Tạp chí Khoa học Nông nghiệp Việt Nam 2026, 24(4): 506-513 https://doi.org/10.31817/tckhnnvn.2026.24.04.08
vie.vjas.vn

NHẬN DIỆN CẢM XÚC KHUÔN MẶT CỦA TRẺ EM BẰNG MÔ HÌNH HỌC SÂU

Nguyễn Quỳnh Trang, Phạm Văn Thắng, Nguyễn Trọng Kương*

Khoa Công nghệ thông tin, Học viện Nông nghiệp Việt Nam

**Tác giả liên hệ: ntkuong@vnua.edu.vn*

Ngày nhận bài: 13.10.2025 Ngày chấp nhận đăng: 12.03.2026

TÓM TẮT

Trí tuệ nhân tạo đã có những bước tiến mạnh mẽ và trở thành công cụ hữu ích hỗ trợ giải quyết các bài toán phức tạp trong mọi lĩnh vực của cuộc sống. Nghiên cứu này trình bày ứng dụng của mạng nơron tích chập để nhận diện hình ảnh khuôn mặt cảm xúc của con người. Quá trình tiền xử lý và nhận diện khuôn mặt đã sử dụng mạng nơron tích chập đa nhiệm để phát hiện và và nhận dạng các khuôn mặt. Mạng tích chập đa nhiệm giúp phát hiện đặc trưng khuôn mặt, sau đó để phân loại cảm xúc. Bộ dữ liệu sử dụng trong nghiên cứu này bao gồm các video lớp học của học sinh mầm non trong độ tuổi từ 3 đến 5 tuổi, được ghi lại trong các tiết học. Những video này cung cấp hình ảnh cảm xúc khuôn mặt của trẻ trong môi trường lớp học để huấn luyện mô hình đào tạo, phát hiện các khuôn mặt trên màn hình và đưa ra kết quả cảm xúc khi nhận diện. Nghiên cứu cho thấy mạng nơron tích chập đạt độ chính xác là 92% trên tập huấn luyện và 95% trên tập kiểm tra. Mạng nơron tích chập đã chứng tỏ khả năng học tốt hơn so với các mạng nơron thông thường cho cùng bài toán.

Từ khóa: Trí tuệ nhân tạo, mạng nơron tích chập, nhận diện cảm xúc khuôn mặt.

Children's Facial Emotion Recognition using Deep Learning Models

ABSTRACT

Artificial Intelligence (AI) has been making great strides and is increasingly becoming a useful tool in supporting the solution of complex problems in all areas of life. This study presented the application of convolutional neural networks (CNN) to recognize users' emotional images from videos of children's activities. The pre-processing and facial recognition process used Multi-Task Cascade Convolutional Network (MTCNN) to detect and recognize faces. MTCNN helps to detect distinctive faces, then provides data to CNN for emotion classification. The dataset used in this study consisted of videos recording the activities of children aged 3 to 5 years old at a preschool. These videos contained emotional images of children in a classroom, they were used to train the learning model to detect how emotional the children express from the videos. As the results, the CNN network achieved 92% accuracy on the training and 95% on the test. The CNN network demonstrated the better learning than conventional neural networks for the same tasks.

Keywords: Deep learning, convolutional neural networks, facial emotion recognition.

1. ĐẶT VẤN ĐỀ

Nhận diện cảm xúc trên khuôn mặt là phương thức giao tiếp quan trọng, phản ánh những suy nghĩ và trạng thái tâm lý bên trong của mỗi người. Khuôn mặt của con người biểu hiện nhiều cảm xúc mà không cần phải nói ra, đó là một trong những phương tiện mạnh mẽ và tự nhiên nhất để con người truyền đạt thể hiện cảm xúc. Trong nghiên cứu của Mehrabian & cs. (1971) chỉ ra rằng, về mặt hiệu quả giao tiếp, thông tin trao đổi qua các phương tiện ngôn ngữ, qua giọng điệu và qua phương tiện không bằng lời khác như ngôn ngữ cơ thể chiếm 7%, 38% và 55%, tương ứng. Bài toán nhận diện cảm xúc khuôn mặt là một trong những bài toán thú vị và thu hút nhiều nghiên cứu với kết quả tích cực trong lĩnh vực thị giác máy tính, ứng dụng rộng rãi của các bài toán này như giám sát trạng thái người lái xe, giám sát người dùng điện thoại, hệ

thống giám sát tại các cơ sở y tế và trong giáo dục (Ekman, 2007). Dù vậy, nhận diện cảm xúc khuôn mặt vẫn đối mặt với không ít trở ngại, một phần là do khuôn mặt của mỗi người mang nét riêng biệt, khiến cùng một cảm xúc lại được thể hiện theo nhiều cách khác nhau.

