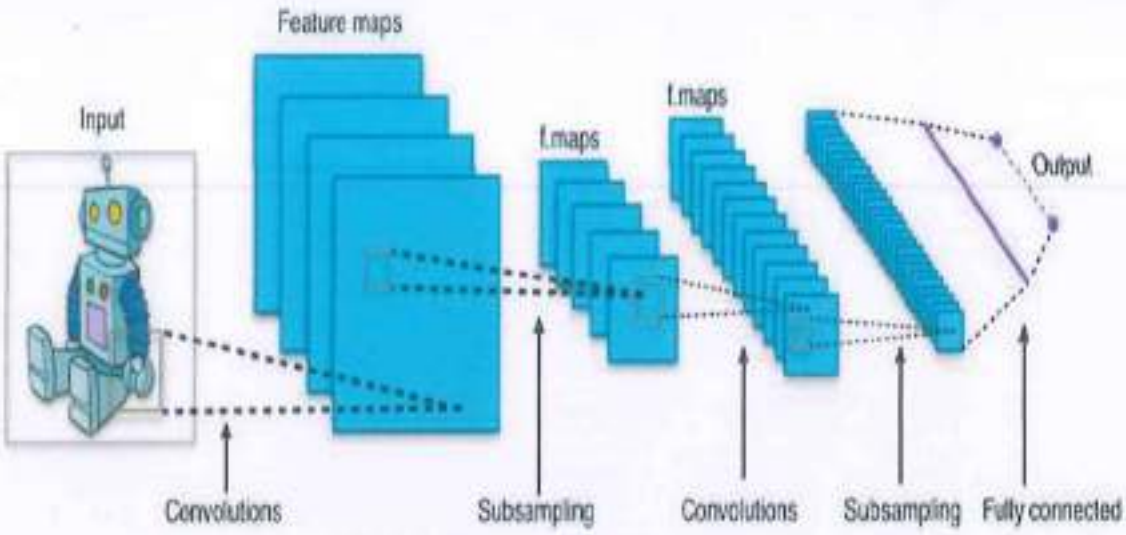
✓ **K-Nearest Neighbors (KNN)**: KNN là phương pháp dựa trên khoảng cách giữa các mẫu ảnh, giúp phân biệt ảnh thật và giả dựa trên các đặc điểm đặc trưng đã được tiền xử lý. Mặc dù KNN không phù hợp với các tập dữ liệu lớn, nó có thể sử dụng hiệu quả trong các trường hợp yêu cầu phân tích nhanh và đơn giản.

✓ **Cây quyết định và Rừng ngẫu nhiên:** Phương pháp này có thể sử dụng để phát hiện các đặc điểm bất thường trong ảnh. Rừng ngẫu nhiên giúp tăng cường tính tổng quát và giảm thiểu hiện tượng quá khớp, phù hợp với các tập dữ liệu đa dạng.

2.2.2 Kiến trúc CNNs và các biến thể

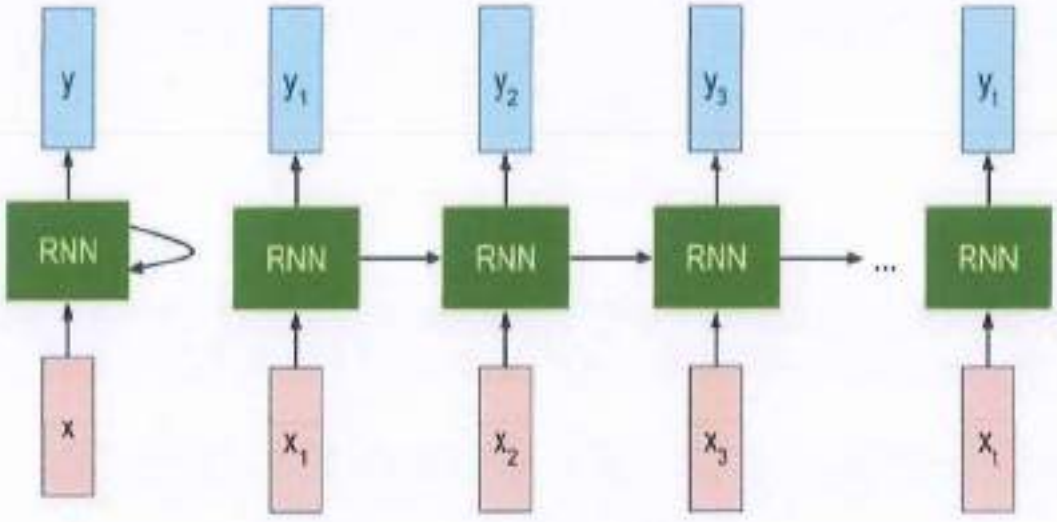
Học sâu là phương pháp phổ biến trong việc phát hiện ảnh sinh bởi AI nhờ khả năng xử lý hình ảnh và học các đặc điểm phức tạp từ dữ liệu:

✓ **Mạng nơ-ron tích chập (CNNs)**: CNNs là một trong những kiến trúc mạng học sâu được sử dụng nhiều nhất để phát hiện ảnh giả. CNNs hoạt động dựa trên việc phát hiện các đặc điểm ở nhiều lớp khác nhau, từ các đặc điểm bề mặt như màu sắc, cạnh, đến các đặc điểm sâu hơn như kết cấu và hình thái. Các mô hình CNNs tiêu biểu cho bài toán nhận diện ảnh sinh bởi AI bao gồm ResNet, EfficientNet và VGG. Những mô hình này có khả năng nhận diện và phân loại tốt các đặc điểm bất thường trong ảnh do AI sinh.



Hình 2.1: Mô hình CNNs

✓ **Mạng nơ-ron xoắn tích hợp (Recurrent Neural Networks - RNNs)**: Mặc dù RNNs chủ yếu dùng cho dữ liệu tuần tự, khi kết hợp với CNNs trong các mô hình như CRNNs (Convolutional Recurrent Neural Networks), nó cũng có thể giúp phân tích các đặc điểm không gian - phụ thuộc, từ đó tăng độ chính xác cho việc phát hiện các bất thường trong ảnh.



Hình 2.2: Mô hình RNNs

2.2.3 Các mô hình CNNs trong phát hiện ảnh Deepfake

2.2.3.1 Mô hình XceptionNet

XceptionNet được François Chollet đưa ra năm 2017 như một biến thể của Inception v3. XceptionNet thay thế các khối Inception bằng tích chập phân chia theo

chiều sâu để tăng hiệu quả tính toán [15]. Mô hình này phù hợp cho các hệ thống phát hiện Deepfake yêu cầu độ chính xác cao, sử dụng máy chủ GPU hoặc đám mây, được sử dụng làm phần lõi của FaceForensics++ [57].

Đặc điểm: Sâu (~36 lớp), sử dụng tích chập phân chia theo chiều sâu giúp giảm số lượng tham số mà vẫn giữ được hiệu năng cao.

Nguyên lý hoạt động: Áp dụng tách biệt hai bước: tích chập theo từng kênh và kết hợp kênh để trích xuất đặc trưng và học phân loại giả/thật từ ảnh khuôn mặt.

✓ **Ưu điểm:**

- Độ chính xác cao.
- Hiệu quả trong phát hiện giả mạo tinh vi.

✓ **Hạn chế:**

- Mô hình lớn, tiêu tốn tài nguyên.
- Kém phù hợp cho thiết bị nhúng/di động.

2.2.3.2 Mô hình MesoNet

MesoNet được đề xuất bởi Afchar và các cộng sự (2018) nhằm cung cấp giải pháp nhẹ cho phát hiện giả mạo mặt trong video và ảnh [2], thích hợp cho các thiết bị có hạn chế về tài nguyên như điện thoại, máy ảnh thông minh, có khả năng thời gian thực.

Đặc điểm: Nông (~4 khối conv), gồm hai biến thể chính là Meso - 4 và MesoInception-4 với tổng tham số nhỏ.

Nguyên lý hoạt động: Khai thác các đặc trưng ở mức trung gian như mép, vùng da, ánh sáng không đồng nhất để phân biệt ảnh thật/giả.

✓ **Ưu điểm:**

- Gọn nhẹ, nhanh.
- Dễ triển khai trên thiết bị hạn chế tài nguyên.
- Tốc độ dự đoán nhanh.

✓ **Hạn chế:**

- Độ chính xác thấp hơn các mô hình sâu.
- Kém hiệu quả với Deepfake tinh vi.

2.2.3.3 Mô hình ResNet-50

ResNet (Residual Network) do nhóm tác giả He phát triển năm 2015, sử dụng cơ chế bỏ qua các kết nối để huấn luyện mạng sâu hiệu quả [32]. ResNet phù hợp cho phát hiện ảnh giả mạo trong môi trường có GPU, dễ dàng tinh chỉnh lại cho từng bộ dữ liệu mới.

Đặc điểm: Sâu (50 lớp) với các khối dư gồm conv 1×1, 3×3, 1×1; dễ mở rộng và dùng làm phần lõi cho mô hình chú ý hoặc lai.

Nguyên lý hoạt động: Sử dụng khối dư và kết nối tắt xuyên qua nhiều lớp, giúp mạng tránh suy giảm gradient và giúp học đặc trưng tốt hơn.

✓ **Ưu điểm:**

- Ổn định, dễ tinh chỉnh.
- Dễ tích hợp vào pipeline hiện có.
- Hiệu quả với các ảnh Deepfake tổng hợp đơn giản.

✓ **Hạn chế:**

- Tốc độ chậm hơn so với EfficientNet.
- Số lượng tham số lớn hơn mô hình nhẹ khác.

2.2.3.4 Mô hình EfficientNet (B0–B7)

Được phát triển bởi Tan và Le tại Google AI (2019), mô hình sử dụng compound scaling để mở rộng mạng theo chiều sâu, chiều rộng và độ phân giải [65]. Thích hợp cho cả môi trường tài nguyên hạn chế (B0–B2) và server mạnh (B5–B7). Hiệu quả tốt trên các ảnh độ phân giải cao.

Đặc điểm: Dựa trên kiến trúc MobileNetV2 với các khối tích chập nhẹ và các khối nén – kích hoạt.

Nguyên lý hoạt động: Kết hợp chuẩn hóa ba chiều và tối ưu hóa hiệu suất thông qua khối tích chập nhẹ, tăng khả năng nhận diện đặc trưng mà không tăng đáng kể chi phí tính toán.

✓ **Ưu điểm:**

- Tối ưu tốt giữa hiệu năng và độ chính xác.
- Nhiều phiên bản phù hợp nhiều cấu hình hệ thống.
- Giảm chi phí dự đoán mà vẫn duy trì độ chính xác.

✓ **Hạn chế:**

- Thiết kế phức tạp hơn ResNet.
- Cần tinh chỉnh kỹ kỹ nếu tinh chỉnh trên tập dữ liệu nhỏ.

Bảng 2.1: So sánh một số mô hình CNNs trong phát hiện ảnh Deepfake

Tiêu chí	XceptionNet	MesoNet	ResNet-50	EfficientNet-B0
Số lớp	~36	~4	50	~82
Tham số	~22M	<1M	~25M	~5M
Độ chính xác (trung bình)	★★★★★	★★★	★★★★	★★★★
Tốc độ xử lý	Trung bình	Rất nhanh	Chậm	Nhanh
Yêu cầu tài nguyên	Cao	Thấp	Trung bình	Thấp – Trung bình
Khả năng triển khai di động	Kém	Tốt	Kém	Tốt
Ưu điểm nổi bật	Độ chính xác cao	Gọn nhẹ, realtime	Dễ tinh chỉnh	Hiệu năng cân bằng
Hạn chế chính	Nặng, không realtime	Độ chính xác chưa cao	Cồng kềnh	Cần tinh chỉnh phức tạp

2.2.4 Các mô hình Transformers trong phát hiện ảnh Deepfake

2.2.4.1 Mô hình Vision Transformer (ViT)

ViT được đề xuất bởi Dosovitskiy và cộng sự năm 2020 [22], là mô hình đầu tiên áp dụng kiến trúc Transformer thuần túy (không dùng CNNs) vào thị giác máy tính, đặc biệt là phân loại ảnh [22]. Mô hình này hiệu quả khi huấn luyện trên tập dữ liệu lớn (ví dụ: JFT-300M), được sử dụng trong phát hiện Deepfake để khai thác quan hệ toàn cục trong ảnh khuôn mặt.

Đặc điểm: Chia ảnh đầu vào thành các vùng nhỏ (ví dụ 16×16), ánh xạ mỗi vùng thành véc - tơ, thêm embedding vị trí, rồi đưa vào khối Transformer encoder.

Nguyên lý hoạt động: Sử dụng cơ chế chú ý để tính toán mối quan hệ giữa tất cả các vùng ảnh, cho phép học đặc trưng không cục bộ, phù hợp với các dạng giả mạo tinh vi phân bố khắp khuôn mặt.

✓ **Ưu điểm:**

- Mạnh trong học đặc trưng toàn cục.
- Có thể thay thế CNNs hoàn toàn trong nhiều nhiệm vụ.
- Dễ mở rộng quy mô mô hình.

✓ **Hạn chế:**

- Cần dữ liệu rất lớn để huấn luyện hiệu quả.
- Tốn tài nguyên và thời gian huấn luyện.
- Không tốt với ảnh độ phân giải nhỏ hoặc bộ dữ liệu hạn chế.

2.2.4.2 Mô hình Swin Transformer

Mô hình được nhóm tác giả Liu đề xuất năm 2021 nhằm khắc phục điểm yếu về chi phí tính toán của ViT với ảnh lớn, đồng thời giữ tính cục bộ và toàn cục [44]. Mô hình này thích hợp với ảnh có độ phân giải cao và bài toán cần cả đặc trưng cục bộ lẫn toàn cục, điển hình như Deepfake có chi tiết giả mạo tinh vi.

Đặc điểm: Áp dụng cơ chế tự chú ý trong các vùng cửa sổ trượt để học đặc trưng trong khu vực nhỏ trước, sau đó mở rộng ra toàn ảnh.

Nguyên lý hoạt động: Học đặc trưng tuần tự từ cục bộ đến toàn cục bằng cách xếp lớp chú ý theo vùng và dịch chuyển cửa sổ giữa các lớp để đảm bảo liên kết toàn cục.

✓ **Ưu điểm:**

- Hiệu suất tốt hơn ViT trên ảnh độ phân giải cao.
- Tối ưu hơn về chi phí tính toán.
- Dễ kết hợp với các mô hình phân loại, phân vùng, phát hiện.

✓ **Hạn chế:**

- Kiến trúc phức tạp hơn.
- Cần tinh chỉnh tốt để đạt hiệu năng tối đa.
- Chưa có nhiều ứng dụng thực tế với Deepfake như CNNs.

2.2.4.3 Mô hình TimeSformer (Time-Space Transformer)

Mô hình được nhóm của Bertasius đưa ra năm 2021. TimeSformer là biến thể của ViT dành riêng cho video, đặc biệt phù hợp cho phát hiện Deepfake trong video [9]. TimeSformer áp dụng cho dữ liệu là video chứ không phải ảnh tĩnh, ví dụ video Deepfake từ TikTok, YouTube.

Đặc điểm: Mô hình tách riêng attention theo không gian và thời gian, giúp giảm chi phí tính toán so với chú ý toàn phần.

Nguyên lý hoạt động: Trích xuất đặc trưng từ từng khung hình bằng ViT, sau đó học mối quan hệ giữa các khung hình theo thời gian để phát hiện sự không liên tục, giật, hay bất thường về biểu cảm.

✓ **Ưu điểm:**

- Hiệu quả với video dài, biểu cảm giả.
- Giảm chi phí so với chú ý toàn bộ trong không gian–thời gian.
- Có thể học mối quan hệ ngữ nghĩa theo thời gian.

✓ **Hạn chế:**

- Mô hình rất nặng, khó triển khai thời gian thực.
- Cần tiền huấn luyện trên các tập video lớn.
- Cần đồng bộ hóa và phân tích khung chính xác.

2.2.4.4 Mô hình DETR (Detection Transformer)

DETR được Facebook AI giới thiệu năm 2020, là sự kết hợp giữa kiến trúc lõi của CNNs và kiến trúc Transformer cho nhiệm vụ phát hiện đối tượng, song được điều chỉnh cho phát hiện vùng ảnh giả mạo Deepfake [11]. DETR có thể xác định vị trí vùng bị làm giả, phù hợp cho các ứng dụng giám định hoặc pháp lý.

Đặc điểm: Gồm hai phần chính: kiến trúc lõi CNNs (thường là ResNet-50) và Transformer giải mã để dự đoán hộp giới hạn.

Nguyên lý hoạt động: Trích xuất đặc trưng bằng CNNs, sau đó dùng cơ chế chú ý để dự đoán các vùng ảnh có khả năng bị làm giả dưới dạng hộp giới hạn và nhãn tương ứng.

✓ **Ưu điểm:**

- Cho phép phát hiện và định vị vùng giả mạo.

- Không cần hộp tham chiếu như các phương pháp truyền thống (Faster R-CNN).

✓ **Hạn chế:**

- Chậm hơn phương pháp truyền thống.
- Không tối ưu nếu chỉ cần phân loại toàn ảnh.
- Cần dữ liệu huấn luyện có annotation vị trí.

