MÔ HÌNH MẠNG NƠRON TÍCH CHẬP THỂ NHẸ DỰA TRÊN KIẾN TRÚC DENSENET CHO NHẬN DẠNG BIỂU CẢM KHUÔN MẶT VÀ ỨNG DỤNG HỖ TRỢ ĐÁNH GIÁ QUÁ TRÌNH HỌC TẬP TRỰC TUYẾN

LIGHTWEIGHT DENSE-BASED CNN MODEL FOR FACIAL EXPRESSION RECOGNITION AND APPLICATION FOR ONLINE LEARNING EVALUATION

Dương Thăng Long*, Đỗ Thị Thu Hà†, Trần Văn Nam‡

Ngày tòa soạn nhận được bài báo: 04/10/2022
Ngày nhận kết quả phản biện đánh giá: 04/04/2023
Ngày bài báo được duyệt đăng: 28/04/2023

Tóm tắt: *Mạng nơ-ron tích chập (CNN) được áp dụng cho nhận dạng cảm xúc trên khuôn mặt đang được quan tâm nghiên cứu của nhiều tác giả với những kết quả rất khả quan và có các ứng dụng thành công. Các mô hình CNN hiện đại được thiết kế với các kiến trúc đa dạng như VGG, ResNet, Xception, EfficientNet, DenseNet và các biến thể của chúng được áp dụng rộng rãi cho các bài toán nhận dạng hình ảnh, trong đó có nhận dạng biểu cảm khuôn mặt. Tuy nhiên, các mô hình này có độ phức tạp khá lớn đối với một số ứng dụng trong thực tế hạn chế về tài nguyên tính toán. Bài báo này đề xuất một mô hình CNN thể nhẹ dựa trên kiến trúc kết nối dày đặc của mô hình DenseNet với độ phức tạp vừa phải nhưng vẫn đảm bảo chất lượng và hiệu quả cho nhận dạng cảm xúc trên khuôn mặt. Chúng tôi cũng thiết kế tích hợp mô hình này với hệ thống LMS nhằm hỗ trợ ghi nhận và đánh giá quá trình học tập trực tuyến của người học. Mô hình đề xuất được thử nghiệm để đánh giá trên một số bộ dữ liệu phổ biến, kết quả cho thấy mô hình đem lại hiệu quả và có thể được sử dụng trong thực tế.*

Từ khoá: *Mạng nơron tích chập, kiến trúc mạng DenseNet, nhận dạng biểu cảm khuôn mặt, hệ thống quản lý học tập trực tuyến.*

Abstract: *Convolutional neural networks (CNN) for facial emotion recognition (FER) are being studied by many authors with very positive results and successful applications. State-of-the-art CNN models with diverse architectures such as VGG, ResNet, Xception, EfficientNet, and DenseNet and their variations are widely applied to many image recognition problems,*

---

* Trường Đại học Mở Hà Nội

† Trường THPT Trần Nhân Tông, Hà Nội

‡ IT-VNEH, Vietnam National Eye Hospital

including FER. However, these models have considerable complexity for some real-world applications with limited computational resources. This paper proposes a lightweight CNN model based on DenseNet architectures with moderate complexity but still ensures quality and efficiency for facial emotion recognition. Then, it is designed to be integrated into LMS for recording and evaluating online learning activities. The proposed model is tested to assess some popular datasets; the results show that the model is effective and can be used in practice.

Keywords: *Convolutional neural network, DenseNet architecture, facial expressions recognition, online learning management systems.*

I. Đặt vấn đề

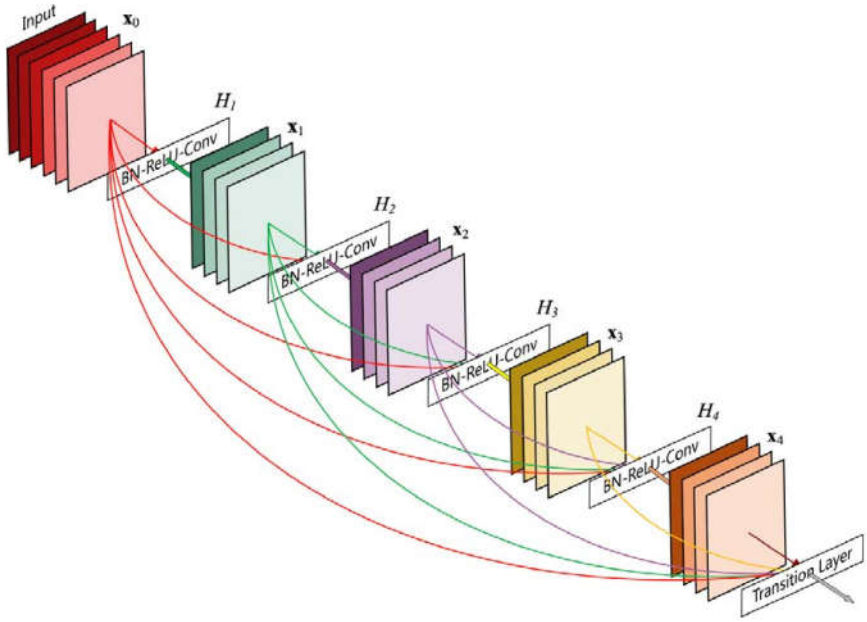
Nhận dạng biểu cảm khuôn mặt (FER) đang rất được quan tâm nghiên cứu rộng rãi hiện nay và nó có tính ứng dụng cao trong lĩnh vực thị giác máy tính. Biểu cảm trên khuôn mặt của con người đóng một vai trò quan trọng trong bất kỳ giao tiếp giữa các cá nhân với nhau, nó có thể giúp người khác hiểu được cảm xúc hoặc thậm chí ý định của một người, khiến nó trở thành một yếu tố giao tiếp không thể thiếu trong tương tác giữa con người với nhau. Với sự phát triển của công nghệ thị giác máy tính và ứng dụng thực tế của nó, các kết quả của các nghiên cứu khác nhau về nhận dạng biểu cảm trên khuôn mặt (Facial Expression Recognition - FER) cho thấy các phương pháp này không chỉ cho kết quả chính xác cao mà còn có thể tiết kiệm chi phí nhân lực một cách hiệu quả trong ứng dụng thực tế, như trong các lĩnh vực giao diện người-máy, hoạt hình, xe tự hành, giao thông, y học và giáo dục [1].

