PHÂN LỚP ÂM THANH MÔI TRƯỜNG SỬ DỤNG ĐẶC TRƯNG KẾT HỢP MFCC VÀ MEL-SPECTROGRAM VỚI MẠNG CNN

Thái Thuận Thương, Phạm Đình Trung Hiếu***

Title: *Environmental Sound Classification Using Combined MFCC and Mel-Spectrogram Features with CNN*

Từ khóa: *Học sâu, Mạng Nơ-ron tích chập, Phân lớp âm thanh môi trường, MFCC, Mel spectrogram ,Bộ dữ liệu âm thanh môi trường*

Keywords: *Deep Learning, Convolutional Neural Networks, Environmental Sound Classification, MFCC, Log-Mel Spectrogram, Environmental Sound Dataset*

Lịch sử bài báo

Ngày nhận bài: 10/03/2025

Ngày nhận kết quả bình duyệt: 10/03/2025

Ngày chấp nhận đăng bài: 03/04/2025

Tác giả: **Trường Đại học Yersin Đà Lạt, **Trường Đại học Việt Đức*

Email liên hệ: *it.vdean@yersin.edu.vn*

TÓM TẮT

Hiện nay, mô hình học sâu được sử dụng rộng rãi trong nhiều bài toán khác nhau và đã chứng tỏ sự vượt trội so với các phương pháp truyền thống. Đặc biệt, trong lĩnh vực phân lớp âm thanh môi trường, nhiều nghiên cứu đã đạt được kết quả đáng kể trong những năm gần đây. Tuy nhiên, hiệu quả của các mô hình sử dụng đặc trưng âm thanh như log-mel spectrogram (LM), hệ số cepstral tần số mel (MFCC) hoặc dạng sóng thô (waveform) để huấn luyện mạng nơ-ron sâu cho bài toán phân lớp âm thanh môi trường (ESC) vẫn chưa đạt yêu cầu. Trong bài báo này, chúng tôi đề xuất phương pháp kết hợp hai đặc trưng MFCC và Mel Spectrogram để tạo ra một biểu diễn toàn diện hơn cho âm thanh môi trường đươc gọi là đặc trưng MMS làm đầu vào để huấn luyện mạng với tên gọi là CNN-MMS, mô hình được huấn luyện trên bộ dữ liệu UrbanSound8K. Kết quả thực nghiệm cho thấy mô hình CNN-MMS được đề xuất cũng đạt 91% độ chính xác so với một số mô hình CNN được đề xuất trước đó trên cùng tập dữ liệu.

ABSTRACT

Nowadays, deep learning models have become widely adopted across various tasks, consistently outperforming traditional approaches. In the domain of environmental sound classification, recent studies have reported significant advancements. However, the effectiveness of existing models that rely on audio features such as log-mel spectrogram (LM), mel-frequency cepstral coefficients (MFCC), or raw waveforms to train deep neural networks for environmental sound classification (ESC) still falls short of expectations. In this study, we introduce a novel approach that integrates MFCC and Mel Spectrogram features to construct a more holistic representation of environmental sounds, termed the MMS feature. This feature is then used as input for training a convolutional neural network, named CNN-MMS, on the UrbanSound8K dataset. Experimental results indicate that the proposed CNN-MMS model achieves an accuracy of 91%, surpassing several previously introduced CNN models on the same dataset.

1. Đặt vấn đề

Nhận dạng âm thanh là một lĩnh vực quan trọng trong nhận dạng mẫu, với nhiều ứng dụng từ nhận dạng tiếng nói, truy vấn thông tin âm nhạc đến phân lớp âm thanh môi trường (ESC). ESC có vai trò thiết yếu trong nhiều ứng dụng thực tế như trợ thính, phân tích bối cảnh video, giám sát âm thanh, và giảm tiếng ồn đô thị.

Các phương pháp truyền thống sử dụng kỹ thuật trích xuất đặc trưng và mô

hình học máy như SVM, GMM, KNN cho ESC nhưng kết quả đạt được vẫn còn hạn chế do không có khả năng tự động trích xuất đặc trưng tối ưu. Trong khi đó, học sâu, đặc biệt là mạng nơ-ron tích chập (CNN), đã đạt nhiều thành tựu trong nhận dạng âm thanh nhờ khả năng tự động học đặc trưng từ dữ liệu thô.