Ekman & cs. (1990) đã xác định niềm vui, sợ hãi, ghê tởm, buồn, ngạc nhiên và tức giận là sáu cảm xúc cơ bản (Ayham & cs., 2014; Mase, 1991; Jyotsna & cs., 2023). Thêm nữa, Fernández & cs. (2021) quan tâm đến biểu hiện bốn loại cảm xúc trên khuôn mặt là hạnh phúc, ngạc nhiên, giận giữ và căm phẫn trong một nghiên cứu về cảm xúc khuôn mặt liên quan đến bệnh Alzheimer. Tùy theo mỗi lĩnh vực quan tâm nghiên cứu mà sự phân lớp cảm xúc có sự khác nhau và cách thức tiếp cận khác nhau. Cũng vậy, việc định lượng và phân loại cảm xúc vẫn là tùy thuộc vào bài toán cụ thể. Trong nghiên cứu này, ba nhóm cảm xúc gồm hào hứng, không hào hứng và khác, tương ứng với vấn đề đánh giá tác động hiệu quả giáo dục của một bài học đến học sinh tham gia dựa trên đo biểu hiện tâm lý ở ba nhóm này.

Nghiên cứu của Jiahong & cs. (2023) đã đưa ra những đánh giá tổng quan về các nghiên cứu liên quan đến bài toán nhận diện khuôn mặt cảm xúc trong giáo dục và những thách thức đối với các bài toán liên quan trong lĩnh vực này. Thông thường, bài toán phân lớp các loại hình ảnh rất phổ biến trong mọi lĩnh vực của công nghiệp và đời sống. Trong đó phải kể đến các phương pháp học máy truyền thống như k láng giềng gần nhất (kNN), cây quyết định (decision tree), Bayes, support vector machine (SVM), rừng ngẫu nhiên (Random forest), hay mạng nơron truyền thống (multiple perceptron). Và, gần đây tập trung sự quan tâm của nhiều nghiên cứu là ứng dụng của học sâu (Deep learning - DL) (LeCun & cs., 2015).

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Thu thập dữ liệu

Dữ liệu trong nghiên cứu là các video về hoạt động trong các lớp học của một trường mầm non. Các tiết học bao gồm hoạt động của trẻ em có độ tuổi từ 3 đến 5 tuổi như học tập, vận động, vui chơi có sự hướng dẫn của giáo viên. Mục tiêu là phân tích biểu cảm khuôn mặt của trẻ trong các tình huống học tập và tương tác, từ đó nhận diện các cảm xúc như hào hứng, không hào hứng để đánh giá vai trò của bài học với nhận thức qua cảm xúc tâm lý. Giải quyết bài toán này góp phần trợ giúp cho việc tăng hiệu quả việc giám sát hoạt động giáo dục và tìm kiếm các bài học tốt hơn cho trẻ.

Quá trình thu thập dữ liệu có thể bao gồm video ghi lại các hoạt động như chơi nhóm, tham gia trò chơi học tập, hoặc trong các tình huống phản hồi giáo viên. Việc giám sát tự động qua các video này sẽ giúp thu thập những biểu hiện cảm xúc tự nhiên của trẻ mà không cần can thiệp hay gây ảnh hưởng đến hành vi của các em. Để có được dữ liệu này, nhóm nghiên cứu đã được sự đồng ý sử dụng dữ liệu từ phía cơ sở mầm non và phụ huynh có con theo học trong lớp.

2.2. Tiền xử lý dữ liệu

Tiền xử lý dữ liệu là bước quan trọng trong quy trình phân tích dữ liệu, đặc biệt là khi làm việc với dữ liệu hình ảnh như ảnh khuôn mặt trong nhận diện cảm xúc. Mục tiêu của tiền xử lý là chuẩn hóa và làm sạch dữ liệu nhằm đảm bảo tính đồng nhất, giảm nhiễu, giúp mô hình học sâu dễ dàng tiếp nhận, phân tích và trích xuất đặc trưng hiệu quả (Afolabi & cs., 2025). Khi làm việc với dữ liệu hình ảnh và video, bài toán khó khăn thường gặp phải là xử lý với dữ liệu lớn và tốn thời gian. Thêm nữa, video thường có độ phân giải cao và chứa các thông tin không cần thiết, chẳng hạn như sự thay đổi màu sắc, ánh sáng, hoặc góc cạnh, điều này có thể ảnh hưởng đến độ chính xác của việc nhận diện và phân lớp.

