GIẢI PHÁP NHẬN DIỆN LOÀI CHIM NGUY CẤP, QUÝ, HIẾM DỰA TRÊN HỌC SÂU

LÊ HẢI HÀ[1], MẠC THỊ MINH TRÀ[2], LÊ HOÀNG ANH[2], NGUYỄN VĂN THÙY[2]
[1]Viện Toán ứng dụng và Tin học, Đại học Bách khoa Hà Nội
[2]Trung tâm Điều tra, Quan trắc Đa dạng sinh học

Tóm tắt
Nhận dạng chính xác các loài chim nguy cấp, quý, hiếm đóng vai trò thiết yếu trong giám sát và bảo tồn đa dạng sinh học (ĐDSH). Nghiên cứu nhằm xây dựng mô hình nhận diện tin cậy cho 26 loài ưu tiên bảo vệ bằng cách tích hợp học sâu với các cơ chế xử lý đặc thù. Phương pháp đề xuất gồm 3 hợp phần: Tiền xử lý bằng YOLOv12 để loại bỏ ảnh không chứa chim; phân loại loài dựa trên ResNet học chuyển giao và tinh chỉnh trên bộ dữ liệu riêng (27.457 ảnh của 171 loài); hậu xử lý sử dụng OpenMax nhằm giảm nhầm lẫn đối với các loài ngoài tập huấn luyện. Kết quả thực nghiệm cho thấy, mô hình đạt chỉ số Macro F1-Score là 83,54%, chứng minh tiềm năng ứng dụng cao trong hệ thống giám sát ĐDSH.
Từ khóa: *Nhận diện loài, học sâu, ResNet, phân loại tập mở, OpenMax.*
Ngày nhận bài: *7/10/2025****; Ngày sửa chữa:*** *10/11/2025****; Ngày duyệt đăng:*** *17/11/2025.*

A deep learning-based solution for identifying endangered and rare bird species

Abstract
Accurate identification of endangered, rare, and precious bird species plays an essential role in biodiversity monitoring and conservation. This study aims to develop a reliable identification model for 26 priority conservation species by integrating deep learning with specific processing mechanisms. The proposed methodology consists of three components: pre-processing using YOLOv12 to discard non-bird images; species classification based on a transfer learning and fine-tuning ResNet model on a in-house dataset (27,457 images of 171 species); and post-processing using OpenMax to reduce confusion with species outside the training set. Experimental results show that the model achieves a Macro F1-Score of 83.54%, demonstrating its strong potential for application in biodiversity monitoring systems.
Keywords: *species identification, deep learning, ResNet, open-set classification, OpenMax.*
JEL Classifications: *O13, O44, P18.*

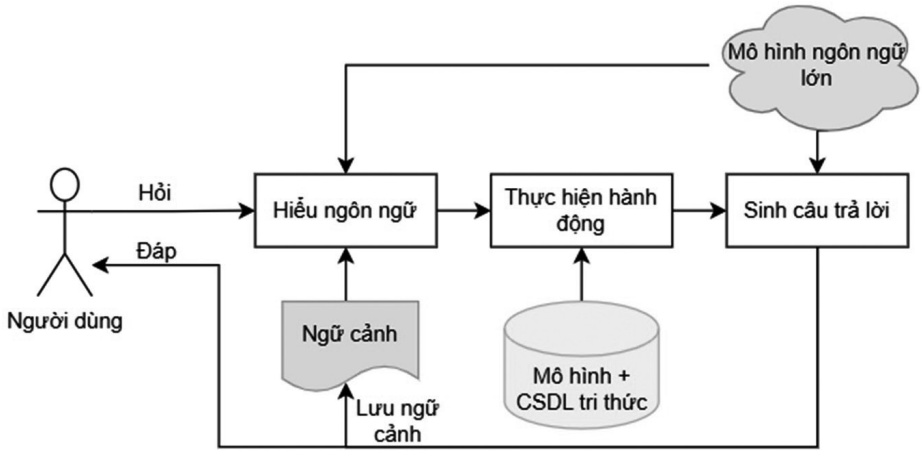
1. ĐẶT VẤN ĐỀ

Nhận diện các loài sinh vật là một trong những bài toán cơ bản và giữ vai trò then chốt trong công tác điều tra, giám sát và bảo tồn ĐDSH. Đối với các loài nguy cấp, quý, hiếm được ưu tiên bảo vệ, việc nhận diện chính xác không chỉ hỗ trợ các nhà nghiên cứu trong phân loại học mà còn giúp nhanh chóng, kịp thời bảo vệ loài cũng như cung cấp dữ liệu phục vụ xây dựng chính sách, kế hoạch quản lý tài nguyên thiên nhiên. Tuy nhiên, quá trình nhận diện gặp nhiều thách thức do điều kiện thu thập dữ liệu, sự tương đồng về hình thái giữa các loài, sự thay đổi theo giai đoạn phát triển, giới tính, mùa sinh sản, hay điều kiện môi trường và quan sát khác nhau.

Các phương pháp nhận diện truyền thống dựa vào đặc điểm hình thái và kinh nghiệm chuyên gia tuy có độ tin cậy cao trong phạm vi hẹp nhưng hạn chế về thời gian, chi phí và khả năng mở rộng khi áp dụng trên quy mô lớn. Với sự phát triển mạnh của học máy và học sâu, việc nhận diện hay giám sát ĐDSH đã có những bước tiến vượt bậc. Trên thế giới, Tuia và cộng sự [1] đã nhấn mạnh vai trò trung tâm của các mô hình học sâu trong giám sát ĐDSH, đặc biệt khi các mô hình này được kết hợp với cảm biến tự động và dữ liệu khoa học công dân (Citizen Science).

Riêng với chim, hướng nghiên cứu nhận diện loài dựa trên hình ảnh đã đạt nhiều bước tiến nhờ sự ra đời của các bộ dữ liệu phân loại chi tiết (fine-grained) chuẩn mực như CUB-200-2011 [2], Birdsnap [3], hay các dự án quy mô lớn được xây dựng với sự đóng góp của cộng đồng như sản phẩm của Van Horn và cộng sự [4]. Sự ra đời của các bộ dữ liệu quy mô lớn như LaSBiRD [5], tiếp tục mở rộng phạm vi và độ đa dạng dữ liệu, tạo điều kiện đánh giá mô hình gần hơn với điều kiện thực địa.