Trong những nghiên cứu về ESC trước đó, thông thường cả hai phương pháp tiền xử lý tín hiệu và máy học đều sử dụng hệ số ma trận, học từ điển, ngân hàng bộ lọc sóng con và những đặc trưng Cepstral (nghĩa là những đặc trưng có được bằng cách biến đổi Fourier rời rạc DFT- Discrete Fourier Transform hoặc DCT - Discrete Cosine Transform của phổ tín hiệu) như GTCC - Gammatone Cepstral Coefficients (Sang, J., Park, S., & Lee, J., 2018). Quá trình này được gọi là "thiết kế đặc trưng", yêu cầu những kỹ thuật và kiến thức chuyên sâu về bài toán và lĩnh vực nghiên cứu. Ngoài ra, thiết kế đặc trưng thường là thiết kế theo kinh nghiệm và có thể không tối ưu cho ESC.

Nghiên cứu này đề xuất một phương pháp học sâu mới cho bài toán phân lớp âm thanh môi trường đô thị. Cụ thể, tác giả xây dựng mô hình CNN với đầu vào là bộ dữ liệu UrbanSound8K, sử dụng đặc trưng MFCC và Mel Spectrogram kết hợp thành đặc trưng MMS. Hiệu quả mô hình CNN-MMS sẽ được so sánh với các nghiên cứu trước để đánh giá khả năng cải thiện phân lớp âm thanh môi trường đô thị, tạo tiền đề cho các nghiên cứu tiếp theo.

2. Tổng quan nghiên cứu và phương pháp nghiên cứu

2.1 Tổng quan nghiên cứu

Trong những năm gần đây, số lượng các nghiên cứu về mô hình CNN ngày càng tăng. Điều này chứng tỏ rằng các mô hình CNN hoạt động vượt trội hơn phương pháp truyền thống trong những bài toán phân lớp khác nhau. Đáng chú ý là nghiên cứu của K. Piczak (Piczak, September 2015) trong nhận dạng âm thanh, đây là đánh giá đầu tiên về hiệu quả sử dụng CNN trong bài toán ESC. Nghiên cứu này đã đề xuất một hệ thống ESC với mô hình CNN bao gồm hai lớp max-pooling và hai lớp full connections. Đặc trưng âm thanh Log-mel spectrograms được rút trích để huấn luyện cho mô hình CNN này. Kết quả nghiên cứu cho thấy độ chính xác phân lớp đạt 71.7% cao hơn những mô hình truyền thống.

Năm 2018, nhóm nghiên cứu Zhichao Zhang và cộng sự đã để xuất một mô hình CNN dựa trên VGG Net sử dụng bộ lọc Convolution 1-D để học các mẫu qua tần số và thời gian tương ứng (Zhang, Z., Xu, S., Cao, S., & Zhang, S., 2018). Phương pháp này thực hiện tốt hơn CNN sử dụng bộ lọc tích chập 3×3 với mạng cùng độ sâu. Trong nghiên cứu này, nhóm tác giả đã đề xuất một mạng CNN sâu cho hệ thống ESC. Kiến trúc của mạng này là sử dụng các lớp convolutional và pooling để rút trích các đặc trưng cấp cao từ những đặc trưng quang phổ, đồng thời trộn hai mẫu được chọn ngẫu nhiên từ bộ dữ liệu ban đầu để làm dữ liệu huấn luyện cho mô hình CNN. Thực nghiệm được thực hiện trên bộ dữ liệu UrbanSound8K và đạt độ chính xác phân lớp tới 83.7%.

B. Zhu và cộng sự đã đề xuất một mô hình CNN cơ bản gọi là WaveNet sử dụng đặc trưng multi-scale để học toàn bộ

thông tin của âm thanh môi trường (Zhu, và những tác giả khác, 2018). Đầu tiên, đặc trưng được rút trích từ một file ghi âm thông qua lớp tích chập – Conv thứ nhất sử dụng ba kích thước bộ lọc khác nhau. Lớp Conv thứ hai sử dụng pool-size tương ứng để cho số chiều của những đặc trưng này bằng nhau và sau đó ba đặc trưng được nối thành dạng đặc trưng multi-scale. Thêm nữa, những đặc trưng này kết hợp với một Log-Mel Spectrogram và thực hiện tốt hơn những hệ thống khác trên bộ dữ liệu ESC-50.