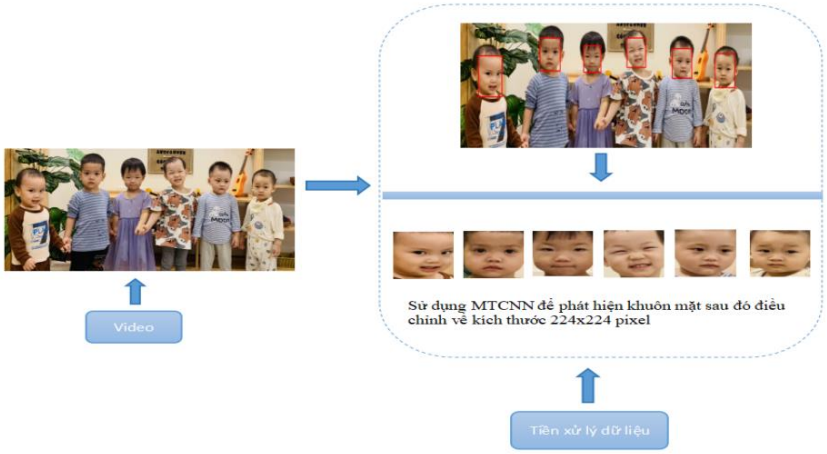
Liên quan đến vấn đề nhận diện đặc trưng khuôn mặt, mạng nơron tích chập đa nhiệm (MTCNN) là mô hình mạng nơron được nhiều nhà nghiên cứu liên quan đến lĩnh vực này quan tâm (Iván & cs., 2024). Mô hình này sử dụng ba bước: phát hiện khuôn mặt, xác định điểm đặc trưng của khuôn mặt, và phân loại các vùng khuôn mặt, giúp định vị và cắt các khuôn mặt trong video đảm bảo khuôn mặt được tách biệt

rõ ràng từ đó thu được một mảng khuôn mặt có trong ảnh để làm hình ảnh đầu vào cho bước nhận diện tiếp theo. Việc trích xuất chỉ lấy vùng hình ảnh khuôn mặt sẽ làm giảm bớt chi tiết dư thừa trong ảnh, nâng cao hiệu suất nhận diện. Trong nghiên cứu này, hình ảnh khuôn mặt sau khi cắt sẽ được điều chỉnh về kích thước chuẩn 224 × 224 pixel, giúp giảm thiểu độ phức tạp và đảm bảo tính đồng nhất cho mẫu dữ liệu. Dữ liệu thuộc tính khuôn mặt này là thuộc tính đầu vào cho các bước huấn luyện và phân lớp cảm xúc tiếp theo.

Hình 1 biểu diễn quá trình tiền xử lý từ đoạn video dữ liệu đầu vào đến trích xuất các frame ảnh, tiếp theo là xác định các vùng khuôn mặt nhờ MTCNN. Cuối cùng, hình ảnh các khuôn mặt trên từng frame ảnh được trích ra. Về nguyên lý, mỗi giây trong video bao gồm 24 frame ảnh. Vì vậy, trong mỗi khoảng thời gian thì chỉ một frame ảnh được trích ra làm đại diện cho biểu hiện cảm xúc tâm lý trong khoảng đó. Nghiên cứu này lấy khoảng trễ t = 10 giây, nghĩa mỗi frame ảnh được lấy ra cách nhau 10 giây để đảm bảo độ thay đổi biểu hiện cảm xúc trên khuôn mặt.

Sau khi trích xuất các đặc trưng khuôn mặt, một bước quan trọng nữa là gán nhãn cảm xúc tâm lý. Cụ thể dữ liệu sẽ được xem xét ở ba nhóm cảm xúc đó là hào hứng, không hào hứng, và những hình ảnh không phải là khuôn mặt trong quá trình MTCNN xác định nhầm. Bảng 2 "hào hứng" và "không hào hứng" bao gồm 2130 hình ảnh mỗi thư mục, thư mục nhãn "khác" bao gồm 1.800 hình ảnh, cả 3 thư mục chứa hình ảnh đều có độ phân giải 224 × 224 pixel. Các ảnh này đã được xác định và phân loại một cách rõ ràng dựa trên đặc trưng biểu cảm khuôn mặt đặc thù của từng loại cảm xúc.

Bảng 1 mô tả các cảm xúc cơ bản cùng với những biểu cảm khuôn mặt đặc trưng về nhận diện cảm xúc qua biểu cảm khuôn mặt dựa theo nghiên cứu của Ekman & cs. (1990). Mỗi cảm xúc sẽ tương ứng với một số dấu hiệu nhận diện cụ thể từ khuôn mặt, giúp việc phân loại và nhận diện cảm xúc trở nên chính xác hơn.



Hình 1. Quá trình tiền xử lý dữ liệu

Bảng 1. Đặc điểm cảm xúc trên khuôn mặt

Lớp cảm xúc	Biểu cảm khuôn mặt
Hào hứng	- Miệng mở rộng và cong lên, tạo thành một nụ cười (miệng cong lên hoặc có thể lộ răng). - Mắt có thể nheo lại hoặc mắt có thể sáng và mở to hơn khi cảm xúc mạnh mẽ. - Gò má nâng cao.
Không hào hứng	- Miệng có thể hơi mím hoặc kéo xuống, không có dấu hiệu cười hay vui vẻ. - Mắt có thể hơi mờ hoặc nhìn chán nản, thậm chí có thể liếc về một hướng khác.
Khác	- Không phải khuôn mặt.