Tại Việt Nam, các nghiên cứu ứng dụng trí tuệ nhân tạo trong nhận diện loài mới xuất hiện trong những năm gần đây, nhưng chủ yếu tập trung vào các nhóm



Hình 1. Kiến trúc của một hệ thống chatbot nhận diện loài

đối tượng khác thay vì các loài chim nguy cấp. Có thể kể đến phần mềm nhận dạng nhanh động, thực vật nguy cấp, quý, hiếm phục vụ bảo vệ rừng tại Thanh Hóa [6], hệ thống nhận diện sinh vật gây hại trên lúa của Cục Bảo vệ thực vật [7], hay ứng dụng nhận biết nhanh các nhóm gỗ bằng điện thoại thông minh của Trường Đại học Lâm nghiệp [8]. Dù đóng góp quan trọng về mặt ứng dụng thực tiễn, các nghiên cứu này chủ yếu tập trung vào các bài toán phân loại tổng quát và chưa đề cập đến bài toán nhận diện chi tiết, cũng như chưa tích hợp cơ chế phân loại tập mở để xử lý các trường hợp ngoài tập huấn luyện.

Sự phát triển của trí tuệ nhân tạo trong những năm gần đây, nổi bật là các mô hình nền tảng như mô hình ngôn ngữ lớn, đã giúp máy tính đạt được khả năng hiểu và sinh tạo ngôn ngữ tự nhiên như con người. Điều này đã thúc đẩy sự ra đời của một thế hệ ứng dụng AI đột phá. Điển hình trong các ứng dụng trí tuệ nhân tạo là ứng dụng trợ lý ảo hay hội thoại trực tuyến (chatbot) cho phép người dùng tương tác, cung cấp thông tin quan sát hay đo đạc được và máy tính với các mô hình và các cơ sở dữ liệu tri thức có thể hỗ trợ xác định loài. Kiến trúc cơ bản của một hệ thống chatbot nhận diện loài như vậy có thể được thể hiện trong Hình 1.

Mô hình ngôn ngữ lớn cùng ngữ cảnh hội thoại cho phép hệ thống hiểu ngôn ngữ người dùng, hiểu ý định người dùng mong muốn xác định loài trong một ảnh đưa vào hay một vài câu mô tả đặc điểm hình thái. Từ hiểu ý định dẫn tới có thể xác định được hệ thống cần thực hiện hành động nào và thực thực hiện hành động đó dựa vào dữ liệu, mô hình hay cơ sở dữ liệu tri thức có trước của hệ thống. Cũng chính mô hình ngôn ngữ lớn sẽ giúp hệ thống sinh ra các câu trả lời tự nhiên và có phong cách đến với người dùng. Hệ thống sẽ trả lời kiểu như "Rất cảm ơn bạn đã cho tôi biết một bức ảnh đẹp nhưng có vẻ bạn đã đưa nhầm ảnh vì bức ảnh đó đang chụp một chú mèo đáng yêu chứ tôi không phát hiện được loài chim nào xuất hiện trong ảnh. Xin vui lòng cung cấp một bức ảnh khác" thay cho các câu trả lời cứng nhắc như "Không biết".

Trong các tiếp cận nhận diện từ hình ảnh, âm thanh hay mô tả, nhận diện từ ảnh được xem là trung tâm của quy trình nhận dạng loài, bởi hình ảnh chứa đựng hầu hết các đặc trưng hình thái trực quan như hình dáng, kích thước, màu sắc, hoa văn, tư thế hay cấu trúc. Câu thành ngữ "trăm nghe không bằng một thấy" thật đúng trong trường hợp này, hình ảnh cung cấp thông tin trực quan có sức phân biệt mạnh mẽ, là nguồn dữ liệu chủ đạo để hệ thống hay cụ thể hơn là các mô hình học sâu rút trích đặc trưng và từ đó xác định loài một cách hiệu quả. Với khung ứng dụng dạng chatbot, mô hình nhận diện loài bằng hình ảnh nên được thiết lập để có độ phủ (recall) lớn hay âm tính giả (false negative) nhỏ bởi sau đó hệ thống có thể yêu cầu người dùng cung cấp thêm thông tin trước khi có các kết luận cuối cùng.

Bài báo trình bày các nghiên cứu và kết quả của nhóm nghiên cứu khi tập trung vào phần xây dựng mô hình nhận diện loài nguy cấp, quý, hiếm được ưu tiên bảo vệ - thử nghiệm đối với lớp chim dựa trên ảnh chụp của chúng. Đến nay, bài toán phân lớp ảnh hay nhận diện loài chim từ ảnh đã đạt được nhiều thành tựu nhờ sự ra đời sự phát triển nhanh và mạnh của học sâu với các kiến trúc mạng nơ-ron tích chập và Transformer cùng các bộ dữ liệu quy mô lớn như ImageNet hay iNaturalist. Tuy nhiên, một thách thức quan trọng của bài toán là khả năng nhận diện các loài chưa từng xuất hiện trong tập huấn luyện. Các mô hình học sâu truyền thống thường có xu hướng gán nhãn sai cho những mẫu nằm ngoài phân bố huấn luyện. Vì vậy, cần có các cơ chế phát hiện và từ chối

dự đoán trong trường hợp không thuộc lớp chim hay không thuộc loài chim nằm trong bộ huấn luyện để từ đó giảm thiểu sai lệch nhận dạng và tăng độ tin cậy tổng thể của hệ thống.

2. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

Đối tượng nghiên cứu trong bài báo là những mô hình học sâu phân lớp ảnh có khả năng phân lớp/nhận diện 26 loài chim nguy cấp, quý, hiếm được ưu tiên bảo vệ theo Nghị định số 64/2019/NĐ-CP ngày 16/7/2019 của Chính phủ về việc sửa đổi Điều 7 Nghị định số 160/2013/NĐ-CP ngày 12/11/2013 của Chính phủ về tiêu chí xác định loài và chế độ quản lý loài thuộc Danh mục loài nguy cấp, quý, hiếm được ưu tiên bảo vệ. 26 loài này gồm: Bồ nông chân xám (Pelecanus philippensis), Cổ rắn (Điêng điểng) (Anhinga melanogaster), Cò trắng trung quốc (Egretta eulophotes), Vạc hoa (Gorsachius magnificus), Cò mỏ thìa (Platalea minor), Quắm cánh xanh (Cò quắm cánh xanh) (Pseudibis davisoni), Quắm lớn (Cò quắm lớn) (Pseudibis gigantea), Già đẫy nhỏ (Leptoptilos javanicus), Hạc cổ trắng (Ciconia episcopus), Ngan cánh trắng (Asarcornis scutulata), Công (Pavo muticus), Gà so cổ hung (Arborophila davidi), Gà lôi lam mào trắng (Lophura edwardsi), Gà lôi tía (Tragopan temminckii), Gà tiền mặt đỏ (Polyplectron germaini), Gà tiền mặt vàng (Polyplectron bicalcaratum), Trĩ sao (Rheinardia ocellata), Sếu đầu đỏ (sếu cổ trụi) (Grus antigone), Ô tác (Houbaropsis bengalensis), Rẽ mỏ thìa (Calidris pygmaea), Choắt mỏ vàng (Tringa guttifer), Niệc nâu (Anorrhinus austeni), Niệc cổ hung (Aceros nipalensis), Niệc mỏ vằn (Rhyticeros undulatus), Hồng hoàng (Buceros bicornis), Khướu ngọc linh (Trochalopteron ngoclinhense).

Mô hình được thiết kế phải đạt được sự cân bằng giữa kích thước nhỏ gọn và hiệu năng xử lý cao. Yêu cầu này nhằm đảm bảo khả năng suy luận nhanh và tiết kiệm tài nguyên tính toán. Nhờ đó, mô hình có thể triển khai linh hoạt trên các hệ thống máy tính có cấu hình thấp mà vẫn duy trì độ chính xác cần thiết. Đồng thời, để đảm bảo khả năng xác định các loài ngoài dữ liệu huấn luyện, mô hình được tiếp cận ở dạng một luồng với 3 thành phần (1) tiền xử lý để đưa ra câu trả lời sớm và các ảnh không chứa lớp chim, (2) mô hình học sâu được huấn luyện trên bộ dữ liệu được thu thập và xử lý cẩn thận của nhóm nghiên cứu và (3) hậu xử lý để phân bổ lại xác suất dự báo các lớp chim trong tập huấn luyện và lớp chim ngoài tập huấn luyện (lớp chưa biết).

2.1. Tiền xử lý

Một trong những phương pháp để sớm có kết luận về ảnh không chứa lớp chim và từ đó tăng khả năng làm việc với các lớp ngoài tập huấn luyện là sử dụng các mô hình nền tảng hay các mô hình được luyện sẵn chuyên biệt đã được luyện trên các bộ dữ liệu lớn để có khả năng tổng quát hóa vượt ngoài tập huấn luyện hay có khả năng nhận biết loài chim mà không cần luyện lại. Ví dụ điển hình về các mô hình nền tảng với quy mô lớn được huấn luyện trên tập dữ liệu đa miền, đa nhiệm và có khả năng tổng quát hóa cao là CLIP, ViT, GPT hay SAM (Segment Anything Model). Tuy nhiên, các mô hình này đều có kích thước lớn và tiêu tốn nhiều tài nguyên tính toán. Chính vì vậy, nhóm nghiên cứu đã lựa chọn mô hình YOLOv12n [9] cho mục đích tiền xử lý này bởi đây là một mô hình kích thước nhỏ, tốc độ cao và vì bản chất của công việc này là loại các ảnh không có lớp chim nếu thực sự chắc chắn nên tham số độ tin cậy phát hiện đối tượng được đặt thấp (0,5). Một điểm nữa là trong giao diện của ứng dụng chatbot cần chỉ rõ vị trí đối tượng trên ảnh nhằm tăng khả năng giải thích kết luận cuối, mô hình YOLOv12n phát hiện đối tượng được chọn cho cả mục đích xác định vị trí loài chim cũng như phân lớp có loài chim hay không.

Đánh giá các chỉ số hiệu năng của mô hình YOLOv12n trên bộ dữ liệu COCO có các kết quả:

Sử dụng kích thước đầu vào 640×640 pixel.

Không phát hiện loài chim *Có loài chim trong ảnh*

Hình 2. Mô hình YOLOv12n luyện sẵn phát hiện loài chim trong ảnh

Đạt mAP_{val}@[.5:.95] = 40.6.

Tốc độ suy luận trên GPU NVIDIA T4 TensorRT là 1.64 ms mỗi ảnh, với 2.6 triệu tham số và 6.5 tỷ phép tính (FLOPs).

Các chỉ số chứng tỏ mô hình có kích thước nhỏ gọn và hiệu năng cao (Hình 2).