Bảng 2.2: So sánh một số mô hình Transformers trong phát hiện ảnh Deepfake

Tiêu chí	Vision Transformer (ViT)	Swin Transformer	TimeSformer	DETR (adapted)
Loại dữ liệu	Ảnh tĩnh	Ảnh tĩnh	Video	Ảnh/Video
Trọng tâm chú ý	Toàn ảnh	Theo vùng trượt	Không gian + thời gian	Không gian
Ưu điểm nổi bật	Các đặc điểm toàn cục mạnh	Tối ưu tài nguyên	Học đặc trưng thời gian	Phát hiện vùng giả cụ thể
Hạn chế chính	Cần dữ liệu lớn	Cấu trúc phức tạp	Nặng, khó theo thời gian thực	Chậm, cần nhãn vị trí
Khả năng phân tích vùng giả	Kém	Trung bình	Trung bình	Tốt
Ứng dụng phù hợp	Phân loại toàn ảnh	Phân loại + phân vùng	Video Deepfake	Phân vùng và định vị giả

2.3 Mô hình học sâu cải tiến cho phát hiện ảnh Deepfake

2.3.1 Mô hình DTN

DTN (Distilled Transformers with Locally Enhanced Global Representations) [78] là một mô hình Transformer chưng cất được thiết kế chuyên dùng cho phát hiện ảnh Deepfake. DTN khắc phục những hạn chế của các mô hình phát hiện ảnh Deepfake đã nêu ở mục 2.2 như sau:

- So với các mô hình CNNs (mục 2.2.3), DTN có thiết kế gọn hơn, sử dụng ít tài nguyên hơn XceptionNet, EfficientNet song vẫn duy trì được hiệu suất. DTN đạt được độ cân bằng giữa thiết kế nhẹ tương đương MesoNet, song có độ chính xác cao hơn.
- So với các mô hình Transformer (mục 2.2.4), DTN có thiết kế nhẹ hơn do sử dụng cơ chế chưng cất (lọc thông tin cần thiết) mà vẫn đạt được hiệu suất cao và độ chính xác. Ngoài ra, DTN tốt hơn các mô hình đã có về khả năng phân biệt các chi tiết giả mạo nhỏ.

Đặc điểm cơ bản của DTN là kết hợp giữa đặc trưng cục bộ và toàn cục để phát hiện các chi tiết giả mạo nhỏ trong khuôn mặt. Mô hình này sử dụng mô-đun chuyên gia hỗn hợp (MoE) để khai thác các chi tiết giả mạo và mô-đun Transformer thị giác tăng cường cục bộ (LEVT) để học các biểu diễn toàn cục. DTN kết hợp các đặc trưng cục bộ và toàn cục đồng thời áp dụng kỹ thuật chưng cất để tạo khả năng áp dụng với các tập dữ liệu huấn luyện hạn chế. Mặc dù vậy, mô hình này vẫn cần đánh giá thêm về hiệu suất theo thời gian thực.

2.3.2 Mô hình kết hợp CNN và Vision Transformer

Trong nghiên cứu [61], nhóm tác giả đề xuất một phương pháp kết hợp giữa CNNs và Vision Transformer để phát hiện Deepfake, tập trung vào các vùng mắt, mũi và toàn bộ khuôn mặt. Mô hình thích hợp cho các hệ thống cần phát hiện Deepfake trong ảnh tĩnh, đặc biệt trong các ứng dụng mạng xã hội.

Mô hình sử dụng CNNs để phát hiện đặc trưng ở các vùng mắt và mũi, kết hợp với Vision Transformer để học các đặc trưng toàn cục từ khuôn mặt. Việc kết hợp kết quả từ các mô hình khác nhau thông qua phương pháp bỏ phiếu đa số để đưa ra quyết định cuối cùng. Ưu điểm nổi trội của mô hình là đạt độ chính xác 97% với các mô hình CNNs đã nêu. Tuy nhiên vẫn cần đánh giá thêm về khả năng tổng quát hóa trên các bộ dữ liệu khác nhau.

2.4 Đề xuất giải pháp sử dụng học sâu phát hiện ảnh Deepfake

Trong nghiên cứu phát hiện ảnh Deepfake, hai vấn đề cơ bản cần quan tâm là lựa chọn mô hình học sâu và tập dữ liệu phù hợp.

2.4.1 Lựa chọn tập dữ liệu

Để phục vụ cho mục đích nghiên cứu của đề án tốt nghiệp, tập dữ liệu hình ảnh cần phù hợp với yêu cầu phát hiện ảnh giả mạo do AI tạo ra. Tập tài liệu cần gồm hai lớp:

- Lớp ảnh thật: sử dụng các ảnh tự nhiên.
- Lớp ảnh Deepfake: được tạo từ các mô hình sinh ảnh, điển hình như với công cụ StyleGAN hay StyleGAN2 như đã trình bày ở Chương 1. Mặt khác, các ảnh này cần có cùng kiểu dữ liệu như lớp ảnh thật.
- Cân bằng giữa kích thước tập dữ liệu, độ phức tạp và khả năng ứng dụng rộng rãi trong thực tiễn.
- Phù hợp với điều kiện tài nguyên, năng lực tính toán.
- Có thể sử dụng để đánh giá nhiều mô hình học sâu khác nhau.

Qua quá trình tổng quan các vấn đề nghiên cứu, có thể sử dụng hai tập dữ liệu trong phạm vi ảnh tĩnh gồm:

- ***Tập dữ liệu CIFAKE***: chứa cả ảnh giả và ảnh tự nhiên có kích thước 32x32 điểm ảnh. Tập CIFAKE có khoảng 60.000 ảnh, rất phù hợp cho huấn luyện nhanh và thử nghiệm mô hình.
- ***Tập FaceForensics++***: Chứa các ảnh khuôn mặt giả, kích thước ảnh 256x256 điểm ảnh. Tập dữ liệu có khoảng 100.000 ảnh. Tuy nhiên, tập dữ liệu này thiên về khuôn mặt, phù hợp với mô hình cỡ lớn và video.

Tập dữ liệu CIFAKE bao gồm cả ảnh giả và ảnh thật tự nhiên có kích thước và định dạng giống nhau, phù hợp với nhiều mô hình học sâu tiên tiến. Do đó, lựa chọn tập dữ liệu CIFAKE phục vụ nghiên cứu trong đề án tốt nghiệp bởi những ưu điểm sau:

- CIFAKE gồm dữ liệu ảnh giả do AI sinh ra, do đó tập dữ liệu có tính tổng quát hơn các tập đã biết. Dữ liệu đa dạng giúp cho việc huấn luyện với nhiều dữ liệu khác nhau, tiến trình học không phụ thuộc vào một miền duy nhất.
- CIFAKE chứa các ảnh có kích thước nhỏ (32x32), nên phù hợp để huấn luyện nhanh trên các mô hình có kiến trúc nhẹ, có hạn chế về tài nguyên và năng lực tính toán. Độ phức tạp và chi phí tính toán là hai tiêu chí của một mô hình học sâu liên quan đến độ chính xác của mô hình, như đã chỉ ra trong [18].

- Ảnh giả sinh ra trong CIFAKE có kiểm soát, có gán nhãn rõ ràng, cho phép đảm bảo tính chính xác của nhãn và kiểm thử mô hình học sâu.
- CIFAKE phù hợp cho việc thử nghiệm các mô hình mới, so sánh đánh giá các thuật toán học sâu, đánh giá khả năng học phân biệt ảnh giả /ảnh thật của mô hình học sâu. Với CIFAKE, có thể thử nghiệm khả năng nhận diện ảnh giả do GANs hay StyleGAN sinh ra.

2.4.2 Lựa chọn mô hình học sâu cho phát hiện ảnh Deepfake

Trong nghiên cứu phát hiện ảnh Deepfake, việc lựa chọn mô hình học sâu phù hợp là yếu tố then chốt ảnh hưởng đến hiệu quả nhận diện và khả năng tổng quát hóa. Những vấn đề kỹ thuật nảy sinh khi lựa chọn mô hình học sâu gồm:

❖ **Mô hình cần phù hợp với tính thiếu đặc trưng rõ ràng và ổn định của dữ liệu**

Ảnh Deepfake thường được tạo ra bằng GANs hoặc các mô hình tổng hợp hình ảnh tinh vi, khiến các dấu hiệu giả mạo ngày càng khó nhận biết bằng mắt thường hoặc thuật toán đơn giản. Các đặc trưng bất thường (như biến dạng khuôn mặt, ánh sáng không tự nhiên, viền nhiễu) dễ bị làm mờ bởi kỹ thuật tổng hợp tiên tiến. Nhiều mô hình, ví dụ như mô hình XceptionNet [57] hoạt động tốt trên các bộ dữ liệu như Face2Face và FaceSwap, nhưng khó nhận diện các DeepFake được tạo bởi StyleGAN2.

❖ **Mô hình cần có khả năng tổng quát hóa với các tập dữ liệu**

Mô hình học sâu phát hiện Deepfake thường nhắm vào một số bộ dữ liệu cụ thể, khiến hiệu năng giảm mạnh khi áp dụng vào nguồn dữ liệu khác. Điều này đặt ra yêu cầu về huấn luyện với dữ liệu đa dạng hoặc áp dụng học không phụ thuộc miền. Có những mô hình hoạt động tốt trên tập huấn luyện, nhưng hiệu suất giảm hơn 30% khi áp dụng sang bộ dữ liệu khác [42].

❖ **Mô hình cần đạt cân bằng giữa hiệu năng và độ phức tạp**

Như đã trình bày ở mục 2.2, các mô hình như XceptionNet, EfficientNet hoặc mạng dựa trên Transformer mang lại độ chính xác cao nhưng có chi phí tính toán lớn, khó áp dụng trong môi trường thực tế. Các mô hình hạng nhẹ thường được áp dụng với mức độ cân bằng giữa độ chính xác và chi phí tính toán [18].

❖ **Khả năng chịu lỗi và thích ứng với kỹ thuật mới**

Deepfake ngày càng tinh vi với các mô hình như StyleGAN2, Diffusion-based, hoặc những mô hình điều khiển pose, ánh sáng, lời nói,... Các mô hình phát hiện hiện tại thường không theo kịp tiến bộ mới trong công nghệ tổng hợp ảnh, vì vậy cần cập nhật liên tục và huấn luyện trên các mẫu giả mạo mới [43].

Mặt khác, mô hình học sâu cần đạt yêu cầu về độ ổn định. Các mẫu Deepfake có thể được điều chỉnh để đánh lừa mô hình phát hiện bằng cách làm nhiễu nhỏ, làm sai lệch kết quả. Điều này đặt ra yêu cầu về tăng cường tính chống chịu cho hệ thống phát hiện. Ví dụ: Các mẫu Deepfake khi được thêm nhiễu bằng Fast Gradient Sign Method (FGSM) có thể khiến ResNet-50 và EfficientNet nhận sai toàn bộ video là thật, dù mắt thường vẫn thấy dấu hiệu giả mạo [10].

2.4.3 Giải pháp phát hiện ảnh Deepfake sử dụng học sâu

Qua kết quả nghiên cứu trong chương 1 và chương 2, từ mục đích nghiên cứu của đề án là nghiên cứu, xây dựng một giải pháp phát hiện ảnh DeepFake sử dụng phương pháp học sâu, giải pháp được đề xuất trong đề án tốt nghiệp gồm:

- Lựa chọn tập dữ liệu CIFAKE để thực hiện đánh giá các mô hình học sâu căn cứ vào những lý do đã nêu ở mục 2.4.1. Tập dữ liệu CIFAKE được sử dụng để huấn luyện mô hình học sâu.
- Ảnh cần phát hiện thật / giả được thu thập, tiền xử lý làm sạch dữ liệu, chuẩn hóa ảnh và sử dụng các mô hình học sâu để trích xuất đặc trưng.
- Lựa chọn ba mô hình học sâu phù hợp với các yêu cầu đã nêu ở mục 2.4.2. Khi triển khai phát hiện ảnh Deepfake, các mô hình sẽ được so sánh đánh giá theo các tiêu chí hiệu năng để chọn ra mô hình tốt nhất.

Ba mô hình khác nhau được lựa chọn xuất phát từ phân tích lý thuyết ở mục 2.2, cụ thể là hai mô hình CNN là **ResNet-50** và **EfficientNet-B0** cùng một mô hình Transformer là **Swin Transformer,** trên tập dữ liệu **CIFAKE.**

- ResNet-50 là một kiến trúc học sâu với cơ chế kết nối dư, cho phép mạng có thể học hiệu quả với độ sâu lớn mà không gặp vấn đề suy giảm gradient. Mô hình này thường được sử dụng làm chuẩn trong các bài toán phân loại hình ảnh, bao gồm cả nhận diện ảnh giả.

- EfficientNet-B0 là phiên bản nhẹ nhất của họ EfficientNet, sử dụng phương pháp cân bằng độ sâu, độ rộng và độ phân giải ảnh đầu vào. Nhờ đó, EfficientNet-B0 đạt hiệu năng tốt với số lượng tham số thấp, rất phù hợp để thử nghiệm trên dữ liệu ảnh kích thước nhỏ như CIFAKE (32x32 điểm).
- Swin Transformer là đại diện cho hướng tiếp cận mới bằng Transformer trong thị giác máy tính. Với thiết kế chú ý theo cửa sổ trượt, mô hình này có khả năng khai thác cả đặc trưng cục bộ và toàn cục, giúp tăng độ nhạy với những khác biệt tinh vi giữa ảnh thật và ảnh giả do AI tạo ra.

Ba mô hình trên sẽ được huấn luyện và đánh giá riêng biệt trên cùng tập dữ liệu CIFAKE, cho phép so sánh chính xác hiệu suất, độ chính xác, khả năng tổng quát hóa và chi phí tính toán của từng kiến trúc. Kết quả thử nghiệm của ba mô hình trên cùng một tập dữ liệu CIFAKE sẽ giúp đánh giá rõ ràng hơn về ưu nhược điểm của từng hướng tiếp cận trong bài toán phát hiện ảnh Deepfake.

2.5 Kết luận chương

Trong chương 2, đề án đã trình bày các nội dung liên quan đến giải pháp sử dụng học sâu trong phát hiện ảnh Deepfake. Các nội dung đã đề cập gồm: đặc điểm của ảnh Deepfake, các dấu hiệu đặc trưng và cách nhận biết; các phương pháp học sâu sử dụng trong phát hiện ảnh Deepfake bao gồm: các phương pháp học máy truyền thống, mô hình CNNs và các biến thể, các mô hình Transformer và các cải tiến.

Bốn mô hình CNNs được đánh giá cao trong phát hiện ảnh Deepfake gồm: Xception, MesoNet, ResNet-50 và EfficientNet-B0. Bốn mô hình Transformer gồm: Vision Transformer (ViT), Swin Transformer, TimeSpace Transformer, Dettection Transformer (DETR). Hai mô hình học sâu cải tiến được trình bày gồm: DTN (Distilled Transformers with Locally Enhanced Global Representations) và mô hình kết hợp CNN và Vision Transformer.

Qua khảo sát, các mô hình CNNs và Transformer đều có các ưu và nhược điểm trong bài toán phát hiện ảnh Deepfake. Để lựa chọn một giải pháp sử dụng học sâu cho phát hiện ảnh Deepfake và tính đa dạng của ảnh Deepfake, học viên đặt ra bốn tiêu chí kỹ thuật cần quan tâm gồm: Mô hình cần phù hợp với tính thiếu đặc trưng rõ ràng và ổn định của dữ liệu; Mô hình cần có khả năng tổng quát hóa với các tập dữ liệu; Mô hình

cần đạt cân bằng giữa hiệu năng và độ phức tạp; Mô hình cần có khả năng chịu lỗi và thích ứng với kỹ thuật mới, điển hình là các kỹ thuật tạo ảnh giả ngày càng đa dạng với công nghệ AI.

Chương 2 đã đưa ra đề xuất giải pháp học sâu gồm: 1) Lựa chọn tập dữ liệu phù hợp để thực hiện đánh giá các mô hình học sâu, cụ thể là tập CIFAKE; 2) Lựa chọn ba mô hình học sâu phù hợp với các yêu cầu đã nêu so sánh đánh giá theo các tiêu chí hiệu năng để chọn ra mô hình tốt nhất. Ba mô hình khác nhau được lựa chọn xuất phát từ phân tích lý thuyết ở mục 2.2, cụ thể là hai mô hình CNNs là ResNet-50 và EfficientNet-B0, một mô hình Transformer là Swin Transformer được so sánh, đánh giá trên cùng một tập dữ liệu ảnh giả CIFAKE.

Trong chương tiếp theo, đề án sẽ trình bày chi tiết về việc triển khai giải pháp, cách thức tiền xử lý dữ liệu ảnh, chuẩn hóa dữ liệu ảnh, phương pháp huấn luyện mô hình, tiến trình huấn luyện, phương pháp đánh giá các mô hình theo các tiêu chí hiệu năng và thực hiện thử nghiệm, đánh giá kết quả.

CHƯƠNG 3. THỰC HIỆN MÔ HÌNH HỌC SÂU TRONG PHÁT HIỆN ẢNH DEEPFAKE

3.1 Sơ đồ khối mô hình học sâu phát hiện ảnh Deepfake

Hình 3.1 là sơ đồ khối mô hình học sâu cho phát hiện ảnh Deepfake được đề xuất trong đề án tốt nghiệp.

Hình 3.1: Các bước thực hiện mô hình phát hiện ảnh Deepfake

Mô hình phát hiện ảnh Deepfake được đề xuất cơ bản tuân thủ đúng 5 bước thực hiện ở trên. Bước 1 được gồm 2 thành phần:

- Tập dữ liệu CIFAKE chứa cả ảnh thật và ảnh giả (do GAN và StyleGAN2 tạo), đã được gán nhãn rõ ràng (real, fake). Tập dữ liệu CIFAKE gồm các ảnh có kích thước 32x32, RGB, theo định dạng png hoặc jpg. Tổng số ảnh khoảng 120.000 ảnh được chia theo tỷ lệ 50% ảnh thật, 50% ảnh giả. Tỷ lệ cân bằng này giúp cho việc huấn luyện mô hình tốt hơn, không bị lệch về ảnh giả hay ảnh thật. Tập dữ liệu CIFAKE sử dụng tỷ lệ chia dữ liệu sẵn có cho huấn luyện là 80%, nghĩa là 96.000 ảnh và cho kiểm thử là 20%, nghĩa là 24.000 ảnh.
- Dữ liệu ảnh cần được kiểm thử thật/giả được thu thập từ thực tế. Dữ liệu này cần được tiền xử lý, chuẩn hóa tương tự như ảnh từ tập CIFAKE để bảo đảm sự tương đồng của dữ liệu.