Trong [2] đề cập việc Ekman và Friesen xác định sáu biểu cảm cơ bản chung của con người mà họ tin rằng đều xuất hiện ở tất cả mọi người bất kể quốc gia, dân tộc hay tôn giáo nào. Sáu biểu cảm gồm hạnh phúc (Ha), buồn bã (Sa), ngạc nhiên (Su), ghê tởm (Di), tức giận (An) và sợ hãi (Fe). Các bộ phận thay đổi, chuyển động trên khuôn mặt như lông mày nhướng lên, lông mày khóa và khóe miệng di chuyển ra ngoài hay mở ra, đóng vào được coi là những đơn vị thay đổi cơ bản của nét mặt. Tuy nhiên, các biểu cảm trên khuôn mặt của mọi người có sự thay đổi đột xuất và các biểu cảm được thể hiện bởi các khuôn mặt khác nhau cũng khác nhau ở mỗi người. Những yếu tố này tác động lớn đến hoạt động và hiệu quả của bất kỳ hệ thống FER nào bao gồm các kỹ thuật thị giác máy tính. Hệ thống cơ sở dữ liệu về các biểu cảm trên khuôn mặt được các tác giả thiết lập, mô tả chi tiết của từng biểu cảm, đặt nền tảng cho việc giải quyết bài toán FER. Hiện nay, các bộ cơ sở dữ liệu cho nghiên cứu về FER được công bố khá nhiều như CK+, JAFFE, Oulu-CASIA (đề cập trong [1], [2]) và chúng được nhiều tác giả sử dụng để đánh giá cho kết quả các mô hình FER.

Phương pháp cho bài toán FER sử dụng các kỹ thuật xử lý ảnh và học máy truyền thống như kỹ thuật biến đổi đối tượng bất biến theo tỷ lệ (SIFT), biểu đồ histogram (HOG) hay phân tích local binary patterns (LBP) [2] đều dựa trên hai loại đặc trưng cục bộ và đặc trưng toàn cục (theo hình học). Các đặc trưng cho biểu cảm khuôn mặt sau khi được trích xuất sẽ sử dụng làm đầu vào cho bộ phân lớp như BP, SVM [2] để phân loại và thu được kết

quả nhận dạng cuối cùng. Tuy nhiên, các phương pháp truyền thống này đạt hiệu quả thấp do sự khác biệt, thay đổi lớn của hình ảnh đối với các góc chụp khác nhau và do tính đơn giản của kiến trúc mô hình nhận dạng. Gần đây, việc sử dụng công nghệ học sâu với mạng nơ-ron tích chập (CNN) được phát triển mạnh mẽ và mang lại hiệu quả cao [2]- [3], nhiều đặc trưng ẩn sâu trong hình ảnh có thể được trích xuất dựa trên huấn luyện mô hình CNN để tạo ra một hệ thống FER mạnh mẽ. Do đó, chúng rất ổn định đối với các hình ảnh với vị trí khuôn mặt có góc chụp và thay đổi tỷ lệ khác nhau [4]. Có khá nhiều mô hình CNN khác nhau cho bài toán FER đã được đề xuất trong các nghiên cứu dựa trên các kiến trúc hiện đại như VGG, SENet, Xception [5], GoogleNet, ResNet [2] hay EfficientNet [6].

Trong đó, kiến trúc DenseNet [7] cung cấp mô hình nhẹ hơn với bộ tham số huấn luyện khá nhỏ và cho kết quả tốt hơn, nó cũng được nhiều tác giả sử dụng như một mạng xương sống cho các mô hình CNN [8], [9], [5]. Kiến trúc này có lưới kết nối dày đặc được thiết kế để tăng khả năng chuyển tải thông tin từ lớp nơron phía trước đến các lớp nơron phía sau và sử dụng tốc độ tăng trưởng thông tin để thiết lập mức độ đóng góp của mỗi lớp vào trạng thái toàn cục của mô hình. Cụ thể, trong mỗi khối của kiến trúc này (gọi là khối kết nối dày đặc, dense block) gồm một số lớp nơron có kết nối dày đặc, tức là lớp nơron phía trước kết nối trực tiếp đến tất cả các lớp nơron phía sau trong khối, cuối mỗi khối có lớp nơron đóng vai trò chuyển tiếp thông tin (transition layers) đến khối tiếp theo (Hình 1.1).



Hình 1.1. Minh họa kiến trúc DenseNet [7]

Mặc dù hầu hết các kiến trúc mô hình CNN được đề xuất đạt hiệu quả cao nhưng chúng có cấu trúc rất phức tạp, chi phí tính toán cao và đòi hỏi hệ thống tính toán lớn trong cả huấn luyện mô hình và sử dụng mô hình của ứng dụng. Do đó, một số nghiên cứu đề xuất sử dụng mô hình CNN ở thể nhẹ [10], [9], [11], [12], [13], [4] nhằm phù hợp với các trường hợp ứng dụng hạn chế về hệ thống tính toán nhưng

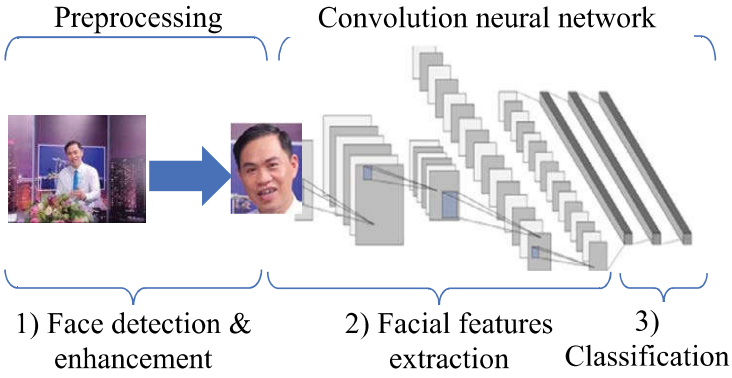
vẫn đảm bảo được hiệu quả của mô hình, đặc biệt mô hình có thể thực thi trên môi trường trực tuyến dạng web.

Trong nghiên cứu này, chúng tôi đề xuất một mô hình CNN ở thể nhẹ dựa trên kiến trúc các khối kết nối dày đặc cho bài toán FER, mô hình có kích thước nhỏ với số lượng tham số của mô hình ít hơn, cho tốc độ thực hiện nhanh hơn và phù hợp trong nhiều ứng dụng thực tế có hệ thống tính toán hạn chế. Hơn nữa, để tăng chất lượng của mô hình trong học máy, chúng tôi áp dụng các phép biến đổi và xử lý hình ảnh nhằm tăng cường thêm dữ liệu (data augmentation) huấn luyện mô hình. Các phần tiếp theo của bài báo này gồm Phần 2 giới thiệu chi tiết về mô hình CNN đề xuất và thiết kế một hệ thống ứng dụng tích hợp của mô hình với hệ thống quản lý học tập (LMS) để ghi nhận và hỗ trợ đánh giá kết quả học tập trực tuyến. Trong Phần 3, các thử nghiệm của mô hình được triển khai trên các bộ dữ liệu khác nhau, các kết quả đạt được và phân tích, so sánh với các mô hình khác để đánh giá hiệu quả. Cuối cùng, Phần 4 là nội dung kết luận.