2.2. Mô hình học sâu

Để có mô hình học sâu nhận diện 26 loài chim nguy cấp, quý, hiếm được ưu tiên bảo vệ, nhóm nghiên cứu cần thu thập bộ dữ liệu ảnh của 26 loài này. Các ảnh thu thập cần đảm bảo tính đại diện như con trưởng thành, con non, con trống, con mái, trong mùa sinh sản hay ngoài mùa sinh sản… Số lượng ảnh cho mỗi loài cần thu thập khoảng hơn 150 ảnh. Nhằm tăng khả năng nhận diện đúng cả các loài ngoài 26 loài nguy cấp, quý, hiếm, cũng như khả năng phân biệt loài của mô hình, nhóm nghiên cứu đã thu thập ảnh của một số loài có nhiều đặc điểm tương đồng với 26 loài nguy cấp, quý, hiếm. Việc lựa chọn loài nào cũng như chất lượng ảnh được chọn được thực hiện theo ý kiến chuyên gia, ảnh được thu thập trên internet và các nguồn chính thống (Sách đỏ Việt Nam; Sách các loài chim Việt Nam) được kiểm duyệt, đảm bảo bản quyền bởi các chuyên gia và có chất lượng đảm bảo các chuyên gia có thể nhận diện được đúng loài. Kết quả bộ dữ liệu riêng để luyện mô hình học sâu có 171 loài thuộc 12 họ gồm: Họ Bồ nông, Họ Cổ rắn, Họ Diệc, Họ Hạc, Họ Cò quắm, Họ Vịt, Họ Trĩ, Họ Sếu, Họ Ô tác, Họ Rẽ, Họ Hồng hoàng và Họ Khướu. Bên cạnh dữ liệu thu thập này, khi luyên mô hình chúng tôi áp dụng các kỹ thuật tăng cường (data augmentation) bằng các phép xoay, lật, cắt, dịch chuyển, thay đổi độ bão hòa nhằm mô phỏng sự đa dạng của điều kiện thực địa đồng thời giúp tăng kích thước hiệu dụng của tập huấn luyện và cải thiện khả năng khái quát hóa của mô hình.

Trong nghiên cứu này, nhóm tác giả đã tiến hành xây dựng và thực nghiệm 4 nhóm mô hình học sâu khác nhau cho bài toán nhận diện loài chim nguy cấp, quý, hiếm từ ảnh. Nhóm mô hình thứ nhất sử dụng kiến trúc mạng nơ-ron tích chập cổ điển tựa mạng AlexNet [10], trong đó các tham số được khởi tạo ngẫu nhiên và mô hình được huấn luyện hoàn toàn từ đầu. Cách tiếp cận này cho phép đánh giá khả năng học đặc trưng trực tiếp từ dữ liệu ảnh thu thập và được sử dụng như một một hình cơ sở để so sánh với các mô hình khác. Cụ thể, kiến trúc mạng tựa AlexNet cho việc nhận diện loài chim dựa trên ảnh như sau:

Đầu vào: 3×224×224.

Khối đặc trưng (Feature Extractor):

Conv1: 64@11×11, s=4, p=2 → BN → ReLU → MaxPool(3×3, s=2)

Conv2: 192@5×5, p=2 → BN → ReLU → MaxPool(3×3, s=2)

Conv3: 384@3×3, p=1 → BN → ReLU

Conv4: 256@3×3, p=1 → BN → ReLU

Conv5: 256@3×3, p=1 → BN → ReLU → MaxPool(3×3, s=2)

Bộ phân loại (Classifier): Flatten (256×6×6) → FC 4096 → Dropout 0.5 → FC 4096 → Dropout 0.5 → FC C (số lớp).

Mô hình được luyện sử dụng hàm mất mát CrossEntropy, bộ tối ưu AdamW, tốc độ học khởi đầu là 10-3 và giảm dần theo lược đồ CosineAnnealingLR.

Nhóm mô hình thứ hai dựa trên kiến trúc ResNet [11] và áp dụng kỹ thuật học chuyển giao (transfer learning) từ các trọng số được huấn luyện trước trên tập dữ liệu ImageNet. Các lớp cuối cùng được điều chỉnh để phù hợp với số lượng loài trong tập dữ liệu nghiên cứu. Cách tiếp cận này giúp mô hình tận dụng được tri thức đã học về đặc trưng hình thái tự nhiên của các loài chim và môi trường, nhờ đó tăng tốc hội tụ và giảm hiện tượng quá khớp trong điều kiện dữ liệu hạn chế. Kiến trúc mạng dựa trên ResNet cho việc nhận diện loài chim dựa trên ảnh như sau:

Đầu vào: 3×224×224.

Khối nền: ResNet50.

Đầu ra sau khối Global Average Pooling (vector có kích thước 2048 chiều).

Bộ phân loại (Classifier): FC 2048 → Dropout 0.2 → FC C (số lớp).

Mô hình được luyện theo hai giai đoạn gồm giai đoạn trích đặc trưng (đóng băng các trọng số của ResNet) và giai đoạn tinh chỉnh (mở khóa dần một số tầng sâu nhất của RestNet).

Nhóm mô hình thứ ba sử dụng kiến trúc Transformer trong lĩnh vực thị giác (Swin Transformer [12]), kết hợp với kỹ thuật fine-tuning nhằm tinh chỉnh các tham số của toàn bộ mô hình. Mục tiêu là tận dụng khả năng biểu diễn toàn cục của Transformer, giúp mô hình nhận diện tốt hơn các mối quan hệ không gian giữa các vùng hình ảnh. Kiến trúc mạng Swin Transformer cho việc nhận diện loài chim dựa trên ảnh như sau:

Đầu vào: 3×224×224.

Khối transformer: SwinTransformer.

Bộ phân loại (Classifier): FC 768 → FC C (số lớp).

Mô hình được khởi tạo từ pretrained weights trên ImageNet và sau đó được luyện trên bộ dữ liệu riêng với tốc độ học 5*10-4, hàm mất mát CrossEntropy và bộ tối ưu AdamW.

Nhóm mô hình thứ tư khai thác đặc trưng nhúng (embedding) từ mô hình CLIP [13] – một mô hình nền tảng học song song giữa ảnh và văn bản. Các vector đặc trưng ảnh được trích xuất và đưa qua một mạng MLP nông để thực hiện phân lớp. Cách tiếp cận này hướng tới khả năng mở rộng cho các bài toán nhận diện mở rộng (open-set

recognition) và tích hợp với nhận diện từ mô tả văn bản. Kiến trúc của mô hình như sau:

Đầu vào: 3×224×224.

Khối embedding: CLIP ViT-B/16.

Embedding với kích thước D=768.

Bộ phân loại (Classifier): FC 768 → GELU → BatchNorm → Dropout 0.2 → FC C (số lớp).

Mô hình được luyện trên bộ dữ liệu riêng với tốc độ học 10-4, hàm mất mát CrossEntropy và bộ tối ưu AdamW.