Trong phạm vi đề án, dữ liệu ảnh để kiểm thử mô hình cũng được lấy từ tập CIFAKE.

Bước 2 gồm khối làm sạch và tiền xử lý dữ liệu thực hiện các nhiệm vụ: hiệu chỉnh kích thước ảnh (đối với ảnh thu thập từ thực tế), chuẩn hóa ảnh, tăng cường chất lượng ảnh.

Bước 3 là khối mô hình học sâu: có thể chọn một trong ba mô hình là: ResNet-50, EfficientNet-B0 hoặc Swin Transformer như đã đề cập ở Chương 2. Module huấn luyện thực hiện huấn luyện theo tập dữ liệu huấn luyện (48.000 ảnh) để có khả năng phân loại ảnh thật/giả. Module phát hiện nhận ảnh đầu vào chưa rõ thật/giả đã qua bước tiền xử lý để nhận diện, phát hiện ảnh thật/giả với mô hình học sâu đã huấn luyện.

Bước 4 và 5: Đầu ra của khối phát hiện trả về kết quả phân loại ảnh thật/giả kèm theo các tham số phục vụ cho đánh giá hiệu năng cho các mô hình. Căn cứ vào kết quả đánh giá hiệu năng theo các tiêu chí, có thể lựa chọn mô hình cho kết quả tốt nhất đối với ảnh cần nhận diện, phát hiện.

3.2 Thu thập và mô tả dữ liệu

Lựa chọn bộ dữ liệu thử nghiệm: Bộ dữ liệu CIFAKE là một bộ dữ liệu cân bằng và bao gồm 120.000 hình ảnh từ hai lớp khác nhau, cụ thể là hình ảnh thực (60.000 ảnh) và hình ảnh tổng hợp do AI tạo ra (60.000 ảnh). Các hình ảnh thực được lấy từ bộ dữ liệu CIFAR-10 (Tập dữ liệu chứa 60.000 hình ảnh màu 32 × 32 trong 10 lớp (máy bay, ô tô, chim, mèo, hươu, chó, ếch, ngựa, tàu, xe tải), với 6.000 hình ảnh cho mỗi lớp; 50.000 hình ảnh đào tạo và 10.000 hình ảnh thử nghiệm thường được sử dụng để đào tạo các thuật toán học máy và thị giác máy tính có sẵn tại [84]. Hình ảnh tổng hợp do AI được tạo ra bằng mô hình khuếch tán tiềm ẩn (Các hình ảnh được tạo ra thuộc về lớp hình ảnh chung, chúng không có nội dung cụ thể như khuôn mặt người hoặc nghệ thuật).

Bộ dữ liệu CIFAKE bao gồm hàng ngàn hình ảnh, trong đó có các cặp ảnh thật và ảnh do AI tạo ra. Mỗi hình ảnh trong bộ dữ liệu được phân loại rõ ràng thành ảnh thật và ảnh giả, giúp cho quá trình huấn luyện mô hình dễ dàng hơn. Đặc điểm của CIFAKE là các hình ảnh giả được thiết kế với mức độ chân thực cao, giúp tạo nên thử thách lớn cho các mô hình phân loại.

Các hình ảnh giả trong CIFAKE được tạo ra từ nhiều mô hình AI khác nhau, bao gồm cả GANs và mô hình Diffusion, giúp tạo nên sự đa dạng về phong cách và độ phức

tạp của ảnh giả. Do đó, bộ dữ liệu này rất hữu ích cho việc huấn luyện các mô hình phát hiện ảnh giả trong các tình huống thực tế.

Đặc điểm: CIFAKE chứa các hình ảnh có kích thước và định dạng chuẩn hóa (thường là 32x32 pixel), giúp dễ dàng tích hợp vào các mô hình học sâu và tối ưu hóa hiệu suất tính toán. Bộ dữ liệu này bao gồm cả hình ảnh thực và hình ảnh giả, cho phép các mô hình học máy có thể học cách phân biệt và phát hiện các đặc điểm khác nhau giữa hai loại hình ảnh này.

Cấu trúc: CIFAKE thường được xây dựng với hai nhãn chính:

- Real (Thật): Bao gồm các hình ảnh được chụp hoặc thu thập từ thế giới thực.
- Fake (Giả): Bao gồm các hình ảnh được tạo ra bởi AI.

Mỗi nhãn sẽ có một tập hợp hình ảnh riêng biệt, giúp phân biệt rõ ràng giữa ảnh thật và ảnh giả.

Cách xây dựng: CIFAKE được phát triển thông qua việc thu thập các hình ảnh từ các nguồn ảnh thực, đồng thời kết hợp các công cụ sinh ảnh AI tiên tiến như GANs hoặc mô hình Diffusion để tạo ra các hình ảnh giả. Việc chọn lọc hình ảnh giả trong CIFAKE tuân theo các tiêu chí nhất định để đảm bảo rằng các ảnh này có chất lượng tương đương và có khả năng gây nhầm lẫn cho người xem, từ đó tăng cường độ khó cho việc nhận diện và phân loại. Để đảm bảo tính chính xác và đa dạng, CIFAKE được xây dựng từ các bộ dữ liệu phong phú, đồng thời sử dụng các phương pháp xử lý và kiểm định để tối ưu hóa chất lượng.

3.3 Tiền xử lý và làm sạch dữ liệu

- **Phát hiện và cắt khuôn mặt**: Sử dụng các thuật toán sẵn có trong thư viện Python để phát hiện khuôn mặt trong ảnh và cắt vùng chứa khuôn mặt, điển hình như thuật toán MTCNN (Multi-task Cascaded Convolution Networks) hoặc dlib.
- **Căn chỉnh khuôn mặt**: Căn chỉnh khuôn mặt theo hướng chuẩn dựa trên các điểm đặc trưng như mắt, mũi, miệng để đảm bảo tính nhất quán trong dữ liệu.
- **Chuẩn hóa ảnh**: Thay đổi kích thước ảnh về kích thước chuẩn (ví dụ: 32×32) và chuẩn hóa giá trị pixel để phù hợp với mô hình học sâu.

- **Tăng cường dữ liệu**: Triển khai các kỹ thuật như lật ảnh theo trục ngang, xoay góc, làm mờ, điều chỉnh độ sáng, nhằm mở rộng và đa dạng hóa tập dữ liệu huấn luyện, từ đó cải thiện khả năng tổng quát hóa của mô hình.
- **Trích xuất đặc trưng**: Áp dụng các kiến trúc học sâu như CNNs hoặc Transformer để tự động trích xuất và học các đặc trưng phân biệt từ dữ liệu hình ảnh sau tiền xử lý, hỗ trợ hiệu quả cho các tác vụ phân loại và nhận dạng.

3.4 Xây dựng mô hình phát hiện ảnh Deepfake

Các mô hình CNNs đã được huấn luyện trước và được thực nghiệm đánh giá trong nghiên cứu này bao gồm: **Resnet-50, Swin Transformer, EfficientNet**.

3.4.1 Mô Hình Resnet-50

Như đã trình bày ở Chương 2, mô hình ResNet (Residual Network) [32] có cơ chế bỏ qua kết nối để huấn luyện mạng sâu hiệu quả. Mô hình có độ sâu không quá lớn, phù hợp cho phát hiện ảnh giả mạo trong môi trường có hạn chế tài nguyên, dễ dàng tinh chỉnh các tham số lại cho từng bộ dữ liệu mới.

Một đặc điểm nổi bật của ResNet-50 là việc sử dụng cơ chế kết nối dư, cho phép tín hiệu trong mạng lan truyền qua các lớp mà không bị suy giảm, giúp giải quyết hiệu quả vấn đề suy giảm gradient và cải thiện khả năng học ở các mô hình sâu, giúp mô hình có thể học và xử lý những đặc điểm tinh vi trong ảnh giả. Do vậy, ResNet-50 thường được dùng làm chuẩn so sánh trong các bài toán phân loại hình ảnh, nhận diện ảnh giả. Mặc dù có nhiều phiên bản ResNet ví dụ ResNet101, song ResNet-50 được lựa chọn trong bài như đã nêu lý do ở Chương 2.

Trong nhận diện ảnh giả, ResNet rất hữu ích vì nó có thể phát hiện những chi tiết nhỏ khó nhận thấy bằng mắt thường nhưng đặc trưng của ảnh do AI tạo ra. Các chi tiết như ánh sáng, độ tương phản và sự nhất quán của kết cấu có thể không hoàn toàn tự nhiên trong ảnh giả. ResNet, nhờ vào cấu trúc sâu và khả năng học đặc trưng phức tạp, có thể phân biệt được những bất thường này với ảnh thật một cách hiệu quả. Ngoài ra, nhờ residual connections, ResNet có khả năng học nhanh hơn và giảm thiểu vấn đề quá khớp, làm cho nó trở thành lựa chọn phù hợp cho các bài toán phân loại phức tạp trên các bộ dữ liệu lớn như CIFAKE.

Một điểm mạnh khác của ResNet là khả năng tương thích cao với các phương pháp học chuyển giao. Trong thực tế, ResNet thường được huấn luyện trước trên các bộ dữ liệu lớn như ImageNet và sau đó được tinh chỉnh trên các bộ dữ liệu chuyên biệt cho nhận diện ảnh giả. Điều này không chỉ giúp tiết kiệm tài nguyên tính toán mà còn tăng cường độ chính xác của mô hình.

Một vấn đề phổ biến trong các mạng nơ-ron sâu, nhất là khi có nhiều lớp là suy giảm gradient, nghĩa là độ của gradient trở nên rất nhỏ (tiệm cận về 0) khi lan truyền ngược từ đầu ra về phía các lớp đầu vào, khiến cho các trọng số ở những lớp đầu hầu như không được cập nhật trong quá trình huấn luyện. ResNet được thiết kế để giải quyết vấn đề này nhờ vào cơ chế kết nối dư. Cấu trúc này cho phép các thông tin quan trọng trong quá trình huấn luyện đi qua các lớp mà không bị suy giảm, giúp mô hình học sâu hơn mà không gặp khó khăn với gradient.

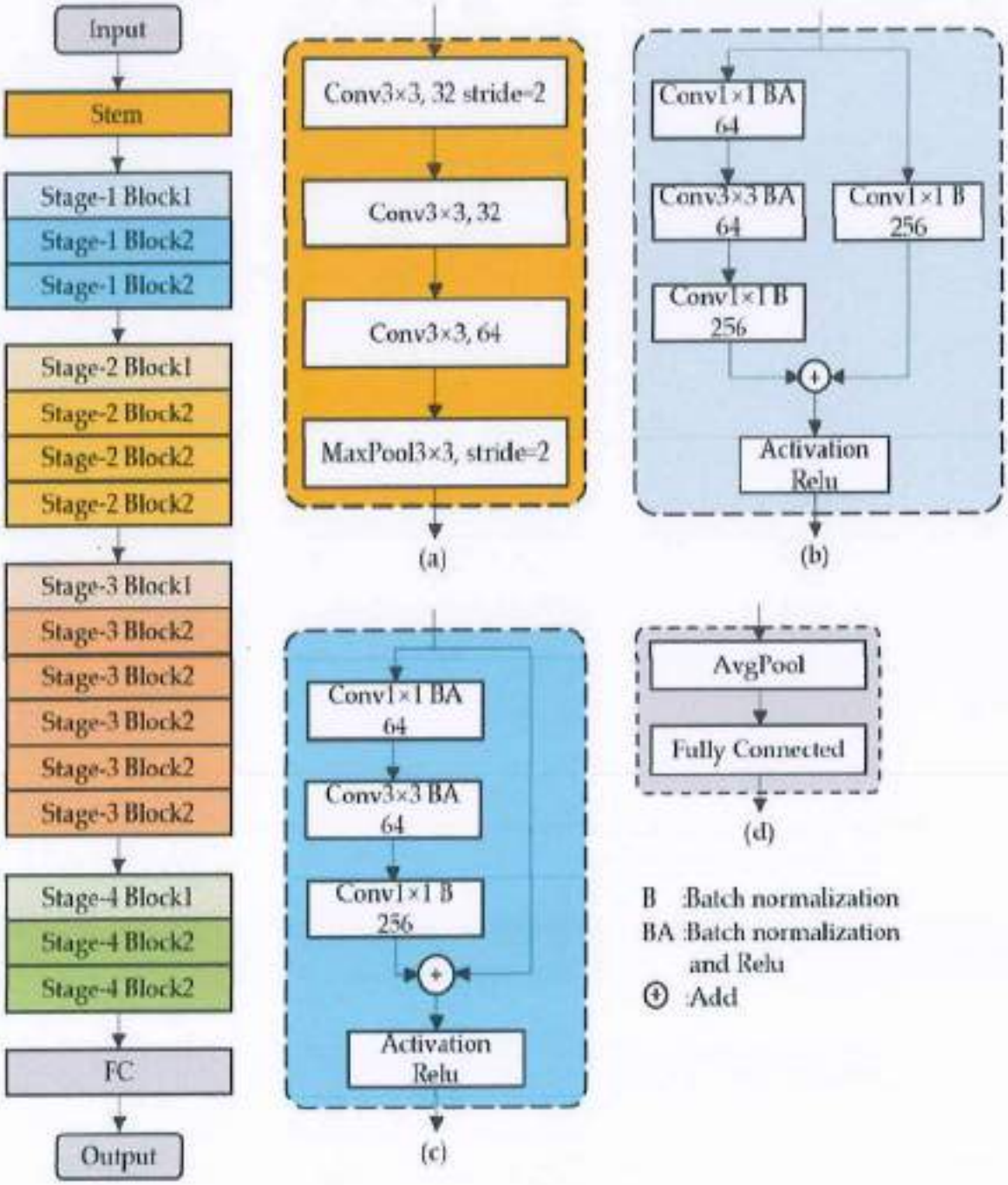
Trong bài toán nhận diện ảnh Deepfake, ResNet có thể phát hiện các bất thường hoặc lỗi nhỏ trong ảnh giả do AI sinh ra, nhờ khả năng trích xuất đặc trưng phức tạp. Cụ thể, ResNet đã chứng minh được hiệu quả cao trong các bộ dữ liệu lớn như CIFAKE, nơi mà việc phân biệt ảnh thật và ảnh giả cần một mô hình có khả năng nhận diện được các chi tiết nhỏ nhưng quan trọng. Ngoài ra, nhờ vào kiến trúc dư, ResNet có thể huấn luyện nhanh hơn và giảm thiểu vấn đề quá khớp so với nhiều mô hình học sâu khác.

Kiến trúc: ResNet-50 là mô hình mạng nơ-ron tích chập sâu với 50 lớp, do Microsoft Research phát triển. Đặc điểm nổi bật của ResNet là sử dụng các khối dư, giúp giải quyết vấn đề suy giảm độ chính xác khi tăng số lượng lớp của mô hình. Các khối dư này cho phép tín hiệu đầu vào có thể được truyền thẳng đến đầu ra thông qua bỏ qua kết nối.

Đặc điểm nổi bật:

- **Khả năng học hiệu quả ở độ sâu cao**: Với các khối dư, ResNet-50 có thể duy trì độ chính xác cao ngay cả khi số lượng lớp tăng lên, giúp mô hình học các đặc điểm phức tạp trong ảnh mà không bị suy giảm hiệu suất.
- **Giảm thiểu vấn đề vanishing gradient**: Bỏ qua kết nối giúp lan truyền gradient tốt hơn trong quá trình huấn luyện, từ đó cải thiện khả năng tối ưu hóa.

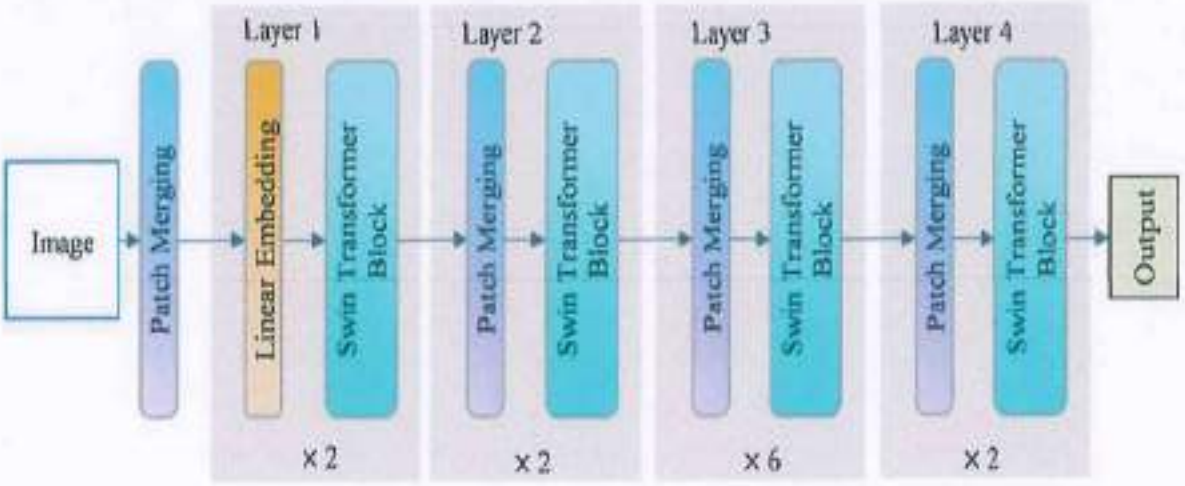
- **Ứng dụng rộng rãi trong nhiều tác vụ**: ResNet-50 được sử dụng rộng rãi trong các ứng dụng về nhận diện ảnh, phân loại và phát hiện đối tượng, nhờ vào tính linh hoạt và hiệu quả trong xử lý hình ảnh phức tạp.



Hình 3.2: Kiến trúc mô hình ResNet-50 [74]

3.4.2 Mô hình Swin Transformer

Trong những năm gần đây, sự nổi lên của kiến trúc Transformer đã mở ra một hướng tiếp cận mới cho các bài toán thị giác máy tính, vốn trước đây chủ yếu dựa trên CNNs. Swin Transformer là một biến thể tiên tiến được đề xuất bởi Liu và cộng sự vào năm 2021 [44], mang lại hiệu quả vượt trội trong nhiều tác vụ thị giác như phân loại ảnh, phát hiện đối tượng và phân đoạn ảnh.