II. Phương pháp

2.1. Mô hình LDFER

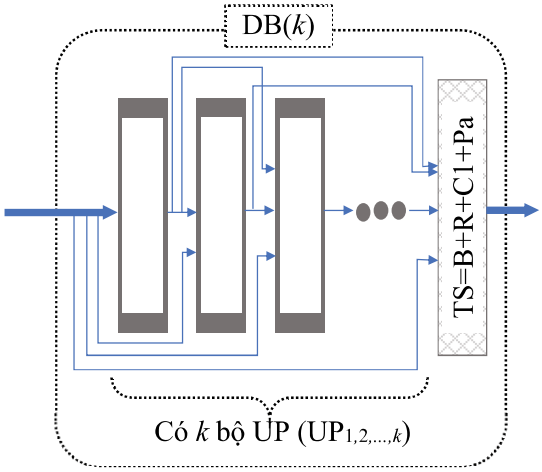
Trong phần này, chúng tôi trình bày chi tiết cho mô hình CNN dựa trên các khối kết nối dày đặc theo kiến trúc DenseNet và tích hợp nó với một hệ thống LMS để ghi nhận và hỗ trợ đánh giá quá trình học tập trực tuyến. Mô hình của chúng tôi, ký hiệu là LDFER (Light-DenseNet Architecture-based for Facial Expressions Recognition), về tổng thể hoạt động được chia thành ba giai đoạn chính (Hình 2.1) gồm: (1) chụp ảnh người học từ thiết bị đầu cuối có kết nối, tiền xử lý hình ảnh để phát hiện vùng khuôn mặt và nâng cao chất lượng của hình ảnh nếu cần; (2) thực hiện trích xuất các đặc trưng của biểu cảm trên khuôn mặt của người học; và (3) phân lớp các đặc trưng để nhận dạng các trạng thái biểu cảm trên khuôn mặt nhằm ghi nhận và hỗ trợ đánh giá kết quả học tập.



Hình 2.1. Sơ đồ tổng thể mô hình LDFER

Phần lõi của mô hình LDFER là kiến trúc CNN thực hiện hai chức năng chính: trích xuất các đặc trưng biểu cảm trên khuôn mặt bằng các khối nơron dưới dạng kết nối dày đặc (gọi là khối dense-connectivity, DB) và phân lớp các đặc trưng được trích xuất thành một trong những biểu cảm cần nhận dạng trên khuôn mặt. Cấu trúc mỗi khối DB trong mô hình này được thiết kế gồm nhiều lớp nơron mà mỗi lớp nơron ở phía trước có liên kết dạng tăng trưởng kênh thông tin (channel-wise concatenate) đến

tất các lớp nơron phía sau. Nói cách khác, đầu ra của lớp nơron phía trước đóng góp thêm cho kênh đầu vào của các lớp nơron ở sau trong khối. Để tránh hiện tượng bùng nổ gradient khi tính toán của quá trình huấn luyện mô hình, mỗi lớp nơron trong khối DB đều được áp dụng cơ chế chuẩn hoá thông tin theo gói (batch normalization) và do đó nó giúp ổn định phân phối của dữ liệu huấn luyện về phân phối chuẩn qua tất cả các lớp nơron. Như vậy, mỗi bộ xử lý tín hiệu đặc trưng (unit processing, ký hiệu UP) trong khối DB là một bộ các phép xử lý gồm chuẩn hoá dữ liệu theo gói (B), tính toán kích hoạt bằng phép tuyến tính "relu" (R) và tích chập (C), ký hiệu B+R+C. Trong đó, ký hiệu C1 hoặc C3 thể hiện độ lớn của hàm nhân trong phép tích chập tương ứng là 1x1 hoặc 3x3. Kết thúc khối DB là lớp nơron để kết nộp thông tin theo dạng cộng tín hiệu (element-wise addition) đóng vai trò chuyển đổi các đặc trưng cho khối DB tiếp theo, ký hiệu lớp này là TS (transition layer), nó gồm các phép xử lý B+R+C1 và phép gộp tín hiệu đặc trưng dạng trung bình có kích thước 2x2 (average pooling, ký hiệu Pa). Khối DB(k) được minh hoạ ở Hình 2.2 có tham số k là số lượng các bộ xử lý UP trong khối.



Hình 2.2. Sơ đồ kết nối trong khối DB(k)

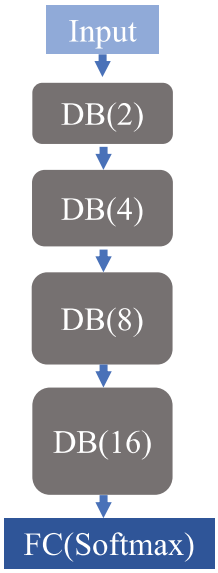
Mỗi khối DB có các kết nối trực tiếp một lớp nơron đến tất cả các lớp nơron tiếp theo trong khối, điều này giúp cải thiện luồng thông tin chuyển tải giữa các lớp nơron. Do đó, lớp nơron thứ h, tương ứng là bộ xử lý tín hiệu đặc trưng UP_h, nhận được các bản đồ đặc trưng của tất cả các lớp nơron trước đó trong khối, x_1, x_2, ..., $x_{(h-1)}$, dưới dạng đầu vào, và tín hiệu đặc trưng đầu ra của lớp này, x_h, được hình thức hoá như sau:

$$x_h = UP_h([x_1, x_2, \dots, x_{h-1}]) \qquad (1)$$

trong đó $[x_1, x_2, \dots, x_{(h-1)}]$ thể hiện việc ghép nối liên tục các bản đồ đặc trưng được tạo ra từ các lớp trước đó (từ *1* đến *h-1*). Như vậy, đầu ra của mỗi khối DB(k) được hình thức hoá dưới dạng hàm hợp thành từ nhiều bộ xử lý tín hiệu đặc trưng cùng với lớp nơron chuyển tiếp (TS) như sau:

$$DB(k) = TS(UP_k, UP_{k-1}, \dots, UP_1) \quad (2)$$

Trong kiến trúc này, số lượng và kích thước mỗi bộ xử lý tín hiệu đặc trưng (UP, số lượng nơron của mỗi UP) trong các khối DB tác động và ảnh hưởng đến chất lượng của tính năng trích xuất đặc trưng của mô hình đối với ảnh đầu vào, đồng thời chúng là những yếu tố tạo nên mức độ phức tạp của mô hình. Các tác giả thường điều chỉnh những yếu tố này để tạo nên một sơ đồ kết nối đầy đủ của mô hình CNN nhằm cân bằng giữa chất lượng nhận dạng và những điều kiện tính toán của môi trường ứng dụng thực tế. Mô hình LDFER trong nghiên cứu này (Hình 2.3) sử dụng 4 khối DB với kích thước lần lượt ở các khối là 2, 4, 8 và 16 có ký hiệu tương ứng là DB(2), DB(4), DB(8) và DB(16).