2.2 Phương pháp nghiên cứu

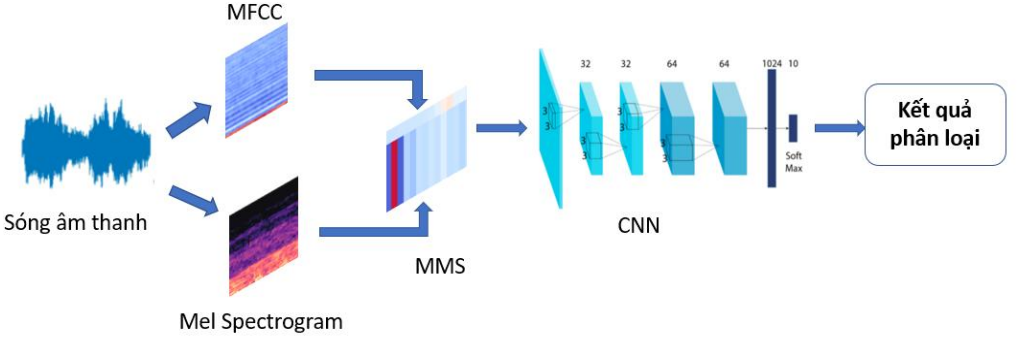
Trong bài báo này, chúng tôi áp dụng kỹ thuật Deep Learning, cụ thể là mô hình CNN-MMC (Hình 1), để phân lớp âm thanh môi trường. Khi một mẫu âm thanh có thời lượng vài giây dưới định dạng máy tính có thể đọc (file .wav) được đưa vào mô hình CNN- MMC sẽ xác định xem âm thanh đó có thuộc về lớp dữ liệu mục tiêu hay không. Nếu không nhận diện được lớp âm thanh cụ thể, mẫu đó sẽ được phân lớp vào nhóm unknown.

Dữ liệu đầu vào sẽ được rút trích các đặc trưng quan trọng như MFCC và Mel Spectrogram, sau đó kết hợp chúng để làm đầu vào huấn luyện cho mô hình. Để đánh giá hiệu suất mô hình, chúng tôi sử dụng độ chính xác phân lớp, được xác định bằng tỷ lệ dự đoán chính xác so với tổng số mẫu dự đoán.

$$\text{Độ } chính\ xác = \frac{Số\ lượng\ phân\ dự\ đoán\ đúng}{Tổng\ số\ dự\ đoán}$$

hay $\text{Độ } chính\ xác = \frac{TP+TN}{TP+TN+FP+FN}$;

trong đó TP = True Positives, TN = True Negatives, FP = False Positives, và FN = False Negatives.



Hình 1. Mô hình CNN với đặc trưng MMS

Độ chính xác phân lớp là thước đo tối ưu, đặc biệt trong trường hợp tập dữ liệu có sự phân bố đồng đều giữa các lớp. Hiệu suất của hệ thống chủ yếu phụ thuộc vào việc rút trích đặc trưng phù hợp, giúp biểu diễn chính xác các đặc điểm của âm thanh môi trường.

3. Nội dung

3.1. Tập dữ liệu và đặc trưng

3.1.1. Tập dữ liệu âm thanh môi trường đô thị UrbanSound8K

Để thực nghiệm và đánh giá mô hình, bài báo sử dụng tập dữ liệu *UrbanSound8K*, được xây dựng dựa trên các hệ thống phân lớp đã được đề xuất trước đó, với tổng thời lượng 27 giờ. Tập dữ liệu này bao gồm 8.732 đoạn âm thanh có nhãn, với độ dài và tần số lấy mẫu khác nhau. Các âm thanh trong tập dữ liệu được phân thành 10 lớp: *air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, police sirens, and street music.* Hầu hết các lớp có khoảng 1.000 đoạn âm thanh. Tuy nhiên, hai lớp *car horn* và *gunshot* có số lượng mẫu ít hơn đáng kể, lần lượt là 429 và 374 đoạn. Do đó, một số thuật toán học máy cơ bản được áp dụng trên tập dữ liệu cho thấy độ chính xác thấp hơn đối với hai lớp này. Tệp CSV của bô dữ liệu UrbanSound8K bao gồm 8 cột: *slice_file_name, fsID, start, end, salience, fold, classID* và *class*.