2.3. Mạng nơron học sâu

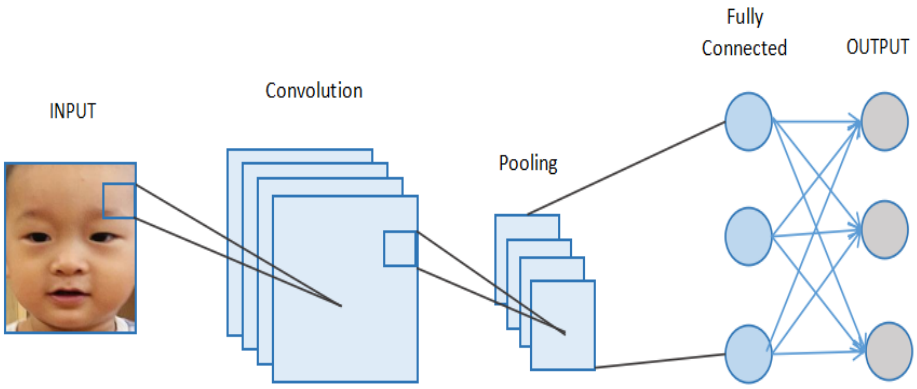
2.3.1. Mạng nơron tích chập

Mạng nơron tích chập (Convolutional Neural Network - CNN) là loại mạng nơron học sâu được thiết kế đặc biệt để xử lý dữ liệu có cấu trúc dạng lưới, điển hình là hình ảnh. Khác với các mạng nơron truyền thống yêu cầu đầu vào dạng vector phẳng, CNN có khả năng trực tiếp xử lý dữ liệu đầu vào 2D (hoặc 3D), giữ nguyên cấu trúc không gian của thông tin. CNN là một kiến trúc mạng được xây dựng từ các lớp đặc biệt như lớp chập (Convolution layer), lớp gộp (Pooling) và lớp kết nối đầy đủ (Fully-Connected), giúp mô hình học được các đặc trưng trong hình ảnh một cách hiệu quả. Lớp chập là thành phần quan trọng nhất trong mô hình mạng nơron tích chập. Lớp này có nhiệm vụ tiếp nhận và xử lý các hình ảnh đầu vào. Khi áp dụng phép toán tích chập (Convolution) vào xử lý, lớp chập giúp biến đổi thông tin đầu vào thành các yếu tố đặc trưng, từ đó phát hiện ra các mẫu hình trong hình ảnh như đường nét, màu sắc, hình dạng. Các yếu tố này rất quan trọng để mô hình có thể nhận diện các đối tượng trong ảnh. Lớp gộp là một lớp trong mạng CNN, có nhiệm vụ giảm kích thước dữ liệu sau khi qua lớp chập, nhằm giảm số lượng tham số và tính toán cần thiết mà vẫn giữ lại các đặc trưng quan trọng. Lớp này sử dụng một cửa sổ trượt quét qua toàn bộ ảnh dữ liệu và mỗi lần trượt sẽ lấy giá trị từ một vùng ảnh con. Các phương pháp phổ biến trong lớp gộp bao gồm MaxPooling (lấy giá trị lớn nhất trong vùng quét), MinPooling (lấy giá trị nhỏ nhất) và AveragePooling (lấy giá trị trung bình), giúp giảm thiểu sự ảnh hưởng của nhiễu và tạo ra các đặc trưng có độ bền cao hơn khi mô hình học.

2.3.2. Cấu trúc mô hình

Kiến trúc của mô hình CNN sử dụng phương pháp chia sẻ tham số nhằm giảm kích thước và độ phức tạp của mô hình để ứng dụng mô hình cho các bài toán có định dạng hạn chế. Để giảm kích thước mô hình, chúng tôi nghiên cứu thiết kế số lớp tích chập (Conv) là 3, sau mỗi loại tích chập sử dụng phép gộp tín hiệu ở dạng lớn nhất (MaxPool). Đầu tiên là ảnh đầu vào, để phù hợp với camera của hình ảnh thiết bị đầu cuối có độ phân giải vừa phải và giảm kích thước tham số của mô hình, hình ảnh được đặt kích thước đầu vào là H (cao) × W (rộng) × D (sâu) = 48 × 48 × 1.

Các bộ lọc của lớp Conv có kích thước là 3 × 3, và ở lớp MaxPool có kích thước là 2 × 2. Các lớp Conv sử dụng hàm kích hoạt "ReLu" nhằm cho phép loại bỏ các giá trị âm, tăng tốc độ huấn luyện của mô hình. Ngoài ra để giảm thiểu hiện tượng quá khớp trong trọng số của các nơron trong học sâu sử dụng kỹ thuật loại bỏ ngẫu nhiên kết nối của các nơron theo tỷ lệ 50% trong mỗi lớp Conv tăng dần kích thước của bộ lọc từ 32, 64, 128 nhằm tăng cơ hội trích chọn được nhiều các đặc trưng ẩn sâu bên trong hình ảnh ở các nơron mức độ sâu hơn.