Hình 2.3. Sơ đồ các khối DB của LDFER

Như vậy, mô hình LDFER có tổng cộng 30 lớp noron tích chập xử lý trích chọn tín hiệu đặc trưng và được chia thành 4 khối DB. Mô hình này có 2.4 triệu tham số, ở mức thấp so với các mô hình CNN cho bài toán FER (Bảng 2.1). Mặc dù mô hình LDFER có nhiều lớp tích chập nhưng chúng tôi sử dụng ít số hàm nhân và kích thước hàm nhân nhỏ trong lớp tích chập, dẫn đến số lượng tham số mô hình ở mức thấp.

Bảng 2.1. So sánh độ lớn các mô hình

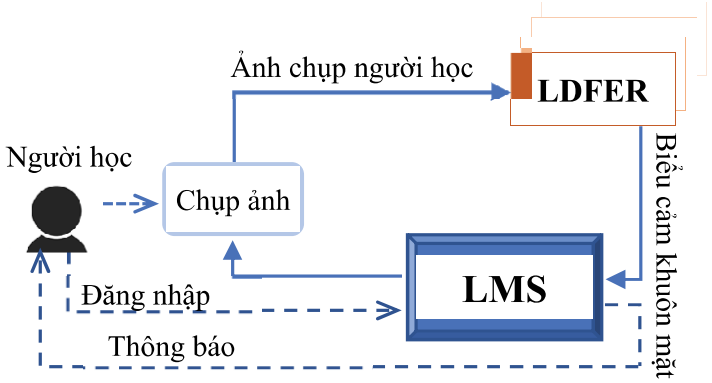
Mô hình	**Lớp tích chập**	**Tham số**
Zhao et al.16 in [20.Deng]	22	6.8M
Kuo et al.18 in [20.Deng]	6	2.7M
Liu et al.14 in [20.Deng]	6	2M
Dynamic Multi-task [20. Zhao]	20	13M
Deep Multi-task learning [20.Lam]	35	-
Lightweight CNN for FER [22.Devaram]	55	1.6M
Efficient CNN for FER [22.Lai]	23	2.5M
EfficientB3 [19.Tan]	16ϕ	10.7M
LDFER	60	2.4M

2.2. Thiết kế tích hợp với ứng dụng LMS

Dựa trên mô hình LDFER này, chúng tôi áp dụng thiết kế tích hợp với hệ thống LMS có sẵn để tự động chụp ảnh người học và nhận dạng biểu cảm khuôn mặt phục vụ cho việc ghi nhận, đánh giá các hoạt động của người học trong quá trình học tập trực tuyến [1]. Việc tích hợp được thực hiện theo kết nối giao diện lập trình ứng dụng (API) giữa các ứng dụng, phần mô hình LDFER được xuất bản thành mô-đun chạy trên thiết bị cá nhân (có thể web-client hoặc các app). Người học đăng nhập LMS thông qua tài khoản của học tập (định danh và mật khẩu) để xác thực cho việc học, hệ thống sẽ yêu cầu mở máy ảnh hoặc webcam của người học để ghi lại hoặc chụp ảnh khuôn mặt từ thiết bị học tập. Những hình ảnh này được gửi đến mô hình LDFER để xử lý và nhận dạng biểu cảm khuôn mặt. Quá trình này được lặp lại theo chu kỳ thời gian nhất định nhằm ghi nhận toàn bộ quá trình học tập. Tổng hợp kết quả nhận dạng được ghi nhận góp phần đánh giá chất lượng học tập, đánh giá nội dung học tập, đánh giá hoạt động giảng dạy của giảng viên và từ đó có các thông báo cho người học, giảng viên, người quản lý biết điều chỉnh hoạt động của mình đạt chất lượng cao hơn. Sơ đồ kết nối và quy trình vận hành của hệ thống tích hợp được trình bày trong Hình 2.4.

Hệ thống tích hợp này được thiết kế dưới dạng kết nối độc lập, không yêu cầu hệ thống LMS hiện có phải sửa đổi nhiều để kết nối với mô hình LDFER. LMS có thể được thực thi độc lập như vốn có của nó mà không cần kết nối với mô hình LDFER. Khi LMS được kết nối với mô

hình LDFER thì nó sẽ nhận được kết quả nhận dạng biểu cảm khuôn mặt của người học trong quá trình học tập và sử dụng kết quả này để tổng hợp, đánh giá, thông báo cho người học là do LMS quyết định thực hiện. Với cách thiết kế này, chúng ta có thể dễ dàng tích hợp mô hình LDFER vào bất kỳ LMS hiện có nào.



Hình 2.4. Sơ đồ kết nối mô hình LDFER với LMS

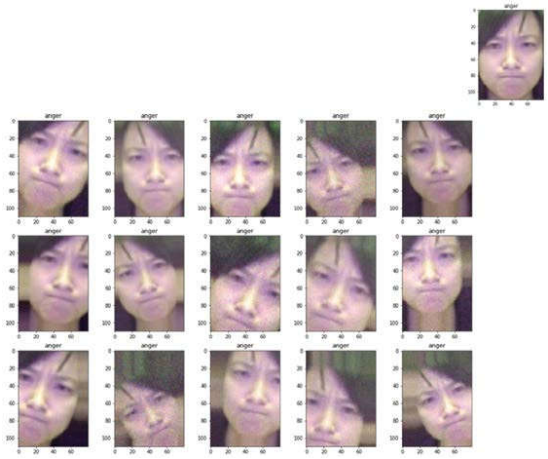
2.3. Tăng cường dữ liệu huấn luyện mô hình

Trong các ứng dụng thực tế, hình ảnh đầu vào thường được chụp từ thiết bị người dùng, chúng bao gồm nền với bất kỳ vật thể nào bên trong ảnh. Nghiên cứu này sử dụng mô hình dựa trên CNN nổi tiếng được gọi là MTCNN như trong [1] để xác định vùng ảnh có chứa khuôn mặt, sau đó cắt bỏ phần nền của ảnh và chỉ giữ lại vùng ảnh chứa khuôn mặt.