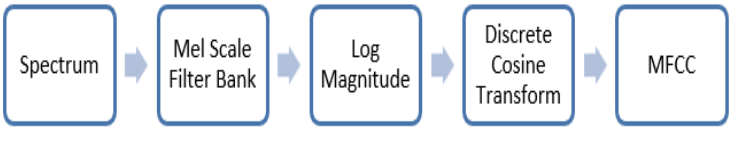
Bảng 1. Minh họa File Metadata

Num-ber	slice_ file_name	fsID	start	end	Salie-nce	fold	Class ID	class
0	100032-3-0-0.wav	100032	0	0.317551	1	5	3	dog_bark
1	100263-2-0-117.wav	100263	58.5	62.5	1	5	2	children_playing
2	100263-2-0-121.wav	100263	60.5	64.5	1	5	2	children_playing
3	100263-2-0-126.wav	100263	63	67	1	5	2	children_playing
4	100263-2-0-137.wav	100263	68.5	72.5	1	5	2	children_playing

3.1.2. Rút trích và kết hợp đặc trưng MFCC và Mel - Spectrogram

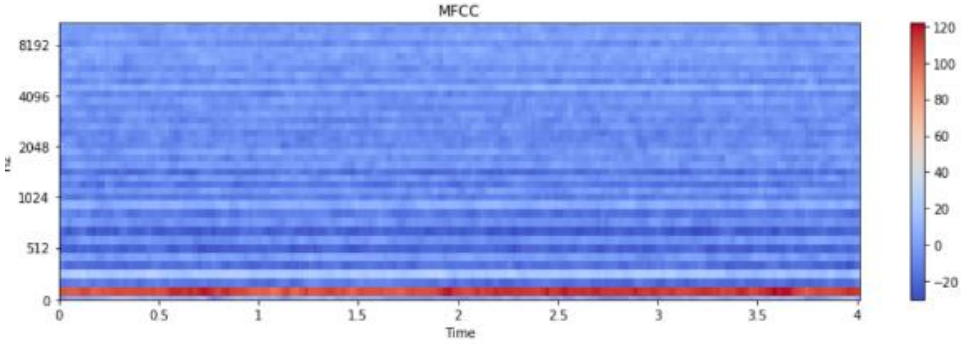
3.1.2.1 MFCC

Truyền một phổ qua ngân hàng bộ lọc Mel, sau đó tính toán độ lớn logarit và áp dụng Biến đổi Cosine Rời rạc (DCT), sẽ tạo ra Mel-Cepstrum. Quá trình DCT giúp trích xuất các thành phần thông tin chính của tín hiệu, trong đó các đỉnh phổ biểu thị những nội dung quan trọng nhất của âm thanh. DCT có thể được xem như một phép biến đổi nhằm chuyển đổi biên độ logarit của đầu ra ngân hàng bộ lọc Mel thành các hệ số cepstral.



Hình 2. Quy trình rút trích đặc trưng MFCC

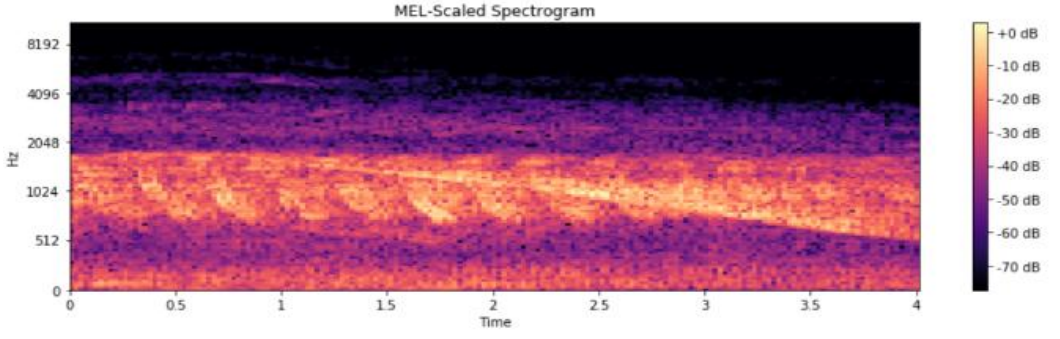
Thông thường, 13 hệ số đầu tiên được trích xuất từ Mel-Cepstrum được gọi là Mel-Frequency Cepstral Coefficients (MFCCs). Các hệ số này chứa thông tin quan trọng về đặc điểm quang phổ của tín hiệu âm thanh và thường được sử dụng làm đầu vào cho các mô hình học máy trong các bài toán nhận dạng và phân lớp âm thanh.