Hình 2. Sơ đồ kiến trúc mạng CNN

Phần tiếp theo của bài toán là phân loại ảnh đầu vào đến các lớp tiếp theo. Khối này có hai lớp nơron được kết nối đầy đủ cho mỗi bài toán cần thực hiện, lớp đầu vào ẩn sử dụng hàm kích hoạt phi tuyến "ReLu" và lớp đầu ra kích hoạt hàm "softmax" để tính xác suất thuộc từng lớp cho mỗi hình ảnh đầu vào. Các đặc trưng đầu ra từ các lớp tích chập và pooling sẽ được chuyển đổi thành một vector phẳng thông qua lớp Flatten, trước khi được đưa vào các lớp phân loại fully connected. Đầu ra của mỗi nơron trong lớp phân loại được tính theo hàm kích hoạt softmax như trong công thức (1) dưới đây.

$$O_j^t = \text{soft}\max\left(y_j^t\right) = \frac{e^{y_j^t}}{\sum_{k=1}^{M^t} e^{y_k^t}} \quad (1)$$

Trong đó O_j^t là đầu vào của nơron thứ j^t của lớp ra tương ứng với t ∈{1, 2,..., T} là tổng các tín hiệu đầu vào của nơron thứ j^t trong lớp phân loại tương ứng với nhiệm vụ t, M^t là số nơron lớp ra của nhiệm vụ t. Ở đây, tổng các giá trị đầu ra được chuẩn hóa hay tổng $O_j^t = 1$

Mô hình huấn luyện được theo phương pháp tối ưu hóa Adam (Diederik & cs., 2015), đây là kỹ thuật tối ưu hóa được sử dụng rộng rãi bằng cách sử dụng các giá trị bình phương gradient để chia tỷ lệ học và sử dụng trung bình động của bước thay đổi gradient (Iván & cs., 2024). Cơ chế điều chỉnh trọng số của Adam thể hiệu theo công thức (2) dưới đây.

$$W_{ij^t} = W_{ij^{t-1}} - \frac{\eta}{\sqrt{v^t + \in}} \times m^t \quad (2)$$

Trong đó m^t và v^t là giá trị trung bình giảm theo cấp số nhân của gradient và của các bình phương gradient tại thời điểm học thứ t, η là hệ số học (hệ số huấn luyện). Trong nghiên cứu này, n được chọn bằng 10^{-8}.

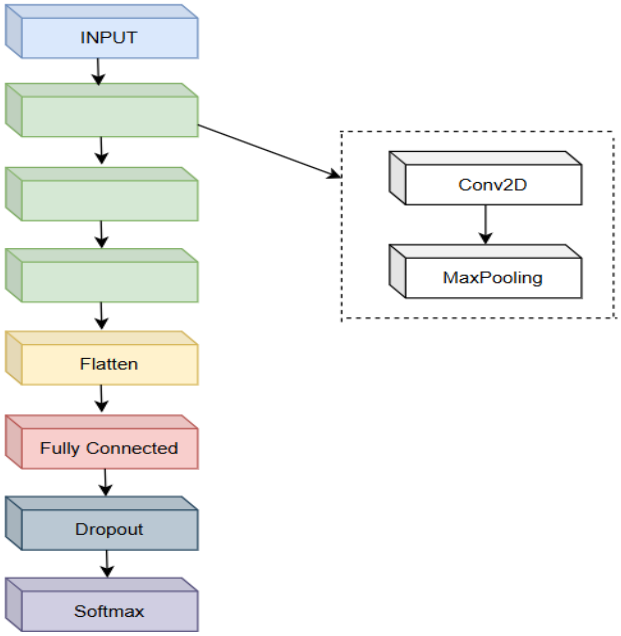
2.4. Huấn luyện mô hình

Quá trình xử lý dữ liệu là bước quan trọng giúp máy tính học từ hình ảnh. Tuy nhiên, do kích thước dữ liệu hình ảnh thường rất lớn, việc xử lý từng bức ảnh một cách riêng lẻ có thể làm giảm hiệu suất của mô hình và tiêu tốn tài nguyên không cần thiết. Điều này là do nhiều điểm ảnh trong ảnh có thể không mang lại thông tin hữu ích, chẳng hạn như những điểm ảnh không liên quan đến khuôn mặt. Thêm vào đó, mỗi điểm ảnh trong ảnh màu có thể có những giá trị màu sắc khác nhau, gây thêm độ phức tạp cho quá trình xử lý. Khi tập dữ liệu hình ảnh đã được tiền xử lý, bước tiếp theo là duyệt qua thư mục chứa dữ liệu để lấy thông tin cần thiết rồi dùng vòng lặp để duyệt từng ảnh trong thư mục. Sau đó thực hiện thay đổi kích thước hình ảnh thành 48 × 48 pixel rồi truyền xuống công đoạn xử lý bước tiếp theo là chuyển đổi các nhãn cảm xúc từ dạng chuỗi thành dạng số nhị phân, thuận tiện cho việc huấn luyện mô hình phân loại. Sử dụng phương pháp mã hóa one-hot sẽ chuyển mỗi nhãn thành một vector nhị phân, trong đó mỗi nhãn được biểu diễn bởi một vị trí duy nhất trong vector có giá trị 1, còn các vị trí còn lại có giá trị 0. Quy tắc mã hóa này tuân theo công thức tổng quát với nhãn y và phần tử thứ i của vector one-hot υ được xác định bởi công thức (3):