Để tránh hiện tượng quá khớp trong huấn luyện mô hình và giúp cho mô hình có khả năng nhận dạng cao hơn, chúng tôi tăng cường hình ảnh huấn luyện như trong [1] bằng cách sử dụng một số kỹ thuật xử lý hình ảnh 2D như thêm nhiễu, xoay, cắt và dịch chuyển, tăng cường độ sáng hoặc làm tối hình ảnh. Với hình ảnh đầu vào α, kết quả nhận được sau các phép tiền xử lý tăng cường ảnh như sau:

$$\{\Im^{\alpha}(f^{D}(a), p^{\alpha})\} \qquad (3)$$

trong đó, f^{D} là bộ dò tìm và phát hiện khuôn mặt trên ảnh, chẳng hạn MTCNN, p^{α} là các tham số cho hoạt động tăng cường hình ảnh với một phép xử lý α={nhiễu, xoay, co giãn, dịch chuyển, độ tương phản,...}, $\Im^{\alpha}$ biểu thị sự biến đổi của hình ảnh đối với phép xử lý tăng cường α. Chẳng hạn, Hình 2.5 dưới đây cho kết quả 15 hình được tăng cường với các tham số ngẫu nhiên từ một ảnh gốc ban đầu (nằm ở dòng đầu) trong dữ liệu OuluCASIA. Các hình ảnh được tạo ra rất đa dạng, do đó, khi huấn luyện sẽ tạo cho mô hình độ ổn định trích chọn đặc trưng khi thay đổi kiểu dáng, vị trí,... của ảnh chụp.



Hình 2.5. Một số hình ảnh tăng cường

Các tham số của các phép biến đổi tiền xử lý tăng cường hình ảnh được

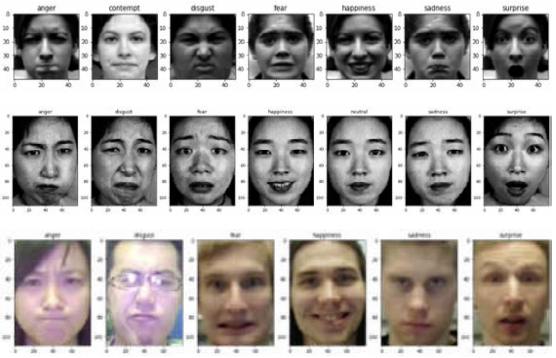
lựa chọn ở mức độ vừa phải để đảm bảo những thông tin chính trên ảnh được duy trì cho việc trích chọn đặc trưng cho bài toán. Chẳng hạn, ảnh thứ 4 ở dòng cuối có mức độ quay và dịch chuyển mạnh và có thể làm mất thông tin biểu cảm khuôn mặt nên rất khó để trích chọn đặc trưng và nhận dạng. Một hình ảnh tăng cường có thể áp dụng cùng lúc đồng thời các phép xử lý và trong nghiên cứu này chúng tôi áp dụng ngẫu nhiên các giá trị tham số điều chỉnh của các phép xử lý.

III. Thử nghiệm và kết quả

3.1. Dữ liệu và kịch bản thử nghiệm

Nghiên cứu này sử dụng ba bộ dữ liệu để thử nghiệm đánh giá mô hình LDFER gồm CK+ (Extended Cohn-Kanade), OuluCASIA và JAFFE.

Tập dữ liệu CK+ có 981 hình ảnh được thu thập từ 118 người khác nhau với bảy biểu cảm cơ bản gồm tức giận (anger), ghê tởm (disgust), sợ hãi (fear), hạnh phúc (happiness), buồn bã (sadness), ngạc nhiên (surprise) và sự khinh thường (contempt). Hình ảnh trong tập dữ liệu này có màu đa cấp xám (Hình 3.1, dòng đầu, tiêu đề của hình ảnh là nhãn biểu cảm tương ứng của người trong ảnh).



Hình 3.1. Một số ảnh trong các tập dữ liệu

Tập dữ liệu JAFFE chứa 213 ảnh biểu cảm khuôn mặt của 10 người phụ nữ khác nhau ở Nhật Bản. Mỗi người có các hình ảnh với sáu biểu cảm cơ bản khuôn mặt gồm tức giận, ghê tởm, sợ hãi, hạnh phúc, buồn bã và ngạc nhiên và hình ảnh có biểu cảm trung tính (neutral). Tập dữ liệu là một thách thức cho huấn luyện mô hình nhận dạng vì nó chứa quá ít hình ảnh cho mỗi loại biểu cảm, trung bình chỉ là 30. Dòng thứ 2 trong Hình 3.1 cho thấy hình ảnh của 7 biểu cảm khuôn mặt khác nhau trong tập dữ liệu này. Đây cũng là hình ảnh đa cấp xám.

Tập dữ liệu OuluCASIA có 1440 hình ảnh gồm sáu loại biểu cảm như trong tập dữ liệu CK+ trừ biểu cảm sự khinh thường (contempt). Nó được thu thập từ 80 người khác nhau và trong các điều kiện ánh sáng khác nhau, đây là hình ảnh màu. Dòng cuối trong Hình 3.1 cho thấy một số hình ảnh của tập dữ liệu Oulu-CASIA. Bảng 3.1 mô tả chi tiết phân bố các hình ảnh theo từng loại biểu cảm trong các tập dữ liệu thử nghiệm.

Bảng 3.1. Số ảnh của các tập dữ liệu

Biểu cảm	Số hình ảnh		
	CK+	JAFFE	Oulu CASIA
anger	135	30	240
contempt	54	-	-
disgust	177	29	240
fear	75	32	240
happiness	207	31	240
neutral	-	30	-
sadness	84	31	240
surprise	249	30	240
Tổng số	981	213	1440

Để chạy thử nghiệm, chúng tôi chia ngẫu nhiên mỗi tập dữ liệu thành 5 phần (fold) có kích thước tương đương nhau giữa các lớp nhận dạng (loại biểu cảm) của bài toán FER. Kịch bản thử nghiệm áp dụng kiểm tra chéo (cross-validation).

Trong mỗi lượt chạy huấn luyện mô hình, chúng ta sử dụng một 5 phần dữ liệu để kiểm tra và đánh giá kết quả mô hình (D^{te}), còn lại 4 phần để xây dựng mô hình, trong đó một phần dùng cho thẩm định và lựa chọn mô hình (D^{va}) và 3 phần còn lại được sử dụng để huấn luyện mô hình (D^{tr}). Kịch bản này được chạy lặp lại 5 lần theo thứ tự lần lượt các phần được chọn để kiểm tra mô hình, kết quả đánh giá cuối cùng là trung bình và độ lệch của 5 lần chạy.