Kết quả so sánh hiệu năng giữa các nhóm mô hình cho thấy mô hình ResNet kết hợp học chuyển giao đạt độ chính xác cao nhất và ổn định nhất trên tập kiểm định. Mô hình này thể hiện khả năng cân bằng giữa độ chính xác, tốc độ huấn luyện và tính khả chuyển, đồng thời duy trì hiệu quả khi mở rộng sang các họ hoặc loài chim chưa có trong tập huấn luyện. Do đó, ResNet với học chuyển giao được lựa chọn làm mô hình chính cho giai đoạn nhận diện loài từ ảnh trong hệ thống đề xuất.

2.3. Hậu xử lý

Phân loại mở (open-set classification) là hướng tiếp cận trong học sâu cho phép mô hình không chỉ phân loại các mẫu thuộc những lớp đã được huấn luyện, mà còn có khả năng nhận biết và loại bỏ các mẫu đến từ những lớp chưa từng thấy trong tập huấn luyện. Khác với phân loại đóng, giả định rằng mọi mẫu đầu vào đều thuộc một trong các lớp đã biết, phân loại mở phản ánh thực tế phức tạp hơn của môi trường tự nhiên, nơi mô hình có thể gặp các loài chưa được gán nhãn.

Trong bài toán nhận diện 26 loài chim nguy cấp, quý, hiếm, việc áp dụng phân loại mở là cần thiết vì hệ thống có thể nhận các ảnh chứa các loài chim khác không nằm trong danh sách cần bảo vệ. Nếu không có cơ chế phát hiện "ngoài tập huấn luyện", mô hình có thể gán nhầm các loài thường gặp thành loài nguy cấp, gây sai lệch nghiêm trọng cho công tác giám sát và bảo tồn. Do đó, việc tích hợp nguyên lý phân loại mở giúp hệ thống nhận diện an toàn hơn, chỉ đưa ra kết luận khi có đủ độ tin cậy, đồng thời hỗ trợ phát hiện các trường hợp tiềm năng của loài mới hoặc chưa được định danh.

Trong những năm gần đây, bài toán phân loại mở (open-set classification) đã thu hút nhiều sự quan tâm trong cộng đồng học sâu, đặc biệt trong các ứng dụng có tính đa dạng và không kiểm soát hoàn toàn như nhận diện đối tượng tự nhiên, an ninh sinh học hay giám sát ĐDSH. Các hướng tiếp cận chính có thể chia thành 3 nhóm: (i) Tiếp cận dựa trên ngưỡng xác suất (thresholding-based), trong đó mô hình sử dụng giá trị softmax hoặc entropy để quyết định xem mẫu có thuộc tập huấn luyện hay không; (ii) tiếp cận dựa trên không gian đặc trưng (feature space-based), khai thác phân bố của các đặc trưng ẩn để phát hiện các điểm bất thường; và (iii) tiếp cận dựa trên mô hình hóa xác suất (probabilistic modeling), trong đó các lớp được biểu diễn bởi các phân bố thống kê cho phép đánh giá độ xa lạ của mẫu mới.

Trong số các phương pháp tiêu biểu, OpenMax [14] được xem là một trong những giải pháp nền tảng cho bài toán phân loại mở trong học sâu. OpenMax mở rộng hàm Softmax truyền thống bằng cách mô hình hóa phân bố các đặc trưng lớp trên không gian kích hoạt của lớp cuối cùng thông qua phân bố Weibull. Với mỗi lớp đã biết, mô hình học các thông số Weibull mô tả mức độ "xa lạ" của một mẫu so với trung tâm lớp. Khi dự đoán, OpenMax hiệu chỉnh lại xác suất của các lớp bằng cách giảm trọng số cho các mẫu có độ lệch cao, đồng thời bổ sung một lớp "unknown" biểu diễn các mẫu có khả năng nằm ngoài tập huấn luyện.

Cách tiếp cận này giúp mô hình không chỉ nhận diện chính xác các lớp đã biết mà còn phát hiện được các mẫu bất thường hoặc thuộc loài chưa từng gặp, phù hợp với yêu cầu của hệ thống nhận diện 26 loài chim nguy cấp, quý, hiếm, nơi khả năng phân biệt giữa "loài đã biết" và "loài ngoài tập" là yếu tố then chốt để đảm bảo độ tin cậy và an toàn trong ứng dụng thực tế.

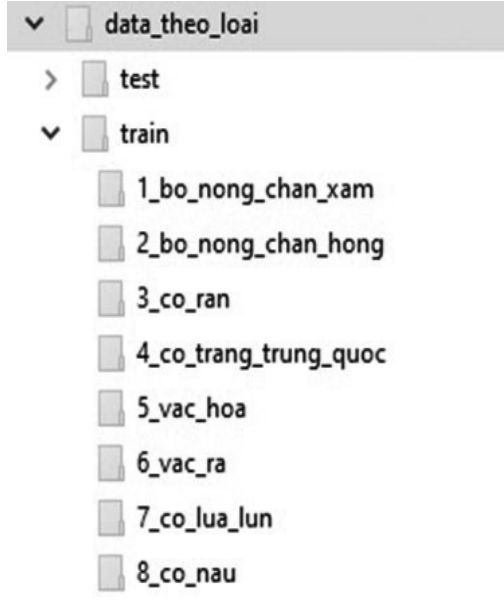
3. KẾT QUẢ VÀ THẢO LUẬN

3.1. Dữ liệu

Nhóm nghiên cứu đã thu thập được một bộ dữ liệu của 171 loài thuộc 12 họ với tổng số 27.457 ảnh chia tương đối đều cho các loài. Các dữ liệu ảnh này được xáo trộn và chia ngẫu nhiên theo tỷ lệ 80%:10%:10% vào 3 thư mục:

- Train: Thư mục dữ liệu để luyện mô hình.
- Val: Thư mục dữ liệu để kiểm định mô hình, hiệu chỉnh các siêu tham số.
- Test: Thư mục dữ liệu để đánh giá mô hình.

Trong cả 3 thư mục là các thư mục con được đặt theo tên loài, các ảnh của loài được đặt trong các thư mục con tương ứng này.