$$y \in \{0, 1, ..., n-1\} \rightarrow \upsilon = \underbrace{(0, ..., 1, ..., 0)}_{\text{vị trí i thì bằng 1}} \quad (3)$$

Sau khi đã chuyển đổi dữ liệu và nhãn thành dạng phù hợp cho quá trình huấn luyện, bước tiếp theo là xây dựng và huấn luyện mô hình học sâu. Trước tiên, dữ liệu sẽ được chia thành hai phần: một phần dùng để huấn luyện mô hình và một phần dùng để kiểm tra hiệu quả của mô hình. Quá trình chia dữ liệu được thực hiện bằng cách sử dụng hàm train_test_split từ thư viện Scikit-learn, với mục tiêu phân chia dữ liệu sao cho 80% dữ liệu được sử dụng cho việc huấn luyện và 20% còn lại để kiểm tra mô hình.

Việc chia dữ liệu này giúp kiểm soát quá trình học của mô hình, đảm bảo rằng mô hình có thể dự đoán chính xác trên dữ liệu chưa được sử dụng trong quá trình huấn luyện. Điều này giúp tối ưu hóa tham số của mô hình, ngăn chặn hiện tượng overfitting (học quá mức), giúp mô hình tổng quát tốt hơn khi xử lý dữ liệu thực tế.



Hình 3. Cấu trúc mô hình

Hình 3 mô tả cụ thể cấu trúc mạng nơron tích chập được xây dựng để phân loại ảnh đầu vào thành nhiều loại cảm xúc. Dữ liệu đầu vào của mạng là các ảnh có kích thước 48 × 48 pixel và có một kênh màu (ảnh xám). Lớp tích chập đầu tiên (Conv), mạng bắt đầu với một lớp tích chập sử dụng 32 bộ lọc, mỗi bộ lọc có kích thước 3 × 3, lớp này thực hiện quét ảnh đầu vào để trích xuất ra các đặc trưng cục bộ như cạnh, góc hoặc đường biên. Hàm kích hoạt được sử dụng là Relu, nhằm đưa tính phi tuyến vào mô hình, giúp mô hình học được các quan hệ phức tạp hơn giữa các điểm ảnh. Tiếp theo là một lớp Max pooling với kích thước 2 × 2, giúp giảm giảm kích thước dữ liệu sau khi qua lớp chập, nhằm giảm số lượng tham số và tính toán cần thiết mà vẫn giữ lại các đặc trưng quan trọng.

Lớp tích chập thứ hai sử dụng 64 bộ lọc, cũng có kích thước 3 × 3, và tiếp tục áp dụng hàm kích hoạt Relu. Việc tăng số lượng bộ lọc giúp mô hình học được nhiều đặc trưng hơn ở mức độ phức tạp cao hơn từ ảnh đã qua xử lý. Một lớp Max pooling khác với kích thước 2 × 2 tiếp tục được sử dụng để giảm chiều không gian của đặc trưng đầu ra, giúp làm cho mô hình nhỏ gọn hơn và ít nhạy cảm với dịch chuyển nhỏ trong ảnh. Mô hình có thêm một lớp tích chập (Conv) nữa với 128 bộ lọc để tiếp tục trích xuất các đặc trưng sâu hơn từ ảnh. Sau đó, lớp Flatten sẽ chuyển đổi đầu ra của các lớp tích chập thành một vector một chiều, giúp các lớp Fully connected xử lý.

Để giảm hiện tượng overfitting, mô hình sử dụng lớp Dropout với tỷ lệ 0,5, trong quá trình huấn luyện 50% số nơron trong lớp trước sẽ bị bỏ ngẫu nhiên trong mỗi lần cập nhật trọng số, nhằm tránh việc mô hình quá phụ thuộc vào một số đặc trưng nhất định.

Cuối cùng, một lớp Dense nữa được sử dụng với số lượng nơron bằng với số lớp cảm xúc cần phân loại. Lớp này sử dụng hàm kích hoạt softmax, giúp đưa ra một phân phối xác suất cho từng lớp, từ đó mô hình có thể chọn ra cảm xúc phù hợp nhất với ảnh đầu vào.