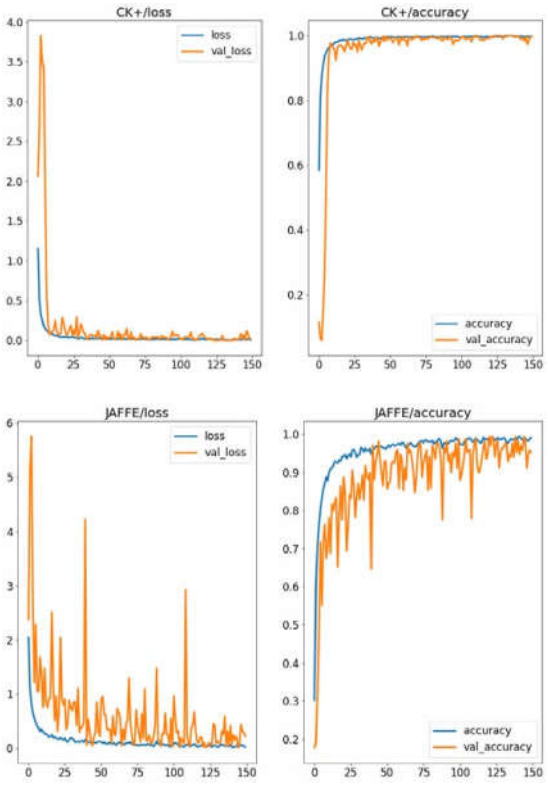
Trong mỗi lần chạy thử nghiệm, các phần dữ liệu huấn luyện mô hình (D^{tr}) được tăng cường bằng cách áp dụng các phép biến đổi hình ảnh $\Im^{\alpha}$. Các tham số cho mỗi phép biến đổi hình ảnh được chọn ngẫu nhiên trong khoảng giới hạn. Hệ số tăng cường là 10 cho mỗi ảnh gốc tạo nên tập dữ liệu huấn luyện lớn gấp 10 lần dữ liệu ban đầu nhằm đảm bảo độ đa dạng của dữ liệu, tránh bị hiện tượng quá khớp và kỳ vọng đạt được độ chính xác cao của mô hình. Các tham số tăng cường hình ảnh và huấn luyện mô hình được thể hiện chi tiết trong Bảng 3.2.

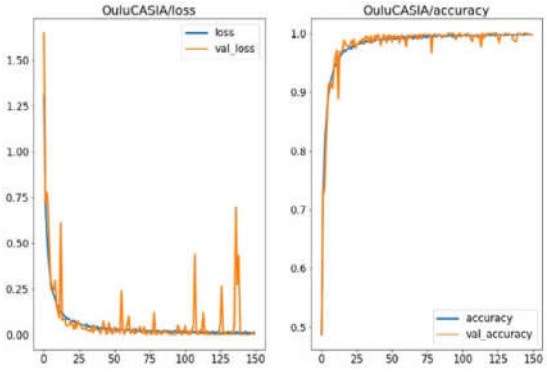
Bảng 3.2. Các tham số chạy thử nghiệm

Stt	Tham số	Giá trị
1	Góc quay tối đa so với ảnh gốc (radian, âm là quay sang trái)	±0.1π
2	Hệ số dịch chuyển tối đa so với kích thước ảnh gốc (âm là dịch sang trái)	±10%
3	Hệ số tương phản tối đa	0.1
4	Hệ số nhiễu tối đa theo phép nhiễu Gaussian	0.1
5	Hệ số co dãn tối đa so với kích thước ảnh gốc (giá trị âm là thu nhỏ)	±10%
6	Tốc độ học ban đầu (theo phương pháp Adam)	10^{-3}
7	Kích thước mỗi gói (batch) dữ liệu	128
8	Số lượt học mô hình (epoch)	150

3.2. Kết quả thử nghiệm

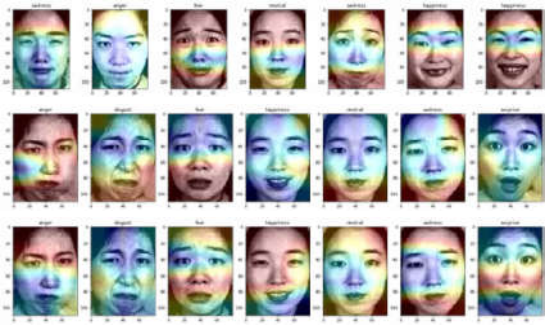
Quá trình huấn luyện mô hình LDFER gồm 150 lượt học (epoch) được tính trung bình trên 5 lần chạy thử nghiệm theo kịch bản cross-validation được thể hiện trong Hình 3.2. Mỗi cặp hình ảnh trên một dòng tương ứng là kết quả của hàm tổn thất (loss) và kết quả nhận dạng đúng (accuracy), chúng được tính trên cả hai phần dữ liệu để huấn luyện (D^{tr}) và dữ liệu để thẩm định, lựa chọn mô hình (D^{va}). Kết quả trên phần dữ liệu D^{va} của JAFFE (hai hình ở dòng giữa) có độ ổn định thấp hơn so với hai tập dữ liệu CK+ và OuluCASIA bởi vì JAFFE có quá ít hình ảnh và các biểu cảm không được thể hiện rõ nét.





Hình 3.2. Quá trình huấn luyện LDFER

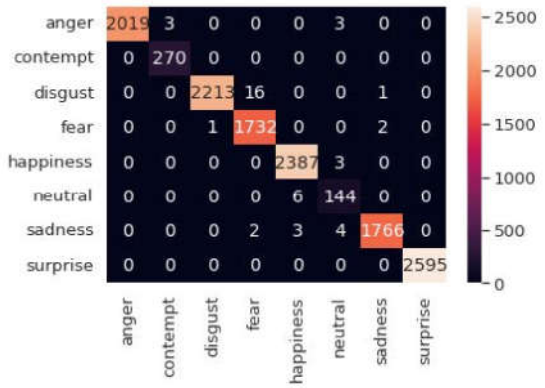
Trong mô hình LDFER, hoạt động của các khối nơron tích chập DB đóng vai trò trích xuất các đặc trưng trên ảnh khuôn mặt để nhận dạng. Ở đây, chúng tôi hiển thị các hình ảnh biểu diễn trực quan của việc hoạt động các khối DB thông qua phương pháp bản địa hóa dựa trên gradient (gradient-based localization). Phương pháp này cho thấy sự tập trung (hoặc sự quan tâm) của lớp nơron tích chập trên ảnh khi trích chọn đặc trưng, nó còn được gọi là bản đồ nhiệt của lớp đối tượng được kích hoạt trên vùng ảnh. Hình 3.3 cho thấy bản đồ nhiệt của lớp tích chập cuối cùng trong mô hình LDFER trên hình ảnh của từng loại biểu cảm trong tập dữ liệu JAFFE. Các hình ảnh đều có bản đồ nhiệt (chỗ màu sắc được sáng lên trên ảnh) hầu như tập trung vào những khu vực quan trọng để diễn tả biểu cảm trên khuôn mặt như vùng miệng, vùng mắt. Điều này trực quan cho thấy rằng mô hình LDFER tập trung vùng ảnh quan trọng để trích chọn đặc trưng mô tả cho biểu cảm khuôn mặt và ngược lại, khi không quan tâm đến những khu vực hình ảnh này thì khó có thể xác định được chính xác biểu cảm của người đó.