Hình 3.3: Mô hình Swin Transformer [79]

Khác với kiến trúc Vision Transformer (ViT) truyền thống – vốn xử lý toàn bộ ảnh đầu vào như một chuỗi các vùng và không có tính cục bộ, Swin Transformer giới thiệu cơ chế cửa sổ trượt nhằm khai thác tính chất cục bộ của hình ảnh. Cụ thể, ảnh đầu vào được chia thành các patch không chồng lấp, sau đó xử lý tuần tự trong các khối Transformer nhỏ với phạm vi tính toán giới hạn trong từng cửa sổ. Ở các tầng kế tiếp, các cửa sổ này được dịch chuyển để mở rộng phạm vi tương tác giữa các vùng của ảnh, từ đó cải thiện khả năng học đặc trưng toàn cục nhưng vẫn giữ được hiệu suất tính toán cao.

Một điểm mạnh của Swin Transformer là khả năng mở rộng theo phân cấp – tương tự như các mạng CNNs – bằng cách giảm dần độ phân giải của các vùng khi đi sâu vào mạng, đồng thời tăng số chiều biểu diễn. Điều này cho phép mô hình học được các đặc trưng từ cục bộ đến toàn cục một cách hiệu quả, và phù hợp với các bài toán yêu cầu xử

lý ảnh ở nhiều mức độ chi tiết khác nhau, như phát hiện vùng giả mạo trong ảnh hoặc video Deepfake.

So với các kiến trúc CNNs hiện đại như EfficientNet hay ResNet, Swin Transformer thể hiện khả năng cạnh tranh mạnh mẽ về độ chính xác, đồng thời có khả năng học biểu diễn dài hạn tốt hơn – một yếu tố quan trọng trong việc phát hiện các bất thường nhỏ hoặc không gian thời gian trong nội dung giả mạo.

Trong các nghiên cứu gần đây, Swin Transformer đã được ứng dụng thành công trong việc trích xuất đặc trưng cho hệ thống phát hiện Deepfake, đặc biệt là khi tích hợp vào các pipeline với phân tích khung hoặc đoạn video liên tiếp. Nhờ cơ chế chú ý giới hạn trong cửa sổ trượt và khả năng mở rộng đa tầng, mô hình này giúp tăng cường hiệu suất và độ chính xác trong các hệ thống nhận diện nội dung giả mạo hiện đại.

3.4.3 Mô hình EfficientNet

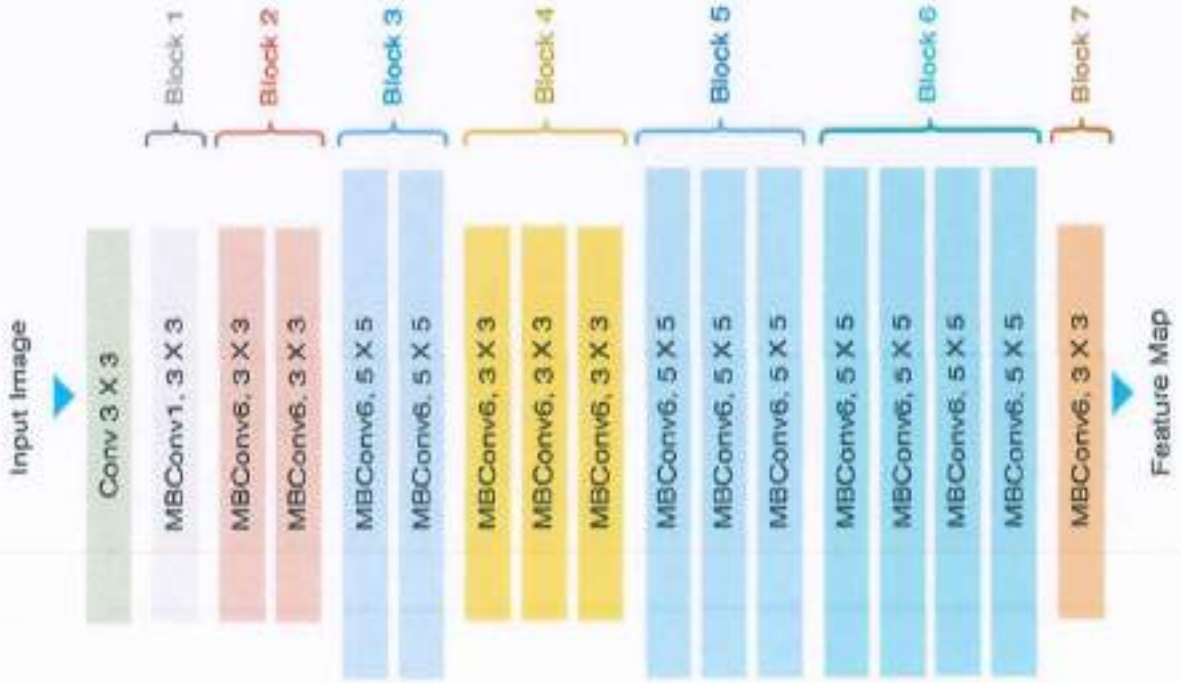
EfficientNet là một họ các kiến trúc mạng CNNs được đề xuất bởi Tan và Le vào năm 2019 tại Google AI [65]. Mô hình này được phát triển với mục tiêu tối ưu hóa hiệu suất phân loại hình ảnh trong khi vẫn duy trì mức độ tính toán hợp lý. Điểm nổi bật của EfficientNet là cách tiếp cận mới trong việc mở rộng kiến trúc mạng bằng chiến lược gọi là compound scaling, đồng thời kết hợp hiệu quả ba yếu tố: chiều sâu, chiều rộng và độ phân giải đầu vào.

Khác với các phương pháp mở rộng truyền thống vốn chỉ điều chỉnh đơn lẻ một trong ba yếu tố trên, EfficientNet tối ưu hóa đồng thời cả ba thông số theo một tỉ lệ được xác định từ mô hình gốc EfficientNet-B0. Cụ thể, mô hình gốc được thiết kế thông qua kỹ thuật AutoML để đạt hiệu suất cao nhất trên tập ImageNet với số lượng tham số và FLOPs tối thiểu. Các phiên bản mở rộng như EfficientNet-B1 đến B7 được tạo ra bằng cách tăng tỉ lệ đồng thời cả ba chiều theo một hệ số φ (phi), đảm bảo sự cân bằng giữa tài nguyên tính toán và độ chính xác mô hình.

Về mặt cấu trúc, EfficientNet sử dụng khối tích chập nút cổ chai đảo chiều cho thiết bị di động kế thừa từ MobileNetV2, cùng với kỹ thuật nén - kích hoạt giúp cải thiện khả năng học các đặc trưng kênh. Việc kết hợp những kỹ thuật này giúp mô hình vừa nhẹ, vừa mạnh, rất phù hợp để triển khai trên các hệ thống giới hạn tài nguyên hoặc cần thời gian phản hồi nhanh như trong các ứng dụng phát hiện giả mạo thời gian thực.

Hiệu suất của EfficientNet được chứng minh là vượt trội so với nhiều mô hình CNNs truyền thống khác như ResNet, DenseNet hay Inception, khi xét trên cả độ chính xác và mức tiêu thụ tài nguyên. Trong các nghiên cứu áp dụng vào bài toán phát hiện ảnh Deepfake, EfficientNet thường được lựa chọn như một backbone mạnh mẽ để trích xuất đặc trưng, đặc biệt khi xử lý các tập dữ liệu có độ phân giải cao hoặc biến đổi phức tạp.

Nhờ khả năng mở rộng hiệu quả, kết hợp với kiến trúc tinh gọn, EfficientNet trở thành một trong những lựa chọn hàng đầu cho các hệ thống nhận dạng hình ảnh hiện đại, trong đó có cả các giải pháp phát hiện nội dung giả mạo.



Hình 3.4: Mô hình EfficientNet-B0 [6]

3.5 Đánh giá hiệu năng của các mô hình

Sau khi hoàn tất quá trình huấn luyện, các mô hình sẽ được đánh giá để kiểm tra hiệu suất dựa trên các độ đo quan trọng là Accuracy (Độ chính xác), Precision (Độ chính xác trong phát hiện nội dung giả), Recall (Khả năng nhận diện toàn bộ nội dung giả) và F1-score (chỉ số đánh giá tổng hợp giữa Precision và Recall).

Accuracy sẽ đo lường tỷ lệ dự đoán đúng trên tổng số mẫu trong tập kiểm thử, trong khi Precision đánh giá mức độ chính xác của các dự đoán mà mô hình cho là nội dung

giả. Recall sẽ đo lường khả năng của mô hình trong việc phát hiện toàn bộ các mẫu giả mạo và F1-score sẽ cung cấp một chỉ số cân bằng giữa Precision và Recall để đánh giá khả năng tổng quát của mô hình trong các tình huống thực tế. Các chỉ số này giúp hiểu rõ hơn về độ tin cậy của từng mô hình (ResNet-50, Swing Transformer, EfficientNet-B0) trong nghiên cứu phát hiện ảnh Deefake, cho phép so sánh và chọn lựa mô hình phù hợp nhất cho mục tiêu của đề án tốt nghiệp.

3.5.1 Độ chính xác (Accuracy)

Độ chính xác được định nghĩa là tỷ lệ giữa tổng số giá trị dự đoán đúng so với tổng số dự đoán. Độ chính xác thường được sử dụng để đánh giá các mô hình phân loại. Độ chính xác của mô hình học máy cho biết tổng số lần mô hình dự đoán đúng trên toàn bộ tập dữ liệu. Tuy nhiên, độ chính xác có thể không phản ánh đầy đủ hiệu suất trong trường hợp dữ liệu không cân bằng (ví dụ, nhiều ảnh thật hơn ảnh giả hoặc ngược lại) [60].

Công thức tính:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.1}$$

trong đó:

- TP (True Positive) – Dương tính thật: Số lượng ảnh giả được dự đoán đúng là giả.
- TN (True Negative) – Âm tính thật: Số lượng ảnh thật được dự đoán đúng là thật.
- FP (False Positive) – Dương tính giả: Số lượng ảnh thật bị dự đoán sai là giả.
- FN (False Negative) - Âm tính giả: Số lượng ảnh giả bị dự đoán sai là thật.

3.5.2 Tỷ lệ trúng (Precision)

Tỷ lệ trúng đo lường tỷ lệ các dự đoán dương tính thật (ảnh giả được dự đoán là giả) trên tổng số dự đoán dương tính. Tỷ lệ trúng phản ánh khả năng của mô hình trong việc không tạo ra quá nhiều kết quả dương tính giả. Chỉ số này rất quan trọng trong các bài toán mà việc dự đoán sai (ảnh thật bị nhận diện là giả) có thể gây ra hậu quả nghiêm trọng. Tỷ lệ trúng cao cho thấy mô hình ít có xu hướng nhầm lẫn ảnh thật thành ảnh giả [81-60].

Công thức tính:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3.2}$$

trong đó:

- TP (True Positive): Số lượng ảnh giả được dự đoán đúng là giả.
- FP (False Positive): Số lượng ảnh thật bị dự đoán sai là giả.

3.5.3 Độ nhạy (Recall)

Recall, hay còn gọi là độ nhạy hoặc độ thu hồi, đo lường tỷ lệ các dự đoán dương tính thật trên tổng số mẫu dương tính thực tế (ảnh giả được nhận diện là giả trên tổng số ảnh giả thực sự). Recall cho biết mô hình có thể nhận diện bao nhiêu phần trăm ảnh giả trong tổng số ảnh giả thực tế. Đây là chỉ số quan trọng để đánh giá hiệu suất của mô hình trong việc phát hiện ảnh giả, đặc biệt là khi việc bỏ sót các ảnh giả (FN) cần được hạn chế tối đa. Recall cao cho thấy mô hình ít bỏ sót ảnh giả [60]

Công thức tính:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3.3}$$

trong đó:

- TP (True Positive): Số lượng ảnh giả được dự đoán đúng là giả.
- FN (False Negative): Số lượng ảnh giả bị dự đoán sai là thật.

3.5.4 F1 Score

F1 Score là độ đo kết hợp giữa Precision và Recall, cung cấp cái nhìn cân bằng hơn trong các trường hợp có sự đánh đổi giữa Rrecision và Recall. F1 Score sẽ cao khi cả Precision và Recall đều cao. F1 Score là chỉ số rất hữu ích khi chúng ta muốn một cân bằng giữa Precision và Recall, nhất là trong các trường hợp dữ liệu không cân bằng. F1 Score cao cho thấy mô hình có thể vừa nhận diện chính xác ảnh giả (Precision) vừa không bỏ sót ảnh giả nào (Recall).

Công thức tính:

$$\text{F1 Score} = 2\,x\,\frac{Precision\ x\ Recall}{Precision + Recall} \tag{3.4}$$

trong đó:

- Precision : Tỷ lệ trúng
- Recall: Độ nhạy

3.6 Môi trường thử nghiệm

3.6.1 Ngôn ngữ lập trình và thư viện

Phần thực nghiệm của đề án tốt nghiệp này được thực hiện bằng ngôn ngữ lập trình Python (Python 3.6.6 :: Anaconda custom 64-bit) và các thư viện đã được phát triển cho nó. Python là một ngôn ngữ mã nguồn mở cấp cao [29], đã trở thành một trong những ngôn ngữ được sử dụng nhiều nhất trong học máy và khoa học máy tính. Nó có nhiều thư viện mã nguồn mở được phát triển mà một số trong đó được sử dụng trong đề án tốt nghiệp này. Các thư viện được sử dụng bao gồm: numpy, scipy, pandas, matplotlib, scikit-learning [53], TensorFlow, xgboost, imblearn, plotly, Keras và seaborn. Các thư viện này cung cấp các công cụ để tiền xử lý, các thuật toán cho học máy và cách thức vẽ biểu đồ dữ liệu.

3.6.2 Cấu hình máy tính thử nghiệm

- Hệ điều hành: Windows 10 Pro
- CPU: Intel(R) Core(TM) i7-4720HQ CPU @2.60GHz
- RAM: 8GB

3.6.3 Cấu trúc tập dữ liệu thử nghiệm

Tập dữ liệu CIFAKE gồm 2 loại ảnh:

- Ảnh thật: Lấy trực tiếp từ tập dữ liệu CIFAR-10, gán nhãn Real
- Ảnh giả: Sinh ra từ mô hình StyleGAN2 huấn luyện trên tập ảnh CIFAR-10, gán nhãn Fake.
- Kích thước ảnh: 32x32 pixel, RGB (3 màu)
- Định dạng ảnh: png, jpg.
- Số lượng ảnh: 120.000 ảnh
- Tỷ lệ ảnh thật: 60.000 ảnh (50%). Tỷ lệ ảnh giả: 60.000 ảnh (50%)
- Tỷ lệ tập huấn luyện: 96.000 ảnh (80%).
- Tỷ lệ tập kiểm thử: 24.000 ảnh (20%).

3.6.4 Các tham số cơ bản của các mô hình

- Input_size: (32, 32, 3)
- Num_classes: 2 (real/fake)
- Learning Rate: $1e^{-3}$

- Batch_size: 64 (mô hình Swin Transformer sử dụng batch nhỏ = 16 để tránh quá tải)
- Epochs: 50
- Loss function: Binary Cross Entropy.

3.6.5 Tiêu chí đánh giá hiệu năng cho các mô hình học sâu

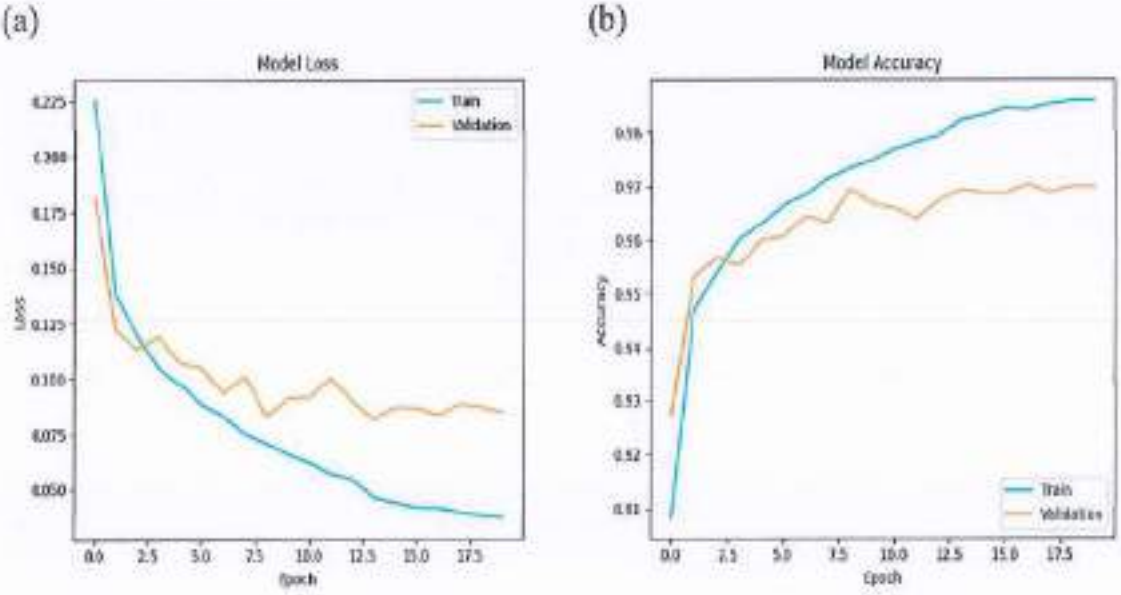
Các tham số dùng để đánh giá hiệu năng của các mô hình bao gồm:

- Độ chính xác (Accuracy)
- Độ dự đoán chính xác (Precision)
- Độ nhạy (Recall)
- F1-score

3.7 Kết quả thử nghiệm

3.7.1 Kết quả huấn luyện và tối ưu tham số các mô hình

a. Mô hình Resnet-50



Hình 3.5: Kết quả huấn luyện và tối ưu tham số của mô hình ResNet-50

Hình 3.5 biểu thị kết quả huấn luyện và tối ưu tham số của mô hình ResNet-50, cụ thể: (a) Biểu đồ quá trình thay đổi hàm mất mát trong quá trình huấn luyện và kiểm thử, (b) Biểu đồ quá trình thay đổi độ chính xác tương ứng theo từng epoch.