Để nâng cao khả năng khái quát và giảm thiểu tình trạng quá khớp (overfitting), dữ liệu huấn luyện được tăng cường bằng cách sử dụng lớp ImageDataGenerator, với các phép biến đổi hình ảnh như xoay, dịch chuyển, thay đổi kích thước và lật ảnh, nhằm giúp mô hình học hiệu quả từ nhiều biến thể của cùng một mẫu dữ liệu.

Để đánh giá hiệu quả học của mô hình CNN, nghiên cứu tiến hành so sánh với một mô

hình mạng nơron truyền thống (NN). Cả hai mô hình đều được huấn luyện trên dữ liệu ảnh xám có kích thước chuẩn hóa 48 × 48. Mô hình CNN được xây dựng với các tầng tích chập, tầng gộp (pooling) và các lớp kết nối đầy đủ (fully connected), cho phép khai thác đặc trưng không gian của ảnh một cách hiệu quả. Trong khi đó, mô hình NN có kiến trúc đơn giản hơn, gồm hai tầng ẩn với số lượng nơron lần lượt là 128 và 64. Cả hai mô hình đều sử dụng thuật toán tối ưu hóa Adam (Adam, 2014) để cập nhật trọng số trong quá trình huấn luyện.

3. KẾT QUẢ VÀ THẢO LUẬN

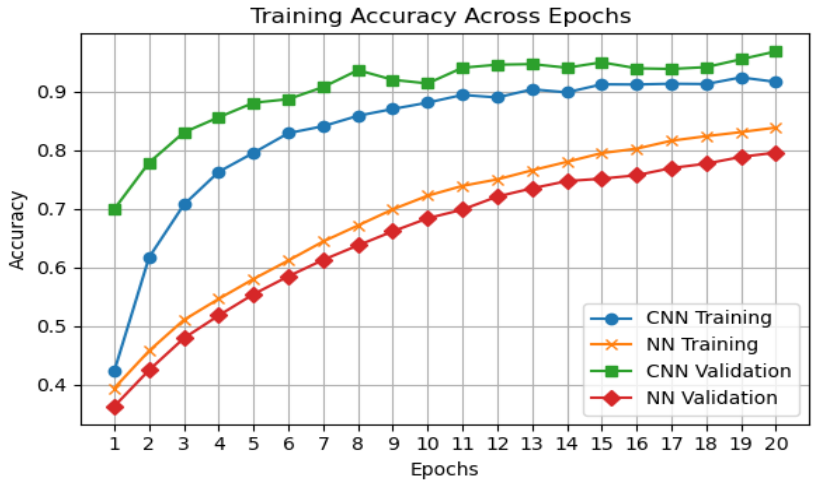
Độ chính xác phân lớp trên cả tập huấn luyện và kiểm tra của hai mô hình được theo dõi qua từng epoch. Kết quả thu được cho thấy CNN thể hiện hiệu suất vượt trội so với mô hình mạng nơron truyền thống trong quá trình học và phân loại cảm xúc từ dữ liệu hình ảnh. Trong quá trình huấn luyện, mô hình sử dụng bộ dữ liệu được thu thập và thực hiện huấn luyện trong 20 epoch. Độ chính xác của mô hình CNN trên tập huấn luyện (train set) đạt 95%, cho thấy mô hình đã học được khá tốt các đặc điểm của dữ liệu huấn luyện, độ chính xác trên tập kiểm tra (test set) đạt 92%. Mặc dù độ chính xác trên tập kiểm tra cao hơn tập huấn luyện, sự chênh lệch này vẫn trong phạm vi hợp lý, không có dấu hiệu overfitting.

Quá trình huấn luyện mạng được thực hiện một cách tuần tự, trong đó mỗi bước đại diện cho một lần cập nhật tham số mô hình dựa trên dữ liệu huấn luyện và thuật toán tối ưu hóa đã lựa chọn. Trong suốt quá trình này, mô hình liên tục cải thiện khả năng học thông qua việc điều chỉnh trọng số nhằm giảm giá trị hàm mất mát và nâng cao hiệu suất phân loại. Sau mỗi bước cập nhật, các chỉ số như độ chính xác và giá trị mất mát được ghi nhận để đánh giá hiệu quả học của mô hình tại thời điểm đó.

Trong nghiên cứu này, đã tiến hành 20 bước huấn luyện liên tiếp, giúp mô hình dần thích nghi với dữ liệu và khai thác được các đặc trưng phức tạp từ đầu vào. Việc theo dõi các chỉ số qua từng bước không chỉ hỗ trợ đánh giá quá trình học mà còn giúp kiểm nghiệm hiệu quả của các siêu tham số và thuật toán tối ưu đã sử dụng. Kết quả thu được cho thấy CNN có khả năng học đặc trưng từ dữ liệu hình ảnh hiệu quả hơn so với mạng NN truyền thống.