Hình 3.3. Bản đồ nhiệt trên dữ liệu JAFFE

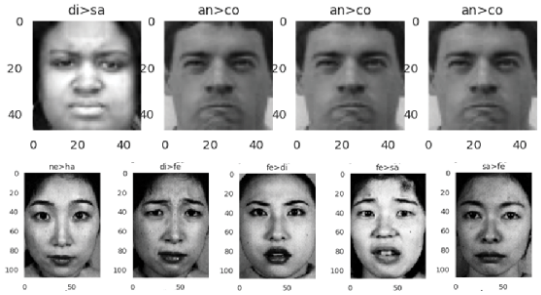
Để minh hoạ kết quả tổng thể của mô hình LDFER, chúng tôi xây dựng ma trận nhầm lẫn từ 5 lần chạy trên toàn bộ ba tập tập dữ liệu CK+, JAFFE và OuluCASIA (Hình 3.4). Mỗi hàng trong ma trận là một nhãn biểu cảm của hình ảnh trong dữ liệu, mỗi cột tương ứng là một nhãn biểu cảm được mô hình nhận dạng. Mỗi lần chạy huấn luyện ra mô hình được áp dụng để nhận dạng trên toàn bộ tập dữ liệu (gồm cả dữ liệu huấn luyện , thẩm định và đánh giá) và tính tổng số kết quả được nhận dạng. Vì vậy, mỗi hình ảnh trong tập dữ liệu được áp dụng 5 lần tương ứng với 5 mô hình được huấn luyện và tổng của một hàng trong ma trận đúng bằng số hình ảnh tương ứng biểu cảm đó trong tập dữ liệu nhân với 5. Ma trận này là tổng số kết quả trên cả 3 tập dữ liệu, trong đó, biểu cảm "contempt" chỉ xuất hiện ở tập dữ liệu CK+ và "neutral" chỉ xuất hiện ở tập dữ liệu JAFFE. Trong ma trận này, biểu cảm "surprise" không có trường hợp nhận dạng sai kể cả hình ảnh thuộc biểu cảm và không thuộc biểu cảm, chứng tỏ nó phân biệt tốt với các biểu cảm còn lại. Biểu cảm "contempt" được nhận dạng đúng đối với tất cả các hình ảnh thuộc biểu cảm, nhưng có 3 hình ảnh thuộc "anger" bị nhận dạng sai thành "contempt". Các biểu cảm còn lại có cả tình huống nhận dạng sai đối với hình ảnh thuộc và không thuộc biểu cảm đó. Tổng số trường hợp nhận dạng sai là 44, chiếm tỷ lệ 0,33% trong toàn

bộ 13170 lượt hình ảnh được nhận dạng, nhiều nhất có 17 hình ảnh thuộc biểu cảm "disgust" bị nhận dạng sai thành "fear" hoặc "sadness". Biểu cảm "fear" có nhiều nhất gồm 18 trường hợp nhận dạng nhầm từ các biểu cảm "disgust" và "sadness". Thông qua phân tích ma trận nhầm lẫn này cho thấy các biểu cảm "fear", "sadness" và "disgust" có mức độ nhầm lẫn giữa chúng là cao hơn các loại biểu cảm khác.



Hình 3.4. Ma trận nhầm lẫn trên 3 tập dữ liệu

Một số trường hợp nhầm lẫn được thể hiện trong Hình 3.5, tiêu đề trên ảnh ghi biểu cảm của ảnh (2 chữ cái đầu, trước dấu ">") và biểu cảm bị nhận dạng sai thành loại khác. Tập dữ liệu CK+ chỉ có tổng cộng 4 trường hợp (dòng đầu) trong khi đó tập dữ liệu JAFFE có 44 trường hợp (chỉ thể hiện 5 hình ảnh, dòng sau). Có thể thấy rằng các hình ảnh nhầm lẫn này cũng rất khó phân biệt bằng trực quan của chúng ta. Riêng dữ liệu OuluCASIA không có trường hợp nhầm lẫn.



Hình 3.5. Một số hình ảnh nhầm lẫn

Kết quả nhận dạng trên tập dữ liệu kiểm tra và đánh giá () của mô hình LDFER sau khi đã được huấn luyện của 5 lần chạy thử nghiệm thể hiện trong Bảng 3.2. Dòng đầu mỗi tập dữ liệu là kết quả trung bình và độ lệch chuẩn (chữ in đậm).

Để so sánh, chúng tôi chạy cùng kịch bản và tham số thử nghiệm đối với mô hình Efficient trong [19.Tan] với phiên bản cơ sở (B0) có kích thước nhỏ nhất. Kết quả thể hiện trong Bảng 3.3. Mặc dù số lượng tham số của mô hình Efficient có hơn 4 triệu tham số, nhiều hơn 70% so với mô hình LDFER nhưng kết quả nhận dạng của LDFER chỉ thấp hơn không đáng kể (0.1%) ở dữ liệu CK+. Đối với dữ liệu JAFFE và OuluCASIA, mô hình LDFER cao hơn đáng kể (2.18% ở JAFFE và 0.14% ở OuluCASIA) so với mô hình Efficient. Trường hợp đạt tối đa 100% ở mô hình LDFER trên dữ liệu OuluCASIA và mô hình Efficient trên dữ liệu CK+.

Bảng 3.2. Kết quả trên dữ liệu kiểm tra ()

Lần chạy	LDFER	Efficient
CK+	**99.90 (±0.002)**	**100**
Lần #1	100	100
Lần #2	100	100
Lần #3	100	100
Lần #4	99.49	100
Lần #5	100	100
JAFFE	**99.08 (±0.009)**	**97.62 (±1.683)**
Lần #1	100	97.62
Lần #2	100	97.62
Lần #3	97.62	95.24
Lần #4	100	97.62
Lần #5	97.78	100
Oulu CASIA	**100**	**99.86 (±0.309)**
Lần #1	100	100
Lần #2	100	100
Lần #3	100	100
Lần #4	100	99.31
Lần #5	100	100

So sánh với một số kết quả đã được công bố (Bảng 3.3) cho thấy mô hình LDFER đạt cao nhất ở hai tập dữ liệu JAFFE và OuluCASIA, còn tập dữ liệu CK+ đạt mức cao thứ hai. Các kết quả trong [19.Wang] và [20.Deng] được lấy trường hợp tốt nhất trong các mô hình được so sánh vì đây là nghiên cứu tổng quan. Ký hiệu sau dấu * của mô hình là kịch bản chạy thử nghiệm, 5F và 10F tương ứng là 5-folds và 10-folds trong phương pháp cross-validation, "No" là không có kịch bản chạy thử nghiệm và 0.2T thể hiện 20% số mẫu dữ liệu dùng để kiểm tra, đánh giá mô hình (testing).