Theo như kết quả thể hiện trên Hình 3.5(a), hàm mất mát trên tập huấn luyện và tập kiểm thử theo số lượng epoch liên tục giảm đều, cho thấy mô hình đang học tương đối hiệu quả từ tập dữ liệu huấn luyện. Cụ thể, hàm mất mát trên tập huấn luyện giảm

đều đặn và ổn định, chứng tỏ mô hình dần học tốt hơn các đặc trưng từ dữ liệu đầu vào. Trong giai đoạn đầu (epoch 0 - 4), hàm mất mát của tập kiểm thử cũng giảm tương ứng với hàm của tập huấn luyện, phản ánh khả năng khái quát hóa tốt. Tuy nhiên, từ epoch thứ 5 trở đi, hàm mất mát trên tập kiểm thử bắt đầu dao động và giảm chậm lại, trong khi hàm mất mát trên tập huấn luyện tiếp tục giảm sâu. Hiện tượng này cho thấy mô hình có dấu hiệu bắt đầu quá khớp (overfitting) – tức là học quá chi tiết dữ liệu huấn luyện nhưng không cải thiện hiệu quả trên dữ liệu mới. Dù overfitting chưa nghiêm trọng (do mất mát trên tập kiểm thử vẫn ổn định quanh giá trị 0,075), việc khoảng cách giữa hai đường mất mát ngày càng mở rộng là tín hiệu cần lưu ý. Từ góc độ thực nghiệm, thời điểm mô hình đạt hiệu quả khái quát tốt nhất có thể nằm trong khoảng epoch thứ 4 đến thứ 7 và việc áp dụng kỹ thuật kết thúc huấn luyện sớm tại thời điểm này là hợp lý. Ngoài ra, để cải thiện khả năng tổng quát hóa, có thể cân nhắc thêm các biện pháp như điều chuẩn hoặc tăng cường dữ liệu nhằm tăng độ đa dạng của dữ liệu huấn luyện.

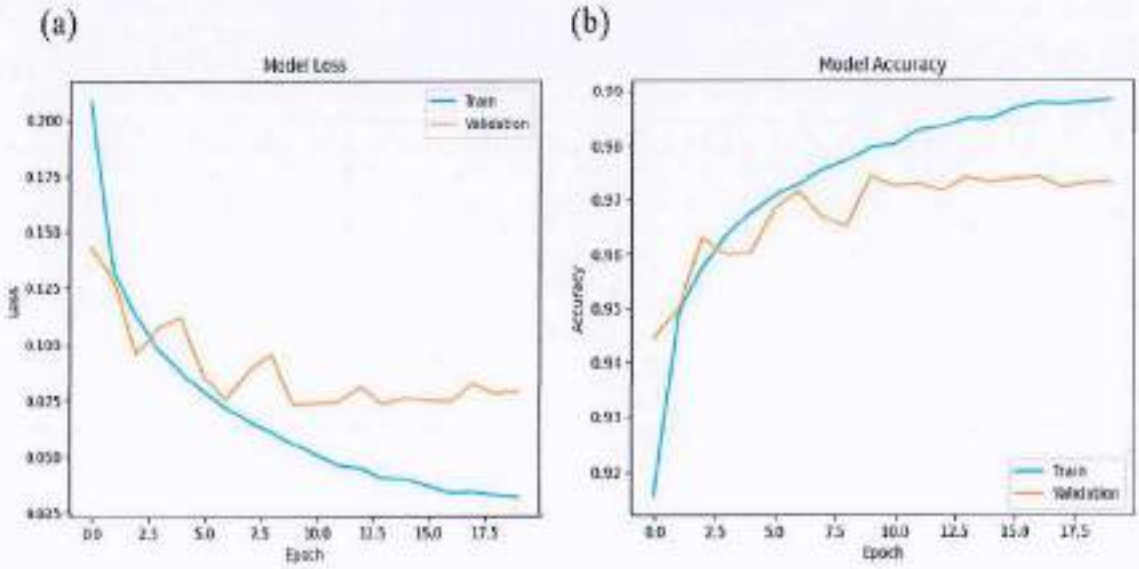
Sự thay đổi độ chính xác của mô hình trên trên tập huấn luyện và tập kiểm thử qua 20 epoch được thể hiện trên Hình 3.5(b), cho thấy quá trình học diễn ra hiệu quả và ổn định. Cụ thể, độ chính xác trên tập huấn luyện liên tục tăng đều, đạt trên 98,5% ở giai đoạn cuối, cho thấy mô hình học tốt từ dữ liệu đầu vào. Trong khi đó, độ chính xác trên tập kiểm thử cũng tăng nhanh trong các epoch đầu, đạt trên 96,5% và giữ ổn định sau đó, dao động nhẹ quanh ngưỡng này. Việc khoảng cách giữa hai đường độ chính xác dần mở rộng sau epoch thứ 10 cho thấy có dấu hiệu mô hình bắt đầu học quá mức dữ liệu huấn luyện (overfitting nhẹ). Tuy nhiên, vì độ chính xác trên tập kiểm thử vẫn giữ ổn định cao và không suy giảm rõ rệt, hiện tượng quá khớp này chưa ảnh hưởng tiêu cực đến khả năng khái quát hóa của mô hình. Nhìn chung, mô hình đạt hiệu suất tốt cả trên tập huấn luyện lẫn kiểm thử và thời điểm hội tụ hợp lý có thể rơi vào khoảng epoch thứ 10 - 12. Nếu mục tiêu là tối ưu hóa khả năng khái quát hóa và tránh lãng phí tài nguyên, việc áp dụng kết thúc huấn luyện sớm, kết hợp với các kỹ thuật như điều chuẩn hoặc tăng cường dữ liệu, là hướng đi phù hợp trong các lần huấn luyện tiếp theo.

Như vậy, có thể thấy mô hình ResNet - 50 cho thấy khả năng học hiệu quả ổn định và hiệu suất mạnh mẽ trên cả hai tập dữ liệu huấn luyện và kiểm thử. Tuy nhiên, do có dấu hiệu quá khớp nhẹ sau khoảng 10 - 12 epoch, có thể xem xét bổ sung các kỹ thuật

như kết thúc huấn luyện sớm, tăng cường dữ liệu hoặc điều chuẩn loại bỏ kết nối ngẫu nhiên khi cả độ chính xác và độ lỗi trên tập kiểm thử đều đạt trạng thái ổn định, qua đó cải thiện khả năng tổng quát hóa hơn nữa.

b. Mô hình EficientNet-B0

Hình 3.6 là kết quả huấn luyện và tối ưu tham số của mô hình EficientNet-B0: (a) Biểu đồ quá trình thay đổi hàm mất mát trong quá trình huấn luyện và kiểm thử, (b) Biểu đồ quá trình thay đổi độ chính xác tương ứng theo từng epoch.



Hình 3.6: Kết quả huấn luyện và tối ưu tham số của mô hình EficientNet-B0

Theo như kết quả thể hiện trên Hình 3.6(a), hàm mất mát trên tập huấn luyện và tập kiểm thử qua 20 epoch cho thấy quá trình học diễn ra tương đối hiệu quả nhưng cũng xuất hiện dấu hiệu overfitting. Cụ thể, hàm mất mát trên tập huấn luyện giảm đều đặn và ổn định trong suốt quá trình huấn luyện, từ hơn 0,20 xuống khoảng 0,025, cho thấy mô hình ngày càng học tốt hơn với dữ liệu đầu vào. Trong khi đó, hàm mất mát trên tập kiểm thử cũng giảm nhanh trong những epoch đầu tiên, đặc biệt đến khoảng epoch thứ 5 - 6, tuy nhiên sau đó bắt đầu dao động quanh mức 0,075 và không tiếp tục cải thiện, mặc dù hàm mất mát trên tập huấn luyện vẫn tiếp tục giảm. Khoảng cách ngày càng lớn giữa hai đường giá trị mất mát là dấu hiệu rõ ràng của hiện tượng quá khớp – khi mô hình học quá chi tiết vào tập huấn luyện nhưng khả năng tổng quát hóa trên dữ liệu mới không còn cải thiện. Trong bối cảnh này, mô hình có thể đạt hiệu suất tổng quát tốt nhất vào khoảng epoch thứ 6 - 8, và việc áp dụng kỹ thuật dừng huấn luyện sớm là cần thiết để

ngăn ngừa lãng phí tài nguyên huấn luyện. Ngoài ra, nên xem xét bổ sung các kỹ thuật chuẩn hóa (như Dropout, L2) và tăng cường dữ liệu để tăng cường khả năng khái quát và làm mượt lại đường mất mát trên tập kiểm thử.

Độ chính xác trên tập huấn luyện tăng đều và đạt gần 99% (Hình 3.6(b)), trong khi độ chính xác trên tập kiểm thử cũng tăng mạnh trong 5 epoch đầu và duy trì ở mức cao khoảng 97,2%. Mức độ dao động của hàm độ chính xác trên tập kiểm thử rất nhỏ, cho thấy hiệu suất ổn định, không có hiện tượng lệch pha nghiêm trọng giữa huấn luyện và kiểm thử.

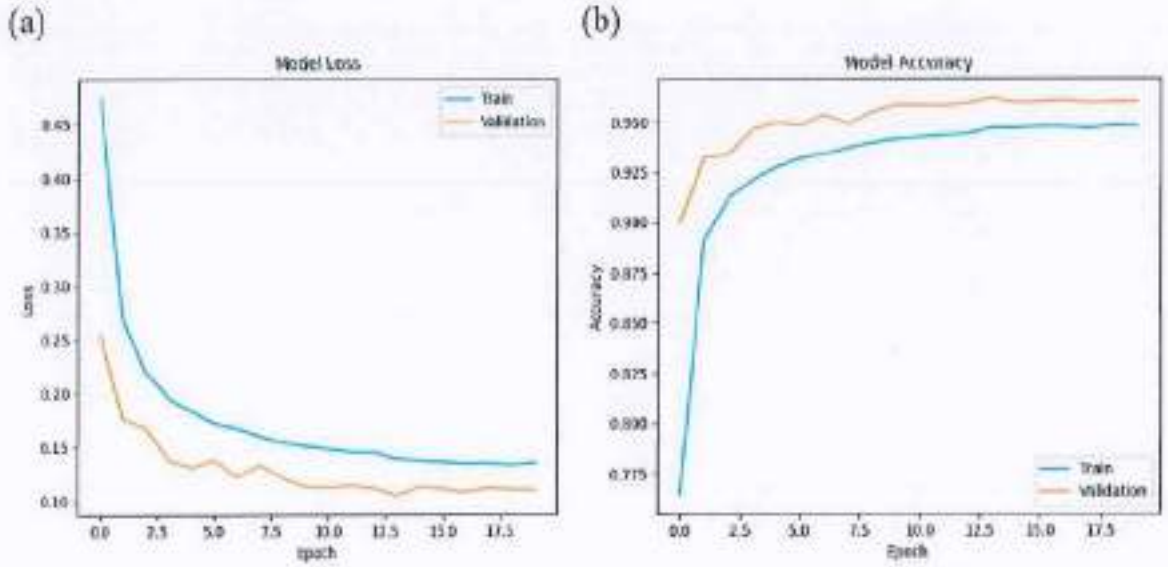
Tổng thể, hình 3.6 cho thấy mô hình học tập hiệu quả trên tập huấn luyện, đạt độ chính xác cao liên tục qua các epoch và hàm mất mát giảm đều đặn. Tuy nhiên, ở tập kiểm thử, mặc dù hàm mất mát giảm nhanh trong giai đoạn đầu và độ chính xác tăng đáng kể, mô hình bắt đầu có dấu hiệu quá khớp nhẹ từ khoảng epoch thứ 6 trở đi. Cụ thể, hàm mất mát trên tập kiểm thử dao động quanh mức 0,075 và không còn giảm, trong khi độ chính xác trên tập kiểm thử giữ ổn định quanh 97,3% nhưng không tiếp tục tăng, trái ngược với đường huấn luyện vẫn đang cải thiện. Khoảng cách giữa tập huấn luyện và tập kiểm thử ngày càng lớn cho thấy mô hình đang dần học quá mức dữ liệu huấn luyện mà không mang lại cải thiện trên dữ liệu chưa thấy. Do đó, thời điểm hội tụ tốt nhất có thể rơi vào khoảng epoch thứ 6 - 8 và cần cân nhắc áp dụng kỹ thuật dừng sớm để dừng huấn luyện đúng lúc. Đồng thời, có thể tăng cường khả năng khái quát hóa bằng các biện pháp như kỹ thuật loại bỏ nút ngẫu nhiên, điều chuẩn L2, hoặc tăng cường dữ liệu. Nhìn chung, mô hình đạt hiệu suất cao và ổn định, nhưng cần điều chỉnh huấn luyện để tối ưu hoá năng lực tổng quát trên dữ liệu thực tế.

Mô hình EfficientNet-B0 cho thấy hiệu quả cao trong quá trình huấn luyện và kiểm thử, với độ chính xác kiểm định trên 97% và hàm mất mát thấp. Tuy xuất hiện dao động nhẹ trong hàm mất mát trên tập kiểm thử, mô hình vẫn duy trì khả năng tổng quát hóa ổn định. Điều này chứng minh EfficientNet-B0 là một mô hình nhẹ nhưng mạnh, phù hợp cho các bài toán phân loại ảnh với hiệu năng cao và chi phí tính toán thấp.

c. Mô hình Swin

Hình 3.7 là kết quả huấn luyện và tối ưu tham số của mô hình Swin: (a) Biểu đồ quá trình thay đổi hàm mất mát trong quá trình huấn luyện và kiểm thử, (b) Biểu đồ quá trình thay đổi độ chính xác tương ứng theo từng epoch.

Biểu đồ bên trái (Hình 3.7(a)) thể hiện sự thay đổi của hàm mất mát, trong khi biểu đồ bên phải (Hình 3.7(b)) biểu diễn độ chính xác trên cả hai tập dữ liệu huấn luyện và kiểm thử qua 20 epoch.



Hình 3.7: Kết quả huấn luyện và tối ưu tham số của mô hình Swin

Quá trình huấn luyện diễn ra hiệu quả với cả hai đường hàm mất mát trên tập huấn luyện và kiểm thử đều giảm mạnh trong những epoch đầu tiên. Đặc biệt, hàm mất mát trên tập kiểm thử có xu hướng giảm nhanh và hội tụ sớm hơn trên tập huấn luyện, đạt mức thấp và ổn định từ khoảng epoch thứ 5 trở đi, dao động nhẹ quanh ngưỡng 0,11 - 0,13. Trong khi đó, hàm mất mát trên tập huấn luyện tiếp tục giảm nhưng chậm dần và có dấu hiệu "chững lại" sau epoch thứ 15, với một chút dao động nhẹ ở cuối chuỗi epoch.

Điểm đáng chú ý là hàm mất mát trên tập kiểm thử luôn thấp hơn tập huấn luyện trong suốt quá trình huấn luyện — điều này khá bất thường nhưng không sai về mặt kỹ thuật. Nó có thể cho thấy rằng tập kiểm thử có độ phức tạp thấp hơn, ít nhiễu hơn hoặc không đủ đa dạng, khiến mô hình dễ khái quát hóa trên đó hơn so với tập huấn luyện. Ngoài ra, điều này cũng có thể phản ánh rằng các kỹ thuật điều chuẩn đang phát huy tác dụng hiệu quả, chẳng hạn như kỹ thuật loại bỏ ngẫu nhiên, L2 hoặc chuẩn hóa theo batch — giúp mô hình học sâu nhưng không quá lệ thuộc vào tập huấn luyện.

Tuy không xuất hiện quá khớp rõ rệt, nhưng độ chênh lệch giữa hai đường hàm mất mát sau epoch 10 –15 có thể được theo dõi thêm trong các lần huấn luyện tiếp theo. Việc áp dụng dừng sớm tại khoảng epoch thứ 15 có thể là hợp lý để tiết kiệm tài nguyên và tránh việc mô hình học thêm nhưng không mang lại cải thiện đáng kể.

Biểu đồ thể hiện độ chính xác trên tập huấn luyện và kiểm thử trong suốt 20 epoch cho thấy mô hình đạt được hiệu suất huấn luyện ổn định và có khả năng tổng quát hóa tốt (Hình 3.7(b)). Độ chính xác trên tập huấn luyện tăng đều từ khoảng 77% ở epoch đầu tiên lên hơn 93% ở giai đoạn cuối, phản ánh quá trình học diễn ra hiệu quả và liên tục. Đáng chú ý, độ chính xác trên tập kiểm thử đạt mức cao hơn cả tập huấn luyện ngay từ giai đoạn đầu (trên 90%) và duy trì ổn định quanh mức 95 – 96% trong suốt quá trình huấn luyện. Điều này cho thấy mô hình không bị quá khớp, thậm chí có thể đang được hưởng lợi từ điều chuẩn tốt hoặc tập kiểm thử ít nhiễu hơn so với tập huấn luyện.

Hiện tượng độ chính xác trên tập kiểm thử cao hơn tập huấn luyện là hợp lý trong một số tình huống, ví dụ như khi áp dụng dropout trong quá trình huấn luyện nhưng không áp dụng trong quá trình đánh giá, hoặc khi dữ liệu kiểm thử mang tính phân bố dễ phân loại hơn. Mặc dù không có dấu hiệu quá khớp, sự khác biệt bền vững này giữa hai đường độ chính xác cho thấy có thể xem xét thêm các chiến lược như điều chỉnh lại dữ liệu huấn luyện, giảm tỉ lệ dropout hoặc kiểm tra độ tương đồng phân phối giữa tập huấn luyện và kiểm thử để đảm bảo đánh giá toàn diện hơn. Điều này cho thấy Swin Transformer học rất hiệu quả và không có hiện tượng học lệch hoặc quá khớp.