Hình 3. Cấu trúc thư mục dữ liệu

Bảng 1. Thông tin luyện giai đoạn 1 (đóng băng tham số của ResNet)

Epoch	train_loss	train@1 (%)	val_loss	val@1 (%)	val@5 (%)
1	2,9541	43,47	1,7588	72,29	93,29
2	1,8572	68,98	1,6504	75,76	93,81
3	1,5558	78,59	1,5426	78,48	94,78
4	1,4107	82,71	1,4842	80,53	95,34
5	1,2583	87,82	1,3722	84,3	96,23
6	1,1413	91,7	1,3347	85,12	96,94
7	1,0557	94,34	1,2913	86,95	96,64
8	1,0131	95,77	1,2674	86,87	96,9

Bảng 2. Thông tin luyện giai đoạn 2 (bỏ đóng băng dần các tham số của ResNet)

Epoch	train_loss	train@1 (%)	val_loss	val@1 (%)	val@5 (%)
1	1,025	95,42	1,2869	86,5	96,68
2	1,0597	93,97	1,3711	83,89	96,34
3	1,0499	94,4	1,3681	83,7	96,01
4	1,0333	94,74	1,3371	85,08	97,09
5	1,0105	95,52	1,3263	85,57	96,23
6	1,0014	95,74	1,3315	85,86	96,12
7	0,985	96,3	1,3466	84,93	96,05
8	0,9689	96,8	1,341	83,77	96,38
9	0,9559	97,12	1,3275	85,64	96,64
10	0,9495	97,23	1,3104	85,68	96,64
11	0,9345	97,61	1,296	85,86	96,64

Thêm nữa, nhằm mục đích đánh giá hiệu suất của mô hình đối với các lớp ngoài tập luyện, nhóm nghiên cứu tạo thêm một thư mục 0_unknown chứa 40 file ảnh của các loài như chó (4 ảnh), mèo (4 ảnh), xe đạp (4 ảnh), ô tô (4 ảnh), hoa ly (4 ảnh), hoa hồng (4 ảnh), gà (4 ảnh), chim bồ câu (4 ảnh), chim sẻ (4 ảnh), và vẹt (4 ảnh) được thu thập trên Internet. Bộ dữ liệu 0_unknown này hoàn toàn không được biết đến trong quá trình luyện mô hình mà chỉ được sử dụng để đánh giá khả năng nhận biết các lớp ngoài dữ liệu luyện. Nếu phân theo ảnh thuộc lớp chim hay không thì 40 ảnh này gồm 24 ảnh không có lớp chim và 16 ảnh có lớp chim.

3.2. Thiết kế, luyện và đánh giá kết quả mô hình học sâu

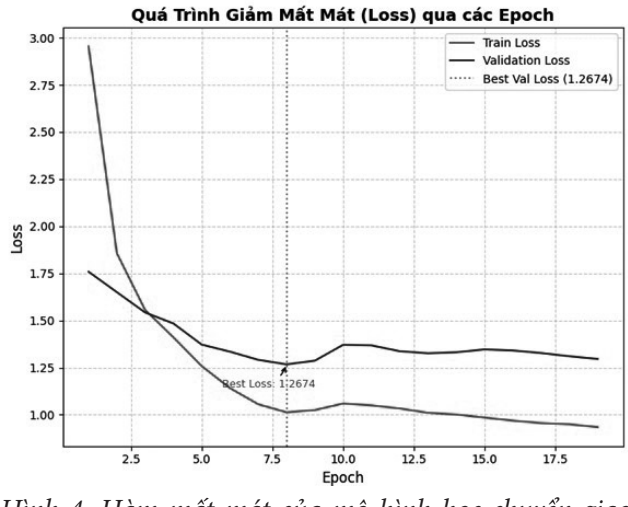
Mô hình học sâu dựa trên mô hình ResNet50 và được học chuyển giao các giá trị tham số từ mô hình đã được luyện trước. Gắn vào ResNet50 là một lớp Dropout và một lớp Linear với số nơ-ron đầu ra là 171 để phân lớp 171 loài chim. Mô hình được luyện theo chiến lược 8 epoch đầu đóng băng các tham số của ResNet và 22 epoch tiếp theo luyện cả các tham số của ResNet theo thứ tự từ các lớp cuối về các lớp đầu. Việc luyện mô hình đã kết thúc sớm ở epoch thứ 19 khi hơn 10 bước cập nhật giá trị tham số mà hiệu năng của mô hình không thay đổi.

Bảng 1 trình bày kết quả luyện mô hình trong giai đoạn 1 (giai đoạn đóng băng tham số của ResNet).

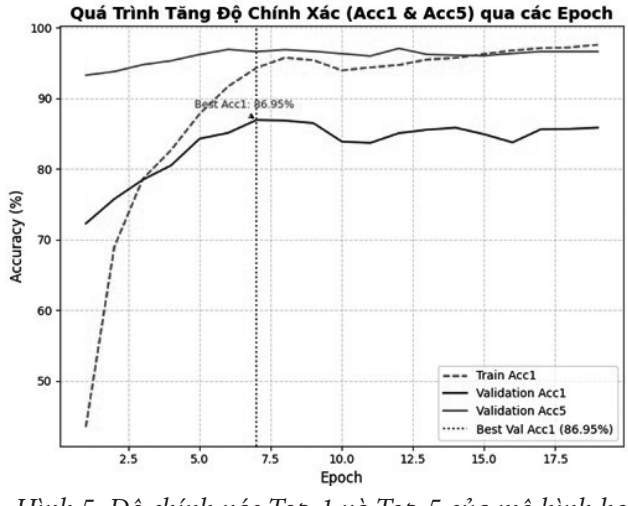
Bảng 2 trình bày kết quả luyện mô hình trong giai đoạn 2 khi dần bỏ đóng băng các tham số của mô hình ResNet. Quá trình luyện đã kết thúc sớm ở epoch thứ 11 của giai đoạn này.

Các biểu đồ của hàm mất mát và độ chính xác theo số epoch luyện mô hình được thể hiện trong các Hình 4 và Hình 5.