Bảng 2. Kết quả độ chính xác của tập huấn luyện và tập kiểm tra

	Huấn luyện	Kiểm tra	Tham số mô hình
CNN	95%	83%	85,830,944
NN	92%	79%	19,276,163



Hình 4. Độ chính xác trong quá trình đào tạo

4. KẾT LUẬN

Bài báo này tập trung nghiên cứu tìm hiểu và phát triển các kỹ thuật học sâu (Deep Learning - DL) kết hợp với các phương pháp tiền xử lý dữ liệu hiện đại. Mạng CNN được lựa chọn để giải quyết bài toán nhận diện cảm xúc, dựa trên những kiến thức nền tảng và kết quả từ các thí nghiệm thực tế. Bằng cách kết hợp giữa mô hình học sâu (CNN) và mô hình nhận diện khuôn mặt MTCNN, nghiên cứu đã đạt được những kết quả thực nghiệm khả quan, chứng minh tầm quan trọng của việc áp dụng mô hình DL đối với dữ liệu thực tế trong việc cải thiện độ chính xác.

Tuy nhiên nghiên cứu vẫn còn tồn tại một số hạn chế đáng chú ý. Trước hết, số lượng ảnh thu thập được còn tương đối ít và chưa thực sự đa dạng về biểu cảm khuôn mặt, chẳng hạn như "thờ ơ", "tức giận" hay "sợ hãi", dẫn đến việc mô hình chưa được huấn luyện đủ để nhận diện hiệu quả trên các trường hợp khác nhau. Bên cạnh đó, mặc dù các biểu cảm đã được phân loại, nhưng do ảnh hưởng của yếu tố ngoại cảnh như điều kiện ánh sáng, góc chụp, độ phân giải và độ che khuất khuôn mặt, mô hình vẫn gặp khó khăn và nhầm lẫn trong quá trình phân loại cảm xúc.

Kết quả của nghiên cứu này sẽ hỗ trợ cho các nghiên cứu sâu hơn trong tương lai để đi đến giải giải quyết bài toán liệu các phương pháp học sâu có thể triển khai trợ giúp cho việc nhận diện các khuôn mặt cảm xúc từ hình ảnh, video trong lớp học của học sinh mầm non hay không? Và tiếp theo nữa là đi đến giải quyết các bài toán về đánh giá hiệu quả của một bài học trong trường mầm non nói riêng và trong các hoạt động giáo dục nói chung dựa vào đánh giá các biểu hiện tâm lý từ học sinh. Điều quan trọng là với khả năng xử lý dữ liệu hình ảnh lớn, nhanh của phương pháp học sâu sẽ góp phần triển vọng xây dựng một công cụ tự động hỗ trợ phân tích, đánh giá tâm lý cảm xúc cho bài toán giám sát và đánh giá chất lượng giáo dục.

TÀI LIỆU THAM KHẢO

Adam K.D.B.J. (2014). A method for stochastic optimization. 1412(6). ArXiv. 1412.6980.

Afolabi I.A., Omolegho A.I., Idama O., Jones U.E. & Michael O.I. (2025). Effective preprocessing techniques for improved facial recognition under variable conditions. Franklin Open. 10: 100225.

Ayham F. & Anis Z. (2014). Novel Solution Based on Face Recognition to Address Identity Theft and Cheating in Online Examination Systems, Advances in Internet of Things. Scientific Research.

Diederik P.K. & Jimmy Lei Ba (2014). Adam: A method for stochastic optimization. ArXiv. 1412.6980.

Ekman P. (2007). Recognizing faces and feelings to improve communication and emotional life. Emotions revealed. Times Books. ISBN 0-8050-7275-6.

Ekman P., Davidson R.J. & Friesen W.V. (1990). The Duchenne smile: Emotional expression and brain physiology: II. Journal of Personality and Social Psychology. 58(2): 342-353.

Fernández R., Redolat M., Serra R.E. & González A.G. (2021). A systematic review of facial emotion recognition in Alzheimer's disease: A developmental and gender perspective. Anal. Psicol.

Iván d.P.C. (2024). Ipazc/mtcnn: v1.0.0., Zenodo. 13901378.

Jiahong S., Davy T.K.N. & Samuel K.W.C. (2023). Artificial Intelligence (AI) Literacy in Early Childhood Education: The Challenges and Opportunities. Computers and Education: Artificial Intelligence. 4: 100124.

Jyotsna C., Amudha J., Amritanshu R., Giandomenico N. (2023). IntelEye: An Intelligent Tool for the Detection of Stressful State based on Eye Gaze Data While Watching Video, Procedia Computer Science. 218: 1270-1279.

LeCun Y., Bengio Y. & Hinton G. (2015). Deep learning. Nature. 521(7553): 436-444.

Mase K. (1991). Recognition of facial expression from optical flow. IEICE Transactions. Information and Systems. 74: 3474-3483.

Mehrabian A. (1971). Silent Messages (1st ed.). Belmont, CA: Wadsworth. ISBN 0-534-00910-7.