Trong quá trình thực nghiệm, ba mô hình ResNet-50, EfficientNet-B0 và Swin Transformer đều thể hiện hiệu suất huấn luyện và kiểm thử tốt, song vẫn có sự khác biệt đáng chú ý về khả năng học và mức độ tổng quát hóa. Mô hình ResNet-50 đạt độ chính xác kiểm thử khoảng 97% và duy trì hàm mất mát trên tập kiểm thử ổn định (~ 0,075), tuy nhiên xuất hiện dấu hiệu quá khớp nhẹ khi hàm mất mát trên tập huấn luyện tiếp tục giảm còn hàm mất mát trên tập kiểm thử dao động nhẹ ở giai đoạn sau. Mô hình EfficientNet-B0 cho thấy khả năng học mạnh mẽ với độ chính xác trên tập kiểm thử đạt ~97,2% và tập huấn luyện gần 99%, tuy có dao động nhỏ ở hàm mất mát trên tập kiểm thử nhưng vẫn giữ được tính ổn định trong tổng thể. Đáng chú ý, Swin Transformer vượt trội ở khả năng tổng quát hóa: mô hình này đạt độ chính xác kiểm thử cao nhất (~95,5%)

ngay từ các epoch đầu, và giữ vững trong suốt quá trình huấn luyện, với hàm mất mát trên tập kiểm thử luôn thấp hơn hàm mất mát trên tập huấn luyện – cho thấy không có hiện tượng quá khớp, ngược lại còn tổng quát hóa tốt hơn trên dữ liệu chưa thấy.

Tổng kết lại, cả ba mô hình đều phù hợp cho bài toán phát hiện ảnh Deepfake, trong đó Swin Transformer thể hiện sự ổn định và khả năng tổng quát hóa vượt trội, EfficientNet-B0 đạt hiệu suất rất cao với chi phí tính toán thấp, còn ResNet-50 vẫn là một mô hình nền tảng mạnh với khả năng học sâu và hiệu suất cao, mặc dù cần điều chỉnh để giảm thiểu quá khớp.

3.7.2 Kết quả đánh giá hiệu năng của các mô hình

Kết quả đánh giá hiệu năng của các mô hình được thể hiện trong Bảng 3.1.

Bảng 3.1: Phân tích đánh giá hiệu năng của các mô hình ResNet-50, EfficientNet-B0 và Swin Transformer

Mô hình	Độ chính xác (Accuracy)	Tỷ lệ trúng (Precision)	Độ nhạy (Recall)	F1-Score
ResNet-50	0,9695	0,9719	0,9669	0,9694
EfficientNet-B0	0,9745	0,9875	0,9611	0,9741
Swin Transformer	0,9615	0,9733	0,9489	0,9610

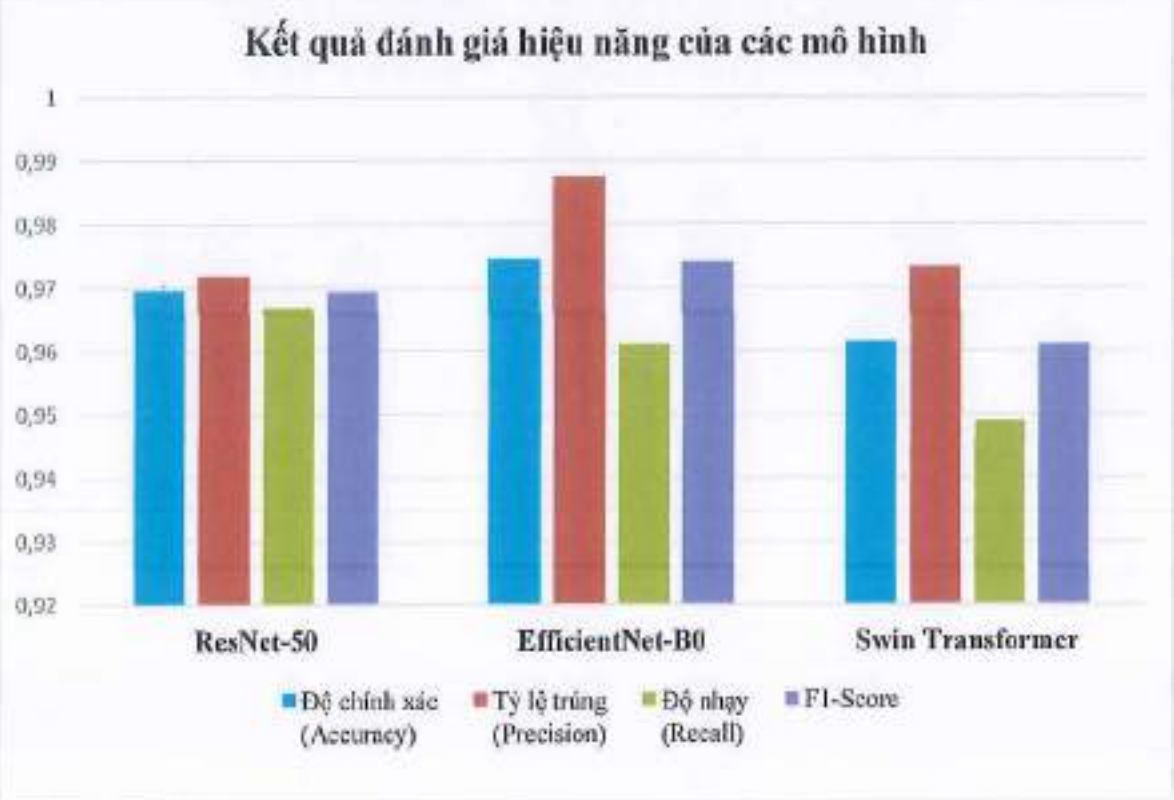
Từ Bảng 3.1, có thể thấy, mô hình Swin Transformer có kết quả hiệu suất thấp hơn so với các mô hình học chuyển giao khác. Trong ba mô hình được triển khai, EfficientNet-B0 đạt độ chính xác cao nhất với giá trị xấp xỉ 97,5%, theo sau là ResNet-50 (~97%) và Swin Transformer (~96,2%). Về mặt tỷ lệ trúng , EfficientNet-B0 tiếp tục dẫn đầu với giá trị gần 98,8%, trong khi Swin Transformer cũng đạt kết quả khá cao (~97,3%).

Thực tế, việc nhận diện sai một mẫu thật là ảnh giả (Deepfake) gây ra hậu quả ít nghiêm trọng hơn so với việc phân loại nhầm một ảnh Deepfake thành ảnh thật. Do đó, mục tiêu chính của các kỹ thuật phát hiện Deepfake là giảm thiểu số lượng False negative, tức là tối ưu hóa chỉ số Recall.

Hình 3.8 biểu thị kết quả đánh giá hiệu năng cho các mô hình đã triển khai. Như thể hiện, ResNet-50 đạt recall cao nhất (~96,7%), trong khi Swin Transformer có giá trị thấp nhất (~94,9%).

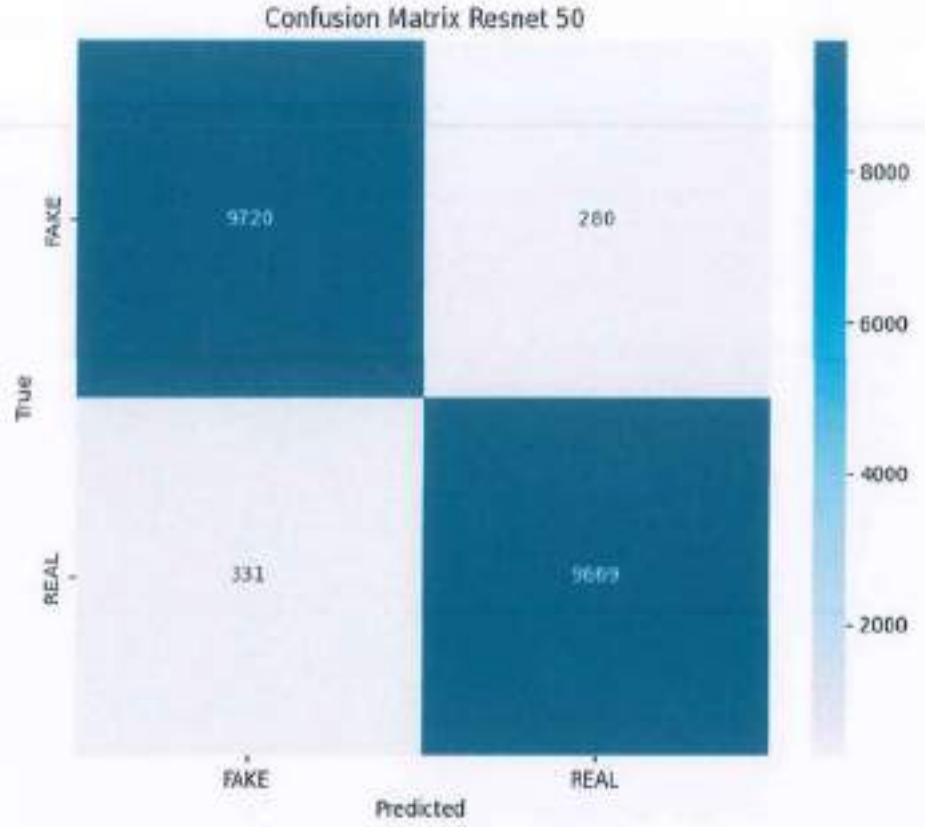
Chỉ số F1-score, đại diện cho sự cân bằng giữa tỷ lệ trúng và độ nhạy, cho thấy EfficientNet-B0 là mô hình tổng thể hiệu quả nhất với giá trị xấp xỉ 97,4%, tiếp theo là ResNet-50 (~97%) và Swin Transformer (~96,1%).

Những kết quả này chỉ ra rằng EfficientNet-B0 không chỉ có độ chính xác cao mà còn duy trì tốt sự ổn định giữa các tiêu chí đánh giá, trong khi Swin Transformer cần được cải thiện ở khả năng nhận diện đúng các ảnh Deepfake.

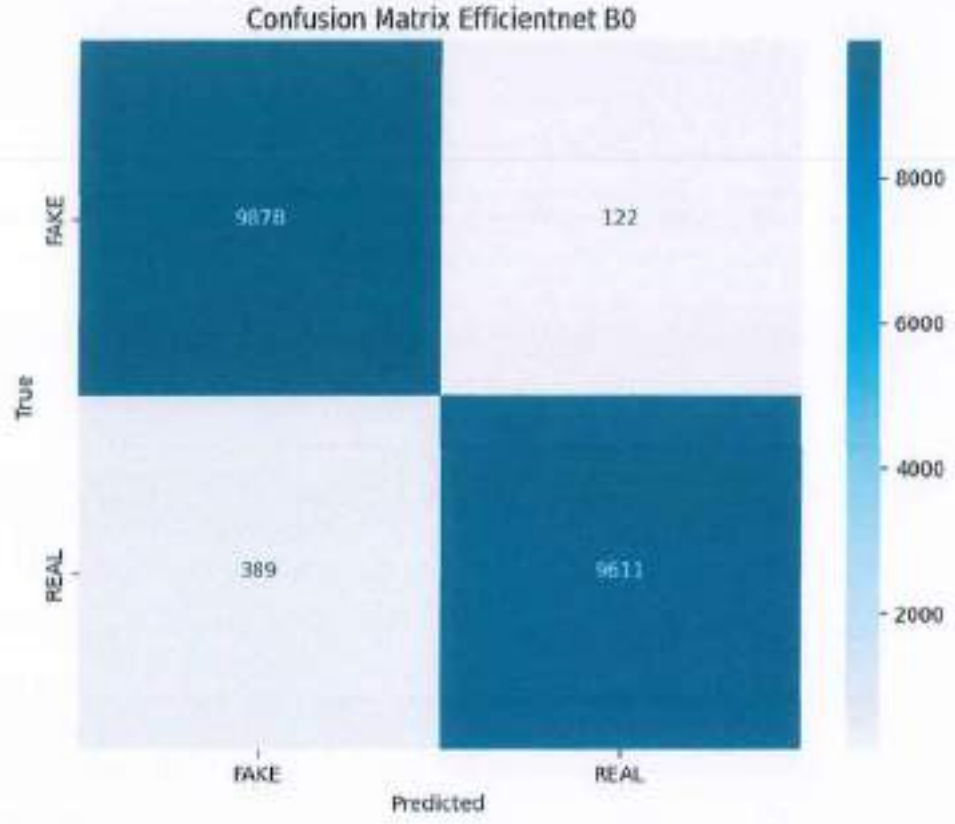


Hình 3.8: So sánh hiệu năng của ResNet-50, EfficientNet-B0 và Swin Transformer

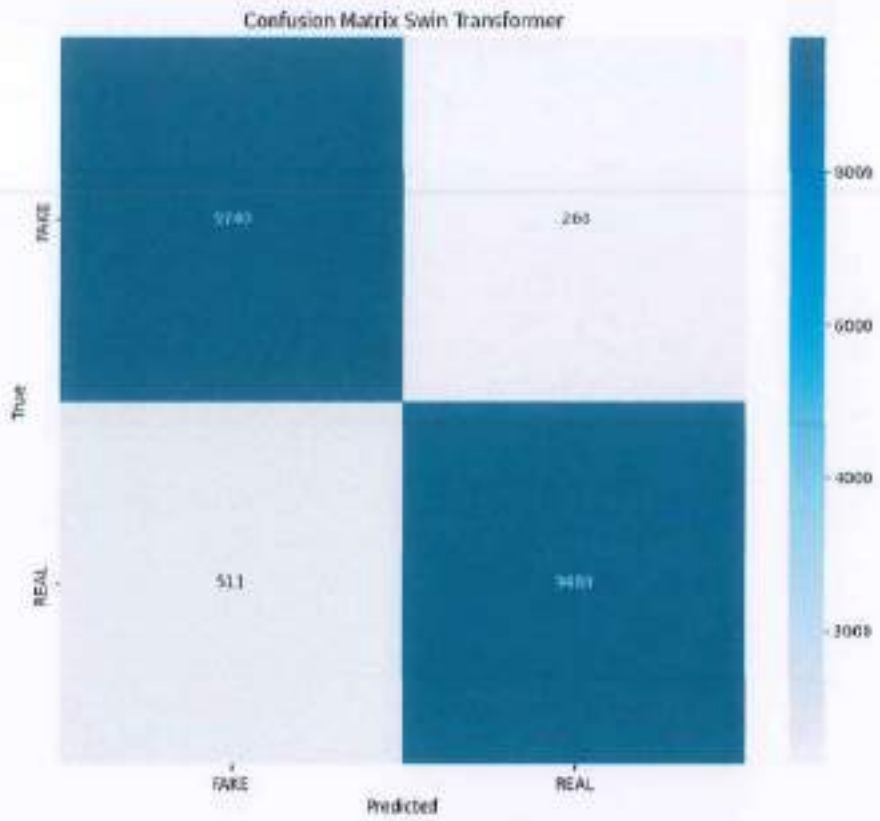
Một chỉ số quan trọng khác được sử dụng trong nghiên cứu này để đánh giá các mô hình là ma trận nhầm lẫn, được trình bày trong Hình 3.9, 3.10 và 3.11 tương ứng với các mô hình ResNet-50, EfficientNet-B0 và Swin Transformer. Chỉ số này cho phép kiểm tra hiệu năng của thuật toán bằng cách so sánh kết quả dự đoán với giá trị thực của nhãn.



Hình 3.9: Phân tích ma trận nhầm lẫn của mô hình ResNet-50



Hình 3.10: Phân tích ma trận nhầm lẫn của mô hình EfficientNet-B0

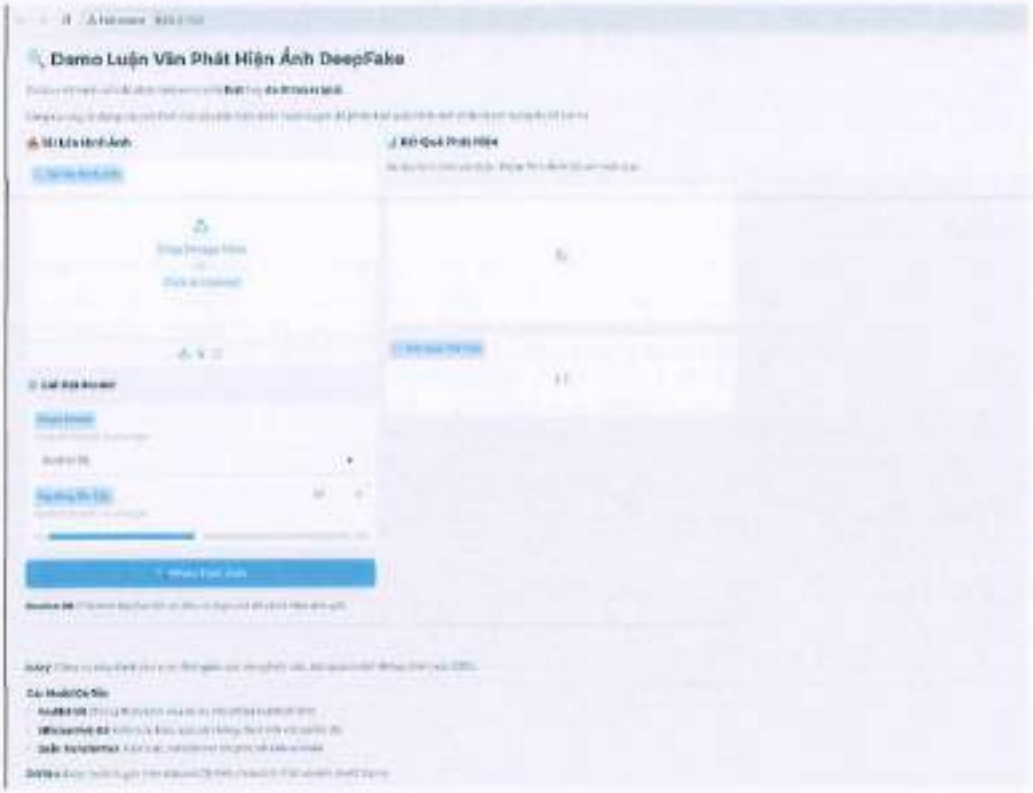


Hình 3.11: Phân tích ma trận nhầm lẫn của mô hình Swin Transformer

Kết quả từ ma trận nhầm lẫn cho thấy EfficientNet-B0 là mô hình có hiệu suất phân loại theo lớp tốt nhất, với số lượng lỗi thấp nhất trong ba mô hình khảo sát. Mô hình này chỉ ghi nhận 122 ảnh Deepfake bị phân loại nhầm là thật và 389 ảnh thật bị phân loại nhầm là giả. ResNet-50 có tổng số lỗi ở mức trung bình, trong khi Swin Transformer ghi nhận số lượng lỗi cao nhất, đặc biệt là tỷ lệ ảnh thật bị nhầm thành Deepfake (511 trường hợp) – yếu tố cần được lưu ý trong các hệ thống yêu cầu độ tin cậy cao.