Bộ dữ liệu kiểm định (test) sau khi được bổ sung thêm 40 ảnh cho nhãn



Hình 4. Hàm mất mát của mô hình học chuyển giao dựa trên ResNet theo số epoch



Hình 5. Độ chính xác Top-1 và Top-5 của mô hình học chuyển giao dựa trên ResNet theo số epoch

"không biết" có tổng số 2.912 ảnh. Sử dụng bộ dữ liệu này để kiểm định mô hình cho kết quả trung bình cho 172 lớp (gồm cả lớp unknown không được biết khi luyện mô hình) như sau:

- Top-1 Accuracy: 0,8465.
- Top-3 Accuracy: 0,9365.
- Top-5 Accuracy: 0,9564.
- Top-1 Macro Precision: 0,8509.
- Top-1 Macro Recall: 0,8360.
- Top-1 Macro F1-score: 0,8352.

Đây là một kết quả tương đối tốt giúp mô hình có thể được áp dụng hiệu quả trong thực tế nhận diện loài chim nguy cấp, quý, hiếm được ưu tiên bảo vệ. Tuy nhiên, do đặc điểm của dạng mô hình đóng là cố gắng dự đoán các ảnh đầu vào thuộc các loài trong bộ dữ liệu luyện (cụ thể là 171 loài) do đó toàn bộ 40 ảnh thuộc nhóm 0_unknown đều được nhận diện là ảnh của một trong 171 loài được luyện.

3.2. Xây dựng thủ tục tiền xử lý và đánh giá kết quả

Áp dụng thủ tục tiền xử lý để xác định ảnh đầu vào có chứa chim hay không giúp kết luận sớm quá trình nhận diện. Ảnh không chứa chim được dự báo với nhãn "không biết". Thực hiện kiểm định với cùng bộ dữ liệu test cho kết quả sau:

- Top-1 Accuracy: 0,8386.
- Top-3 Accuracy: 0,9255.
- Top-5 Accuracy: 0,9433.
- Top-1 Macro Precision: 0,8623.
- Top-1 Macro Recall: 0,8265.
- Top-1 Macro F1-score: 0,8354.

Có thể thấy, khi áp dụng thủ tục tiền xử lý, độ chính xác giảm đi một chút nhưng bù lại mô hình đã giảm dương tính giả từ đó tạo điều kiện để ứng dụng chatbot cuối hỏi người dùng thêm thông tin trước khi đưa ra quyết định cuối. Chỉ số tổng hợp là Top-1 Macro F1-score tăng một chút chứng tỏ về tổng thể áp dụng thủ tục tiền xử lý có hiệu quả. Chi tiết với nhóm 0_know, việc tiền xử lý đã loại 25/40 ảnh khỏi quá trình nhận diện bằng mô hình học sâu và nhanh chóng kết luận 25 ảnh này không thuộc một trong 171 loài.

3.4. Xây dựng thủ tục hậu xử lý và đánh giá kết quả

Để mô hình có thể dự đoán cả lớp "không biết" cũng xác suất của nó, thay vì dùng Softmax trên đầu ra của hàm Logit (véc tơ 171 chiều), mô hình cần "khấu trừ" xác suất từ các 171 lớp cho xác suất của lớp "không biết" này. Việc tính toán các phần cần khấu trừ này sử dụng các phân phối weibull cho từng lớp và do đó cần trích các đặc trưng từ mô hình. Các đặc trưng được thiết kế để lấy ở lớp sau cùng ngay trước lớp phân loại (thường là đầu ra của khối global average pooling (avgpool)). Các vec tơ đặc trưng này có kích thước 2048 chiều cho mỗi mẫu ảnh đầu vào. Lớp này được xem là biểu diễn trừu tượng nhất của dữ liệu, phản ánh thông tin hình thái và ngữ nghĩa tổng quát mà mô hình đã học được.

Thực hiện đưa toàn bộ tập huấn luyện qua mô hình ở chế độ suy luận (inference) để trích xuất các vector đặc trưng. Tuy nhiên, chỉ những mẫu được mô hình dự đoán đúng nhãn mới được giữ lại để đảm bảo rằng các đặc trưng thu được thực sự đại diện cho không gian đặc trưng "chuẩn" của từng lớp. Đối với mỗi lớp c, thu được tập các vector đặc trưng $\{f_i^c\}_{i=1}^{N_c}$ với $f_i^c \in \mathbb{R}^{2048}$

trong đó N_c là số mẫu được dự đoán đúng của lớp đó.

Từ tập đặc trưng này, tính Mean Activation Vector (MAV) cho lớp c bằng công thức:

$$\text{MAV}_c = \frac{1}{N_c}\sum_{i=1}^{N_c} f_i^c$$

MAV đóng vai trò là "trung tâm" của phân bố đặc trưng của lớp trong không gian biểu diễn 2048 chiều.

Thực hiện mô hình hóa mức độ biến thiên của các đặc trưng quanh MAV bằng cách tính khoảng cách Euclid-cosine giữa từng vector đặc trưng và MAV tương ứng theo công thức:

$$d_i^c = \|f_i^c - \text{MAV}_c\|_2 + \frac{1}{2}(1 - \cos(f_i^c, \text{MAV}_c))$$

Trong đó, thành phần chuẩn Euclid phản ánh độ lệch về biên độ, còn thành phần cosine đo mức độ khác biệt về hướng của vector trong không gian đặc trưng. Hàm khoảng cách kết hợp này giúp cân bằng giữa độ lớn và hướng của đặc trưng, cho phép mô hình nhận biết tốt hơn các điểm nằm xa phân bố huấn luyện.

Cuối cùng, các giá trị khoảng cách $\{d_i^c\}$ được sắp xếp theo thứ tự tăng dần và phân phối Weibull được fit lên phần đuôi (tail) của phân bố này, nghĩa là các giá trị lớn nhất thể hiện các điểm xa MAV nhất trong lớp. Phân phối Weibull được sử dụng để mô tả xác suất một mẫu mới nằm "xa bất thường" so với MAV của lớp, làm cơ sở cho các mô hình nhận diện mẫu chưa biết (unknown samples) trong không gian mở.