Bảng 3.3. So sánh kết qủa các mô hình

Datasets Models	**CK+**	**JAFFE**	**Oulu CASIA**
The best in [19. Wang]*No	98.62	98.90	88.92
[19.Abdolrashidi] *0.2T	98.00	92.80	-
The best in [20. Deng]*10F	99.60	95.80	91.67
[20.Zhao]*10F	89.60	-	99.50
[20.Lam]*10F	97.85	-	89.23
[22.Devaram]*5F	84.27	80.09	-
[22.Lai]*5F	97.30	-	-
Efficient [19. Tan]*5F	**100**	97.62	
LDFER*5F	99.90	**99.08**	**100**

V. Kết luận

Trong nghiên cứu này, chúng tôi đã đề xuất một mô hình mạng nơron tích chập cho bài toán nhận dạng biểu cảm khuôn mặt (LDFER). Kiến trúc của mô hình dựa trên chuẩn kiến trúc kết nối lớp nơron tích chập dạng DenseNet. Mô hình này có độ sâu (số lớp nơron tích chập) vừa phải và số lượng tham số của mô hình thấp (ở mức 2.4 triệu), được gọi là mô hình thể nhẹ. Kết quả nhận dạng rất khả quan trên các tập dữ liệu thử nghiệm, đạt mức thấp nhất là 99.08% đối với tập dữ liệu JAFFE, cao nhất là 100% đối với tập dữ liệu OuluCASIA. So sánh với các kết quả khác cho thấy mô hình LDFER cao nhất ở 2 tập dữ liệu JAFFE và OuluCASIA, cao thứ hai ở tập dữ liệu CK+. Mô hình LDFER có thể áp dụng cho kết quả tốt trong các ứng dụng. Đặc biệt, nó ở thể nhẹ nên dễ dàng tích hợp trên các hệ thống có năng lực tính toán không đòi hỏi quá cao, phù hợp với đa dạng điều kiện trong thực tế nhưng vẫn cho kết quả tốt đối với các bài toán ứng dụng.

Chúng tôi cũng đã thiết kế hệ thống tích hợp mô hình LDFER vào hệ thống quản lý học tập trực tuyến (LMS) để hỗ trợ ghi nhận và đánh giá quá trình học tập trực tuyến của người học trên các hệ thống LMS. Theo đó, mỗi người học được ghi nhận chi tiết quá trình học tập, được đo đếm biểu cảm thể hiện trong suốt quá trình học tập, nếu có những bất thường hệ thống có thể tổng hợp báo cáo cho người dạy, người quản lý và hỗ trợ để nhắc nhở, giúp đỡ người học đạt kết quả học tập cao hơn. Việc tích hợp hệ thống này theo cơ chế mở, không gắn chặt với nhau, do đó, hệ thống hoạt động khá độc lập và có thiết kế đảm bảo tính an toàn, an ninh của dữ liệu và hệ thống kết nối tích hợp.

Trong những nghiên cứu tiếp theo, chúng tôi sẽ cải tiến tích hợp lai ghép giữa các kiến trúc hiện đại để đạt chất lượng cao hơn trích chọn đặc trưng của mô hình và thử nghiệm trên các tập dữ liệu phức tạp hơn để đánh giá.

Tài liệu tham khảo:

[1]. D.T.Long, "A Facial Expressions Recognition Method Using Residual Network Architecture for Online Learning Evaluation," *Journal of Advanced Computational Intelligence and Intelligent Informatics,* tập 25, số 6, pp. 1-10, 2021.

[2]. W.Deng và S. Li, "Deep Facial Expression Recognition: A Survey," *IEEE Transactions on Affective Computing,* tập 13, pp. 1195-1215, 2022.

[3]. M.Wang và W.Deng, "Deep Face Recognition: A Survey," *Neurocomputing,* tập 429, pp. 215-244, 2021.

[4]. S.-C. Lai, C.-Y. Chen và J.-H. Li, "Efficient Recognition of Facial Expression with Lightweight Octave Convolutional Neural Network," *Journal of Imaging Science and Technology,* pp. 040402.1-9, 2022.

[5]. A. Greco, N. Strisciuglio, M. Vento và V. Vigilante, "Benchmarking deep networks for facial emotion recognition in the wild," *Multimedia Tools and Applications,* pp. https://doi.org/10.1007/s11042-022-12790-7, 2022.

[6]. M. Tan và Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *Proceedings of the 36th International Conference on Machine Learning,* pp. 6105-6114, 2019.

[7]. G. Huang, Z. Liu, L. V. D. Maaten và K. Q. Weinberger, "Densely Connected Convolutional Networks," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* số ISSN:1063-6919, pp. 1-9, 2018.

[8]. G. Zhao, H. Yang và M. Yu, "Expression Recognition Method Based on a Lightweight Convolutional Neural Network," *IEEE Access,* tập 18, pp. 38528 - 38537, 2020.

[9]. R. R. Devaram và A. Cesta, "LEMON: A Lightweight Facial Emotion Recognition System for Assistive Robotics Based on Dilated Residual Convolutional Neural Networks," *Sensors,* tập 22, số 3366, pp. 1-20, 2022.

[10]. D.T.Long, "A Lightweight Face Recognition Model Using Convolutional Neural Network for Monitoring Students in E-Learning," *I.J. Modern Education and Computer Science,* tập 6, pp. 16-28, 2020.

[11]. N. Zhou, R. Liang và W. Shi, "A Lightweight Convolutional Neural Network for Real-Time Facial Expression Detection," *IEEE Access,* tập 9, pp. 5573 - 5584, 2020.

[12]. P. N. R. Bodavarapu và P. Srinivas, "An Optimized Neural Network Model for Facial Expression Recognition over Traditional Deep Neural Networks," International Journal of Advanced Computer Science and Applications, tập 12, số 7, pp. 443-451, 2021.

[13]. Y. Nan, J. Ju, Q. Hua, H. Zhang và B. Wang, "A-MobileNet: An approach of facial expression recognition," Alexandria Engineering Journal, tập 61, p. 4435–4444, 2022.

Địa chỉ tác giả: Trường Đại học Mở Hà Nội

Email: duongthanglong@hou.edu.vn