Những kết quả này khẳng định hiệu quả nhận diện của EfficientNet-B0, đặc biệt trong việc giảm thiểu false negative, điều rất quan trọng đối với bài toán phát hiện ảnh Deepfake.

3.8 Kết quả demo ứng dụng phát hiện ảnh Deepfake



Hình 3.12: Ảnh giao diện của ứng dụng

Hình 3.13: Giao diện kết quả khi đưa ảnh Deepfake

Hình 3.14: Giao diện kết quả khi đưa ảnh ảnh thật

Hình 3.15: Giao diện ứng dụng cho phép người dùng có thể chọn lựa mô hình xác thực ảnh

3.9 Kết luận chương

Trong Chương 3, đề án đã trình bày các nội dung về triển khai thực hiện mô hình học sâu trong phát hiện ảnh Deepfake. Nội dung chương đã trình bày sơ đồ các khối trong mô hình học sâu, mô tả các thành phần trong mô hình: thu thập dữ liệu, trình bày ba mô hình học sâu áp dụng gồm: ResNet-50, EfficientNet-B0, Swin Transformer. Tiếp đó, đề án đã trình bày các tiêu chí đánh giá hiệu năng cho các mô hình, môi trường và các tham số dùng cho thử nghiệm. Các kết quả thử nghiệm cho ba mô hình cho thấy, với tập dữ liệu CIFAKE sử dụng, mô hình EfficientNet cho độ chính xác cao hơn so với hai mô hình còn lại. Kết quả từ ma trận nhầm lẫn cho thấy EfficientNet-B0 là mô hình có hiệu suất phân loại theo lớp tốt nhất, với số lượng lỗi thấp nhất trong ba mô hình khảo sát. Ngoài ra, đề án cũng minh họa kết quả xác thực ảnh thật/giả của ứng dụng phát hiện ảnh Deepfake - ứng dụng xây dựng dựa trên kết quả thực nghiệm của đề án.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN TIẾP

Kết luận

Phát hiện ảnh Deepfake là một nhu cầu thực tiễn và là vấn đề kỹ thuật khó, có nhiều thách thức do các ảnh Deepfake do AI tạo ra rất đa dạng. Ảnh Deepfake đang bị lợi dụng cho các mục đích xấu, như lừa đảo, phát tán thông tin sai lệch, xâm phạm quyền riêng tư và gây ảnh hưởng đến danh dự cá nhân. Sự lan truyền nhanh chóng của ảnh Deepfake đặt ra những thách thức lớn đối với tính xác thực của thông tin, an ninh mạng, pháp lý và đạo đức xã hội. Việc nghiên cứu, phát triển các phương pháp và xây dựng giải pháp phát hiện ảnh Deepfake trở thành một yêu cầu cấp thiết hiện nay.

Các phương pháp và mô hình học sâu tỏ ra rất hiệu quả trong việc xử lý và phân loại hình ảnh thông qua các kiến trúc mạng nơ-ron nhân tạo tiên tiến như Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) và Transformers [11-13, 15-17]. Các mô hình học sâu có khả năng tự động trích xuất đặc trưng từ hình ảnh mà không cần phải xác định trước các đặc điểm quan trọng, từ đó giúp cải thiện đáng kể độ chính xác trong phát hiện ảnh Deepfake. Trên cơ sở đó, đề án tốt nghiệp hướng tới mục tiêu: Nghiên cứu, ứng dụng các phương pháp học sâu vào xây dựng giải pháp phát hiện ảnh Deepfake.

Các kết quả chính đã đạt được trong bài gồm:

- Nghiên cứu về công nghệ AI sử dụng tạo ảnh Deepfake; Các mô hình học máy và học sâu; Các đặc trưng của ảnh Deepfake do AI tạo ra; Phương pháp nhận diện và phát hiện ảnh Deepfake; Các phương pháp học sâu trong phát hiện ảnh Deepfake.
- Đưa ra được một mô hình áp dụng phương pháp học sâu vào phát hiện Deepfake do AI tạo ra.
- Thực hiện thử nghiệm mô hình, đánh giá mô hình bằng các chỉ số Accuracy, Precision, Recall, F1-score và so sánh đánh giá ba mô hình tiêu biểu là ResNet-50, EfficientNet-B0 và Swin Transformer.

Từ kết quả thực nghiệm và phân tích hiệu năng, có thể thấy rằng cả ba mô hình ResNet-50, EfficientNet-B0 và Swin Transformer đều đạt hiệu quả cao trong bài toán phát hiện ảnh Deepfake. Tuy nhiên, xét trên các tiêu chí toàn diện như độ chính xác, F1-score và đặc biệt là khả năng giảm thiểu sai sót trong ma trận nhầm lẫn, EfficientNet-B0 là mô hình tối ưu nhất. Mô hình này không chỉ đạt độ chính xác và F1-score cao nhất mà còn có số lượng ảnh Deepfake bị nhầm là ảnh thật thấp nhất, đồng thời giữ mức sai lệch tổng thể nhỏ hơn so với các mô hình còn lại. Đây là yếu tố then chốt trong các hệ thống phát hiện nội dung giả mạo, nơi việc bỏ sót Deepfake (False Negative) có thể dẫn đến hậu quả nghiêm trọng. Tuy nhiên, kết quả cũng cho thấy rằng chưa có mô hình nào đạt hiệu suất tuyệt đối, đặc biệt là trong việc duy trì độ nhạy (recall) và hạn chế False positive ở mức tối thiểu.

Hướng phát triển tiếp

Mặc dù đề án đã đạt được một số kết quả khả quan, tuy nhiên vẫn còn một số vấn đề cần tiếp tục giải quyết trong thời gian tới như sau:

- Tăng cường kỹ thuật tiền xử lý ảnh, đặc biệt là các phương pháp làm nổi bật chi tiết khuôn mặt hoặc phát hiện hiện vật bất thường trong ảnh Deepfake.
- Kết hợp nhiều mô hình hoặc sử dụng kiến trúc lai giữa CNNs và Transformer để khai thác ưu điểm của từng loại mô hình.
- Phân tích sâu về lỗi sai, đặc biệt là dương tính giả đối với ảnh thật, nhằm giảm thiểu các cảnh báo sai trong ứng dụng thực tiễn.
- Khám phá thêm các tập dữ liệu mới, có tính đa dạng cao hơn về nguồn ảnh, phong cách sinh ảnh (GANs, Diffusion), để nâng cao tính tổng quát hóa cho mô hình.

Tóm lại, EfficientNet-B0 là lựa chọn tối ưu trong bối cảnh nghiên cứu hiện tại, nhưng việc cải tiến mô hình và mở rộng phạm vi thử nghiệm là cần thiết để hướng tới các hệ thống phát hiện ảnh Deepfake có độ tin cậy cao hơn trong môi trường thực tế.

TÀI LIỆU THAM KHẢO

[1] Abdullah, S. M., Cheruvu, A., Kanchi, S., Chung, T., Gao, P., Jadliwala, M., & Viswanath, B. (2024, May). An analysis of recent advances in deepfake image detection in an evolving threat landscape. In 2024 IEEE Symposium on Security and Privacy (SP) (pp. 91-109). IEEE.

[2] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A Compact Facial Video Forgery Detection Network. IEEE WIFS. https://arxiv.org/abs/1809.00888

[3] Al-Adwan, A., Alazzam, H., Al-Anbaki, N., & Alduweib, E. (2024). Detection of Deepfake Media Using a Hybrid CNN–RNN Model and Particle Swarm Optimization (PSO) Algorithm. Computers, 13(4), 99. https://doi.org/10.3390/computers13040099].

[4] Alsharif, M. H., & Naser, M. A. (2024). The role of AI in medical imaging: Advances and opportunities. Discover Artificial Intelligence, 4(1). https://doi.org/10.1007/s44204-024-00013-2.

[5] Ahmed Khan, S., & Dang-Nguyen, D.-T. (2023). Deepfake Detection: A Comparative Analysis. arXiv preprint arXiv:2308.03471. https://arxiv.org/abs/2308.03471.

[6] T. Ahmed, N. H. Nuri SababAuthors (2022), Classification and Understanding of Cloud Structures via Satellite Images with EfficientUNet, SN Computer Science, Volume 3, Issue 1, https://doi.org/10.1007/s42979-021-00981-2, Jan. 2022.

[7] Anan, K., Bhattacharjee, A., Intesher, A., Islam, K., Fuad, A. A., Saha, U., & Imtiaz, H. (2025). Hybrid Deepfake Image Detection: A Comprehensive Dataset-Driven Approach Integrating Convolutional and Attention Mechanisms with Frequency Domain Features. arXiv preprint arXiv:2502.10682.].

[8] Ba, Z., Liu, Q., Liu, Z., Wu, S., Lin, F., Lu, L., & Ren, K. (2024, March). Exposing the deception: Uncovering more forgery clues for deepfake detection. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 2, pp. 719-728).].

[9] Bertasius et al., "Is Space-Time Attention All You Need for Video Understanding?", ICML 2021].

[10] Carlini et al., "Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods", AISec 2017].

[11] Carion et al., "End-to-End Object Detection with Transformers", ECCV 2020

[12] Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., ... & Etzioni, O. (2025). Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024. arXiv preprint arXiv:2503.02857].

[13] Chalapathy, R., & Chawla, S. (2019). Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407. https://arxiv.org/abs/1901.03407.

[14] Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., et al. (2018), "Opportunities and obstacles for deep learning in biology and medicine", Journal of The Royal Society Interface, 15(141), 20170387. https://doi.org/10.1098/rsif.2017.0387

[15] Chollet, F. (2017), Xception: Deep Learning with Depthwise Separable Convolutions, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1800-1807.

[16] Choudhary, K., DeCost, B., Chen, C., Jain, et.al. (2022), "Recent advances and applications of deep learning methods in materials science", npj Computational Materials, 8(1), 59.

[17] Choy, G., Khalilzadeh, O., Michalski, M., Do, S., Samir, A. E., Pianykh, O. S., ... & Dreyer, K. J. (2018). Current applications and future impact of machine learning in radiology. Radiology, 288(2), 318–328. https://doi.org/10.1148/radiol.2018171820.

[18] Chugh et al., "Not Made for Each Other – Audio-Visual Dissonance-based DeepFake Detection and Localization", CVPRW 2020.

[19] [CyberLink. (n.d.). Facial recognition for security, surveillance & access control. Retrieved May 9, 2025, from https://membership.cyberlink.com/faceme/insights/articles/1214/facial-recognition-for-security-surveillance-access-control./

[20] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT (pp. 4171–4186). https://arxiv.org/abs/1810.04805

[21] Dolhansky, B., Bitton, J., Pflaum, B., et al. (2020). The DeepFake Detection Challenge (DFDC) Dataset. arXiv preprint arXiv:2006.07397.

[22] Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", ICLR 2021.].

[23] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017), "Dermatologist-level classification of skin cancer with deep neural networks", Nature, 542(7639), 115–118. https://doi.org/10.1038/nature21056.

[24] Fielddrive. (2021). Artificial intelligence and face recognition: The basics. Retrieved May 9, 2025, from https://www.fielddrive.com/blog/artificial-intelligence-face-recognition-basics].

[25] Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image Style Transfer Using Convolutional Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2414-2423.].

[26] Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems, 27. https://papers.nips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

[27] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems, 27, 2672-2680..

[28] Google Cloud. (n.d.). OCR: Extract text from images with AI. Retrieved May 9, 2025, from https://cloud.google.com/use-cases/ocr.

[29] Graesser, L., & Keng, W. L. (2019). Foundations of deep reinforcement learning: theory and practice in Python. Addison-Wesley Professional.

[30] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2018 15th IEEE International Conference on Advanced Video and

Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6, doi: 10.1109/AVSS.2018.8639].

[31] Guarnera, L., Giudice, O., & Battiato, S. (2020). Fighting deepfake by exposing the convolutional traces on images. IEEE access, 8, 165085-165098.].

[32] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[33] Janai, J., Güney, F., Behl, A., & Geiger, A. (2020). Computer vision for autonomous vehicles: Problems, datasets and state of the art. Foundations and Trends® in Computer Graphics and Vision, 12(1–3), 1–308. https://doi.org/10.1561/0600000079.

[34] Kendall, A., Hawke, J., Janz, D., et al. (2019). Learning to drive in a day. In International Conference on Robotics and Automation (ICRA). https://arxiv.org/abs/1807.00412

[35] Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114].

[36] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8), 30–37. https://doi.org/10.1109/MC.2009.263.

[37] Kuo, R. J., Lin, Y., & Chang, C. (2024). Deep learning applications in modern supply chain optimization: A review. Journal of Industrial and Production Engineering, 41(2), 87–102. https://doi.org/10.1016/j.jipe.2024.03.014].

[38] Lai Minh Tuan*, Pham Tien Manh, Dong Thi Thuy Linh, Deepfake detection based on deep learning, TNU Journal of Science and Technology, 228(15): 88 – 95].

[39] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. Nature, 521(7553), 436-444.

[40] Le, B. M., Kim, J., Tariq, S., Moore, K., Abuadbba, A., & Woo, S. S. (2024). Sok: Facial deepfake detectors. arXiv preprint arXiv:2401.04364].

[41] Lee, J., Bagheri, B., & Kao, H. A. (2015). A cyber-physical systems architecture for industry 4.0-based manufacturing systems. Manufacturing Letters, 3, 18–23. https://doi.org/10.1016/j.mfglet.2014.12.001.

[42] Li et al., "Celeb-DF: A New Dataset for DeepFake Detection", CVPR 2020].

[43] Li et al., "Face X-ray for More General Face Forgery Detection", CVPR 2020].

[44] Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows", In Proceedings of the IEEE/CVF international conference on computer vision ICCV 2021, pp. 10012-10022.

[45] McCarthy, J. (2007), "What is artificial intelligence?", Stanford University. Retrieved from https://www-formal.stanford.edu/jmc/whatisai.pdf].

[46] Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2016). Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics, 19(6), 1236–1246. https://doi.org/10.1093/bib/bbx044.

[47] Microsoft. (n.d.). Overview of OCR in Azure AI services. Retrieved May 9, 2025, from https://learn.microsoft.com/en-us/azure/ai-services/computer-vision/overview-ocr].

[48] Midjourney. (n.d.). Midjourney. Retrieved May 12, 2025, from https://www.midjourney.com/].

[49] Mordvintsev, A., Olah, C., & Tyka, M. (2015). Inceptionism: Going Deeper into Neural Networks. Google Research Blog].

[50] Nguyen, G., Dlugolinsky, S., Bobák, M., et al. (2021). Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. Artificial Intelligence Review, 54(7), 5213–5253. https://doi.org/10.1007/s10462-020-09841-6

[51] OpenAI. (n.d.). DALL·E 2. Retrieved May 12, 2025, from https://openai.com/index/dall-e-2/].

[52] Patel, Y., Tanwar, S., Bhattacharya, P., Gupta, R., Alsuwian, T., Davidson, I. E., & Mazibuko, T. F. (2023). An improved dense CNN architecture for deepfake image detection. IEEE Access, 11, 22081-22095.

[53] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

[54] Rafique, R., Gantassi, R., Amin, R. et al. Deep fake detection and classification using error-level analysis and deep learning. Sci Rep 13, 7422 (2023). https://doi.org/10.1038/s41598-023-34629-3]

[55] Ramesh, A., Pavlov, M., Goh, G., et al. (2022). Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125. https://arxiv.org/abs/2204.06125

[56] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. arXiv preprint arXiv:2102.12092.]

[57] Rossler, A., Cozzolino, D., Verdoliva, L., et al. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. ICCV 2019. https://arxiv.org/abs/1901.08971

[58] Sallab, A. E., Abdou, M., Perot, E., & Yogamani, S. (2017). Deep reinforcement learning framework for autonomous driving. Electronic Imaging, 2017(19), 70–76. https://doi.org/10.2352/ISSN.2470-1173.2017.19.AVM-023

[59] Sharma, A., Saini, N., & Makkar, A. (2021). Strengthening autonomous vehicle safety with a deep learning-based object detection system. Autonomous Vehicle International. Retrieved from https://www.autonomousvehicleinternational.com/features/strengthening-autonomous-vehicle-safety-with-a-deep-learning-based-object-detection-system.html.

[60] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information processing & management, 45(4), 427-437.

[61] Soudy, A.H., Sayed, O., Tag-Elser, H. et al. (2024), Deepfake detection using convolutional vision transformers and convolutional neural networks. Neural Comput & Applic 36, 19759–19775 (2024). https://doi.org/10.1007/s00521-024-10181-7

[62] Stankov, I. S., & Dulgerov, E. E. (2024, September). Detection of Deepfake Images and Videos Using SVM, CNN, and Hybrid Approaches. In 2024 XXXIII International Scientific Conference Electronics (ET) (pp. 1-5). IEEE.].