Quy trình này được lặp lại cho tất cả các lớp, tạo thành tập các tham số $\{\text{MAV}_c, \text{Weibull}_c\}_{c=1}^{C}$, cung cấp nền tảng cho phương pháp như OpenMax, trong đó mỗi lớp có một mô hình thống kê riêng phản ánh vùng đặc trưng "an toàn" của lớp trong không gian đặc trưng học sâu.

Bằng cách phân phối lại xác suất từ các lớp thông thường sang cho lớp "không biết", thủ tục hậu xử lý cho phép nhận diện cả các lớp "không biết" chưa hề được biết đến trong tập luyện mô hình. Thực hiện kiểm định với cùng bộ dữ liệu test (kèm cả bước tiền xử lý) cho kết quả sau:

- Top-1 Accuracy: 0,8386.
- Top-3 Accuracy: 0,9255.
- Top-5 Accuracy: 0,9433.
- Top-1 Macro Precision: 0,8623.
- Top-1 Macro Recall: 0,8265.
- Top-1 Macro F1-score: 0,8354.

Kết quả kiểm định trên cho thấy áp dụng OpenMax có cải tiến giảm dương tính giả và chỉ số tổng thể F1-score có cải tiến. Với nhóm 0_unknown, OpenMax giúp nhận diện được 22/40 ảnh không thuộc một trong 171 loài.

4. KẾT LUẬN

Nghiên cứu đã phát triển thành công một giải pháp nhận diện đáng tin cậy cho 26 loài chim nguy cấp, quý, hiếm dựa trên học sâu tích hợp. Kiến trúc đề xuất bao gồm: (i) Giai đoạn tiền xử lý bằng YOLOv12 để loại bỏ các ảnh không chứa đối tượng chim; (ii) Mô hình ResNet tinh chỉnh được huấn luyện trên bộ dữ liệu chuyên biệt (27.457 ảnh của 171 loài, do nhóm nghiên cứu tự thu thập); (iii) Giai đoạn hậu xử lý sử dụng nguyên lý phân loại tập mở nhằm giảm sai lệch nhận dạng đối với các loài không có trong tập huấn luyện. Những đóng góp chính của nghiên cứu là bộ dữ liệu chuyên biệt cùng quy trình tích hợp ba bước, cho phép mô hình đạt được độ chính xác cao (Macro F1-Score là 83,54%) và khả năng vận hành ổn định trong môi trường thực tế.

Tuy nhiên, một số hạn chế vẫn tồn tại như: Bộ dữ liệu hiện chưa bao quát đầy đủ biến thiên tự nhiên và khả năng nhận diện tập mở mới chỉ được đánh giá ở mức độ ban đầu. Các hướng nghiên cứu tiếp theo sẽ tập trung vào mở rộng dữ liệu từ các nguồn bẫy ảnh (camera traps) và khoa học công dân, cải thiện cơ chế phát hiện ngoài tập huấn luyện (Out-of-Distribution/OOD), phát triển hệ thống hoàn chỉnh để triển khai tại các khu bảo tồn và vườn quốc gia. Nghiên cứu cũng khuyến nghị cơ quan quản lý cần tăng cường đầu tư xây dựng cơ sở dữ liệu ĐDSH chuẩn hóa quốc gia và tích hợp các hệ thống nhận diện tự động vào hoạt động giám sát thường xuyên. Điều này nhằm nâng cao hiệu quả bảo tồn và quản lý bền vững tài nguyên sinh học.

Lời cảm ơn: *Bài báo sử dụng kết quả nghiên cứu của nhóm tác giả khi thực hiện đề tài "Nghiên cứu ứng dụng công nghệ trí tuệ nhân tạo (AI) xây dựng hệ thống hỗ trợ nhận diện loài nguy cấp, quý, hiếm được ưu tiên bảo vệ - thử nghiệm đối với lớp chim", (mã số TNMT.2024.04.02)*■

TÀI LIỆU THAM KHẢO

1. D. Tuia, B. Kellenberger, S. Beery, B. R. Costelloe, S. Zuffi, B. Risse, A. Mathis, M. W. Mathis, F. Van Langevelde and T. Burghardt, "Perspectives in machine learning for wildlife conservation," Nature communications, vol. 13, no. 1, p. 792, 2022.

2. C. Wah, S. Branson, P. Welinder, P. Perona and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," California institute of technology, 2011.

3.T. Berg, J. Liu, S. Woo Lee, M. L. Alexander, D. W. Jacobs and P. N. Belhumeur, "Birdsnap: Large-scale fine-grained visual categorization of birds," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014.

4. G. Van Horn, S. Branson, R. Farrell, S. Haber, J. Barry, P. Ipeirotis, P. Perona and S. Belongie, "Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.

5.W. Rabhi, F. Eljaimi, W. Amara, Z. Charouh, A. Ezzouhri, H. Benaboud, M. B. Saindou and F. Ouardi, "An integrated framework for bird recognition using dynamic machine learning-based classification," in 2023 IEEE Symposium on Computers and Communications (ISCC), 2023.

6. Sở NN&PTNN tỉnh Thanh Hóa, "Xây dựng phần mềm nhận dạng nhanh một số loài động, thực vật nguy cấp, quý, hiếm phục vụ công tác quản lý, bảo vệ rừng và bảo tồn ĐDSH trên địa bàn tỉnh Thanh Hóa," 2020.

7. Cục Bảo vệ thực vật, "Xây dựng phần mềm nhận diện sinh vật gây hại trên lúa," 2021-2022.

8. Đại học Lâm nghiệp Hà Nội, "Ứng dụng trí tuệ nhân tạo trong nhận biết nhanh gỗ trên điện thoại thông minh," 2020.

9.Y. Tian, Q. Ye and D. Doermann, "Yolov12: Attention-centric real-time object detectors," arXiv preprint arXiv:2502.12524, 2025.

10 A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol. 25, 2012.

11. K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

12. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF international conference on computer vision, 2021.

13. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin and J. Clark, "Learning transferable visual models from natural language supervision," in International conference on machine learning, 2021.

14. A. Bendale and T. E. Boult, "Towards open set deep networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.