[63] Stability AI. (n.d.). Stable Diffusion. https://stability.ai/stable-image].

[64] Sudarsan, A. (2024). Deepfake Characterization, Propagation, and Detection in Social Media - A Synthesis Review. IEEE. https://en.wikipedia.org/wiki/Deepfake].

[65] M. Tan, Quoc V. Le (2019), EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. International Conference on Machine Learning ICML 2019, pp. 6105-6114.

[66] Tao, F., Qi, Q., Liu, A., & Kusiak, A. (2018). Data-driven smart manufacturing. Journal of Manufacturing Systems, 48, 157–169. https://doi.org/10.1016/j.jmsy.2018.01.006.

[67] Thing, V. L. L. (2023). Deepfake Detection with Deep Learning: Convolutional Neural Networks versus Transformers. arXiv preprint arXiv:2304.03698. https://arxiv.org/abs/2304.03698].

[68] Tiwari, A., Dave, R., & Vanamala, M. (2023). Leveraging Deep Learning Approaches for Deepfake Detection: A Review. arXiv preprint arXiv:2304.01908. https://arxiv.org/abs/2304.01908].

[69] Tirkolaee, E. B., Goli, A., Dashtian, M. A., & Sadeghi, S. (2022). A systematic literature review on applications of deep learning in supply chain management. Artificial Intelligence Review, 55, 2041–2071. https://doi.org/10.1007/s10462-022-10289-z.

[70] X. Tong, L. Wang, X. Pan and J. g. Wang, "An Overview of Deepfake: The Sword of Damocles in AI," 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL), Chongqing, China, 2020, pp. 265-273, doi: 10.1109/CVIDL51233.2020.00-88/

[71] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1706.03762

[72] Verdoliva, L. (2020). Media Forensics and DeepFakes: An Overview. IEEE Journal of Selected Topics in Signal Processing, 14(5), 910–932. https://doi.org/10.1109/JSTSP.2020.3002103

[73] Verdoliva, L. (2020). Media Forensics and DeepFakes: An Overview. IEEE Journal of Selected Topics in Signal Processing, 14(5), 910–932. https://doi.org/10.1109/JSTSP.2020.3002103].

[74] Wang, S., Xia, X., Ye, L., & Yang, B. (2021). Automatic detection and] of steel surface defect using deep convolutional neural networks. Metals, 11(3), 388.]

[75] Wang, S. Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). CNN-generated images are surprisingly easy to spot... for now. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8695–8704. https://arxiv.org/abs/1912.11035].

[76] S. Waseem, S. A. R. S. Abu Bakar, B. A. Ahmed, Z. Omar, T. A. E. Eisa and M. E. E. Dalam, "DeepFake on Face and Expression Swap: A Review," in IEEE Access, vol. 11, pp. 117865-117906, 2023, doi: 10.1109/ACCESS.2023.3324403].

[77] Wodajo, D., & Atnafu, S. (2021). Deepfake Video Detection Using Convolutional Vision Transformer. arXiv preprint arXiv:2102.11126. https://arxiv.org/abs/2102.11126].

[78] Y. Zhang, Q. Li, Z. Yu, L.Shen (2025), Distilled transformers with locally enhanced global representations for face forgery detection, Pattern Recognition Journal, 05-2025, DOI: 10.1016/j.patcog.2024.111253.

[79] Zhao, Z., Chuah, J. H., Chow, C. O., Xia, K., Tee, Y. K., Hum, Y. C., & Lai, K. W. (2024). Machine learning approaches in comparative studies for Alzheimer's diagnosis using 2D MRI slices. Turkish Journal of Electrical Engineering and Computer Sciences, 32(1), 93-107.

[80] Zignuts. (n.d.). Artbreeder: AI Image Creation through Collaborative Genetic Editing. Retrieved May 12, 2025, from https://www.zignuts.com/ai/artbreeder.

[81] Zobaed, S. et al. (2021). DeepFakes: Detecting Forged and Synthetic Media Content Using Machine Learning. In: Montasari, R., Jahankhani, H. (eds) Artificial

Intelligence in Cyber Security: Impact and Implications. Advanced Sciences and Technologies for Security Applications. Springer, Cham. https://doi.org/10.1007/978-3-030-88040-8_7, pp.177-201.

[82]https://vietnamnet.vn/hacker-mu-trang-viet-phat-trien-cong-cu-phat-hien-deepfake-2166061.html.

[83]The Laotian Times, (2025), X-PHY Inc Unveils Real-Time Deepfake Detection Tool Ahead of RSA Conference 2025 - Laotian Times]. https://laotiantimes.com/2025/04/24/x-phy-inc-unveils-real-time-deepfake-detection-tool-ahead-of-rsa-conference-2025/

[84] The CIFAR-10 dataset, https://www.cs.toronto.edu/~kriz/cifar.html

[85] DeepArt. https://www.deeparteffects.com/

KiểmTraTàiLiệu

BÁO CÁO KIỂM TRA TRÙNG LẶP

Thông tin tài liệu

Tên tài liệu:	Vilayvone Phimsipasom-Nghiencuu,ungdungphuongphaphocsauvaophathienanhDeepfake_OFFICIAL(final)
Tác giả:	Vilayvone Phimsipasom
Điểm trùng lặp:	5
Thời gian tải lên:	19:03 09/06/2025
Thời gian sinh báo cáo:	19:09 09/06/2025
Các trang kiểm tra:	77/77 trang

Kết quả kiểm tra trùng lặp

Có **5%** nội dung trùng lặp

Có **93%** nội dung không trùng lặp

Có **2%** nội dung người dùng loại trừ

Có **0%** nội dung hệ thống bỏ qua

Nguồn trùng lặp tiêu biểu

arxiv.org vjol.info.vn www.mdpi.com

Người hướng dẫn	Tác giả thực hiện
(ký tên)	*(ký tên)*
PGS.TSKH. Hoàng Đăng Hải	Vilayvone Phimsipasom

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BÁO CÁO GIẢI TRÌNH
SỬA CHỮA, HOÀN THIỆN ĐỀ ÁN TỐT NGHIỆP

Họ và tên học viên: Vilayvone Phimsipasom

Chuyên ngành: KHMT

Khóa: 2023 đợt 2

Tên đề tài: Nghiên cứu, ứng dụng phương pháp học sâu vào phát hiện ảnh Deepfake

Người hướng dẫn khoa học: PGS.TSKH. Hoàng Đăng Hải

Ngày bảo vệ: 19/07/2025

Các nội dung học viên đã sửa chữa, bổ sung trong đề án tốt nghiệp theo ý kiến đóng góp của Hội đồng chấm đề án tốt nghiệp:

TT	Ý kiến hội đồng	Sửa chữa của học viên
1	Bổ sung các mô tả về tham số của mô hình (learning rate, batch size)	Tiếp thu góp ý của Hội đồng, học viên đã kiểm tra lại nội dung trong đề án. Các mô tả về tham số của mô hình (learning rate, batch size) đã được trình bày trong Chương 3, mục 3.6.4.
2	Chỉnh sửa lại các lỗi chính tả	Tiếp thu góp ý của Hội đồng, học viên đã rà soát, chỉnh sửa các lỗi soạn thảo, các lỗi ngữ pháp của toàn bộ đề án tốt nghiệp.
3	Phân tích thêm về kết quả thực nghiệm của một số mô hình học sâu	Tiếp thu góp ý của Hội đồng, học viên đã bổ sung thêm các phân tích về kết quả thực nghiệm của các mô hình học sâu thử nghiệm, trình bày tại Chương 3, mục 3.7.

Hà Nội, ngày 29 tháng 07 năm 2025

Ký xác nhận của

CHỦ TỊCH HỘI ĐỒNG CHẤM ĐỀ ÁN	THƯ KÝ HỘI ĐỒNG	NGƯỜI HƯỚNG DẪN KHOA HỌC	HỌC VIÊN
PGS.TS. Trần Quang Anh	TS. Đào Thị Thúy Quỳnh	PGS.TSKH. Hoàng Đăng Hải	Vilayvone Phimsipasom

BỘ KHOA HỌC VÀ CÔNG NGHỆ
HỌC VIỆN CÔNG NGHỆ
BƯU CHÍNH VIỄN THÔNG

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BIÊN BẢN
HỌP HỘI ĐỒNG CHẤM ĐỀ ÁN TỐT NGHIỆP THẠC SĨ

Căn cứ quyết định số Quyết định số 1098/QĐ-HV ngày 26 tháng 06 năm 2025 của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ. Hội đồng đã họp vào hồi ..10.. giờ ..40.. phút, ngày 19 tháng 07 năm 2025 tại Học viện Công nghệ Bưu chính Viễn thông để chấm đề án tốt nghiệp thạc sĩ cho:

Học viên: **Vilayvone Phimsipasom**

Tên đề án tốt nghiệp: **Nghiên cứu, ứng dụng phương pháp học sâu vào phát hiện ảnh DEEPFAKE**

Chuyên ngành: **Khoa học máy tính** Mã số: **8480101**

Các thành viên của Hội đồng chấm đề án tốt nghiệp có mặt: ..09../05

TT	HỌ VÀ TÊN	TRÁCH NHIỆM TRONG HĐ	GHI CHÚ
1	**PGS.TS. Trần Quang Anh**	Chủ tịch	
2	**TS. Đào Thị Thúy Quỳnh**	Thư ký	
3	**TS. Trần Đăng Công**	Phản biện 1	
4	**PGS.TS. Nguyễn Hà Nam**	Phản biện 2	
5	**PGS.TS. Nguyễn Trọng Khánh**	Uỷ viên	

Các nội dung thực hiện:

1. Chủ tịch Hội đồng điều khiển buổi họp. Công bố quyết định của Giám đốc Học viện Công nghệ Bưu chính Viễn thông về việc thành lập Hội đồng chấm đề án tốt nghiệp thạc sĩ.
2. Người hướng dẫn khoa học hoặc thư ký đọc lý lịch khoa học và các điều kiện bảo vệ đề án tốt nghiệp của học viên. (có bản lý lịch khoa học và kết quả các môn học cao học của học viên kèm theo).
3. Học viên trình bày tóm tắt đề án tốt nghiệp.
4. Phản biện 1 đọc nhận xét (có văn bản kèm theo)
5. Phản biện 2 đọc nhận xét (có văn bản kèm theo)
6. Các câu hỏi của thành viên Hội đồng:

- Vì sao em chọn Xception cho bài toán này? So với Efficientnet, Xception có ưu điểm gì?

- Tại sao lại chọn tham số Swin Transformer và Efficientnet? Có giải thích nào khiến cho kết quả tốt hơn đó?

7. Trả lời của học viên:

- Học viên chạy Resnet 50 và mô tả mô hình đem lại kết quả trong thị giác máy tính

- Học viên tiếp thu và chỉnh sửa theo yêu cầu của phản biện

8. Thư ký đọc nhận xét về quá trình thực hiện đề án tốt nghiệp của học viên *(có văn bản kèm theo).*

9. Hội đồng họp riêng:

- Bầu Ban kiểm phiếu:
  1. Trưởng Ban kiểm phiếu: TS Đào Thị Thúy Quỳnh
  2. Uỷ viên Ban kiểm phiếu: PGS.TS Nguyễn Hà Nam
  3. Uỷ viên Ban kiểm phiếu: TS. Phan Đăng Cầu
- Hội đồng chấm đề án tốt nghiệp bằng bỏ phiếu kín.
- Ban kiểm phiếu làm việc:
- Trưởng Ban kiểm phiếu báo cáo kết quả kiểm phiếu *(có Biên bản họp Ban kiểm phiếu kèm theo)*
- **Điểm trung bình của đề án tốt nghiệp:** 8,5

Kết luận:

1. Các nội dung cần chỉnh sửa, hoàn thiện sau bảo vệ đề án tốt nghiệp:

- Bổ sung các mô tả về các tham số của mô hình (learning rate, batch size)
- Chỉnh sửa lại các lỗi chính tả
- Phân tích thêm về kết quả thực nghiệm của một số mô hình học sâu

2. Đề nghị Học viện công nhận (hoặc không) và cấp bằng (hoặc không) thạc sĩ cho học viên:

3. Đề án tốt nghiệp có thể phát triển thành đề tài nghiên cứu cho NCS……………………………

Buổi làm việc kết thúc vào…………… cùng ngày.

Chủ tịch	Thư ký
PGS.TS. Trần Quang Anh	TS. Đào Thị Thúy Quỳnh

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập - Tự do - Hạnh phúc

BẢN NHẬN XÉT LUẬN VĂN TỐT NGHIỆP THẠC SĨ

(Dùng cho người phản biện)

Tên đề tài luận văn: Nghiên cứu, ứng dụng phương pháp học sâu vào phát hiện ảnh Deepface
Chuyên ngành: Khoa học máy tính Mã số:8.48.01.01
Tên học viên: Vilayvone Phimsipasom
Họ và tên người nhận xét: Nguyễn Hà Nam.....
Học hàm, học vị: PGS. TS Chuyên ngành: CNTT .....
Cơ quan công tác: Ban Khoa học và Đổi mới sáng tạo, ĐHQGHN.....

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:
Đề tài lựa chọn nghiên cứu hướng tới một trong những vấn đề mang tính thời sự và cấp thiết trong lĩnh vực **AI – Computer Vision**, cụ thể là **phát hiện ảnh Deepfake (giả mạo khuôn mặt bằng AI)**. Trong bối cảnh hiện nay, sự phát triển mạnh mẽ của các công nghệ tạo sinh như **DeepFaceLab, FaceSwap, GAN (Generative Adversarial Networks)** đã khiến việc tạo ra các video, ảnh giả mạo khuôn mặt ngày càng tinh vi, gây ra nhiều hệ lụy về **an ninh mạng, pháp lý, truyền thông và quyền riêng tư cá nhân**.Vì vậy, việc nghiên cứu và ứng dụng **học sâu (Deep Learning)** để phát hiện ảnh Deepface có ý nghĩa **thực tiễn cao**, đóng góp vào việc xây dựng các hệ thống bảo mật, kiểm duyệt nội dung số.

II/ Về nội dung, chất lượng của luận văn, các kết quả đã đạt được (so với đề cương đã được duyệt):
Nội dung của luận văn và các kết quả đạt được cơ bản bám sát theo đề cương đã được phê duyệt.
Luận văn được trình bày trong 3 chương bao gồm giới thiệu tổng quan, nghiên cứu giải pháp ứng dụng học sâu trong phát hiện ảnh giả mạo và cuối cùng triển khai thử nghiệm đánh giá mô hình trên tập dữ liệu CIFAKE và đã cho ra một số kế quả tốt.

III/ Những vấn đề cần giải thích thêm:
- Tại sao lại chọn Resnet-50, Swin Transformer và EfficientNet? Có mô hình nào hiệu quả hơn được sử dụng trog thực tế?
-

IV/ Kết luận:
Luận văn đáp ứng được yêu cầu cơ bản của luận văn thạc sỹ chuyên ngành KHMT(theo định hướng ứng dụng). Tôi đồng ý để học viên được bảo vệ luận văn trước Hội đồng chấm luận văn thạc sĩ

Ngày 15 tháng 7 năm 2025
NGƯỜI NHẬN XÉT
(Ký và ghi rõ họ tên)

Nguyễn Hà Nam

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

BẢN NHẬN XÉT ĐỀ ÁN TỐT NGHIỆP THẠC SĨ
(Dùng cho người phản biện)

Tên đề tài đề án tốt nghiệp: **Nghiên cứu, ứng dụng phương pháp học sâu vào phát hiện ảnh Deepfake**

Chuyên ngành: Khoa học máy tính

Mã chuyên ngành: 8.48.01.01

Họ và tên học viên: **Vilayvone Phimsipasom**

Họ và tên người nhận xét: Trần Đăng Công

Học hàm, học vị: Tiến sĩ

Chuyên ngành: Khoa học máy tính

Cơ quan công tác: Đại học Đại Nam

Số điện thoại: 0964981451 E-mail: congtd@dainam.edu.vn

NỘI DUNG NHẬN XÉT

I/ Cơ sở khoa học và thực tiễn, tính cấp thiết của đề tài:

- Đề tài có tính thực tiễn cao, tác giả đã trình bày được từ cơ sở lý thuyết đến kết quả thí nghiệm về việc phát hiện ảnh deepfake dựa trên kỹ thuật học sâu.

- Tác giả đã trình bày được một số mô hình học sâu cơ bản áp dụng phát hiện ảnh deepfake như DTN, mô hình kết hợp CNN và Vision Transformer.

.....

II/ Nội dung của đề án tốt nghiệp, các kết quả đã đạt được:

- Tác giả đã trình bày được các bước từ thu thập dữ liệu, xử lý, làm sạch và xây dựng mô hình, đánh giá.

- Tác giả đã thử nghiệm, phân tích, so sánh kết quả giữa 03 mô hình: Resnet-50, Swin Transformer và EfficientNet, từ đó có những đánh giá.

- Báo cáo gồm 3 chương, 70 trang.

.....

.....

III/ Những vấn đề cần giải thích thêm:

- Hãy phân tích, làm rõ việc so sánh hiệu ả giữa 2 mô hình dựa trên các tiêu chí đánh giá hiệu năng (Accuracy, Precision, Reca ', F1-Score).

.....

.....

IV/ Kết luận:

Đề án tốt nghiệp đã được giải quyết với kết quả cơ bản, sử dụng các mô hình cơ bản, thư viện sẵn để giải quyết.

Tuy nhiên, báo cáo trình bày các mô hình, cũng như phân tích cấu trúc, điều chỉnh tham số các mô hình được áp dụng còn chưa chi tiết.

Đồng ý cho phép học viên bảo vệ đề án tốt nghiệp.

Ngày 15 tháng 7 năm 2025

NGƯỜI NHẬN XÉT

TS. Trần Đăng Công