Hình 3. Hình ảnh đặc trưng MFCC

3.1.2.2 Mel-Spectrogram

Hình 3 minh họa một Mel-spectrogram, là sự kết hợp giữa thang Mel và phổ tần số, trong đó thang Mel biểu diễn phép biến đổi phi tuyến của thang tần số. Trong trường hợp này, tín hiệu âm thanh trước tiên được chia thành các khung nhỏ, sau đó áp dụng cửa sổ Hamming lên từng khung. Tiếp theo, Biến đổi Fourier Rời rạc (DFT - Discrete Fourier Transform) được sử dụng để chuyển tín hiệu từ miền thời gian sang miền tần số. Cuối cùng, phép toán logarit được áp dụng để tạo ra phổ, hỗ trợ trong quá trình tạo Mel-spectrogram



Hình 4. Hình ảnh đặc trưng Mel-Spectrogram

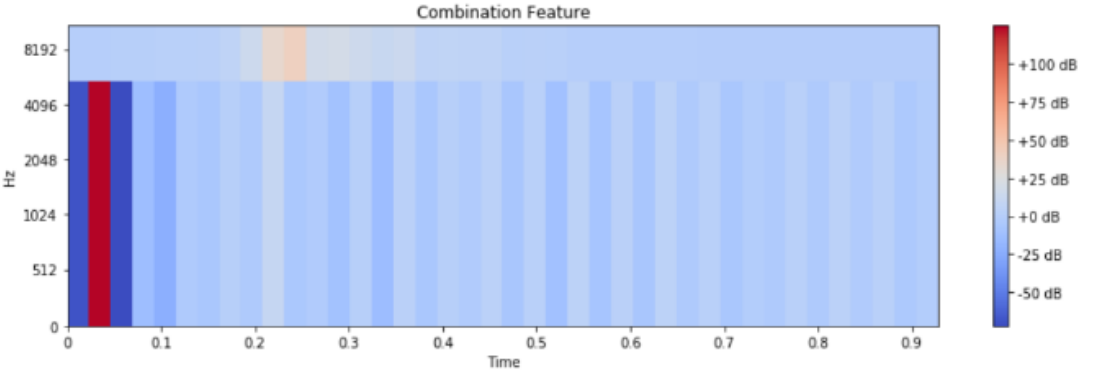
3.1.2.3 Kết hợp đặc trưng

Do sự tương đồng và tính bổ sung giữa Mel-Spectrogram và MFCC, việc hợp nhất các đặc trưng bổ sung giúp cải thiện đáng kể hiệu suất nhận dạng âm thanh. Cụ thể, phương pháp kết hợp giữa Mel Spectrogram - MFCC được thiết kế nhằm khai thác tối đa thông tin từ hai đặc trưng này.

Quy trình xây dựng đặc trưng MMS được thực hiện như sau: Đầu tiên, tín hiệu âm thanh gốc được chuyển đổi thành hệ số Mel-Frequency Cepstral Coefficients (MFCC) 40 chiều và Mel Spectrogram 40 chiều. Sau đó, để đảm bảo tính nhất quán về độ dài của

các đặc trưng đầu vào, cả hai đặc trưng trên được đệm giá trị 0 để phù hợp với tín hiệu âm thanh có độ dài lớn nhất trong tập dữ liệu (4 giây, tương ứng với 174 chiều).

Sau quá trình tiền xử lý, kích thước đầu ra của MFCC và Mel Spectrogram đều là 40 × 174. Tiếp theo, hai đặc trưng này được ghép nối theo chiều ngang để tạo thành đặc trưng MMS với kích thước 40 × 348. Việc kết hợp này giúp duy trì tính toàn vẹn của cả hai dạng biểu diễn âm thanh, đồng thời tăng cường khả năng phân biệt của mô hình nhận dạng trong các bài toán phân lớp âm thanh.



Hình 5. Hình ảnh đặc trưng MMS

3.2. Kiến trúc mô hình CNN-MMS

Kiến trúc mạng CNN-MMS chứa bốn lớp tích chập và một lớp kết nối đầy đủ như trong Hình 6 cụ thể như sau:

(1) Lớp thứ nhất sử dụng 32 hạt nhân (kernel) với 3 x 3 trường tiếp nhận và bước rộng được thiết lập là 2 x 2 và chuẩn hóa lô (batch-normalization) được thực hiện. Rectified Linear Unit (ReLU) được khai thác như là hàm kích hoạt.

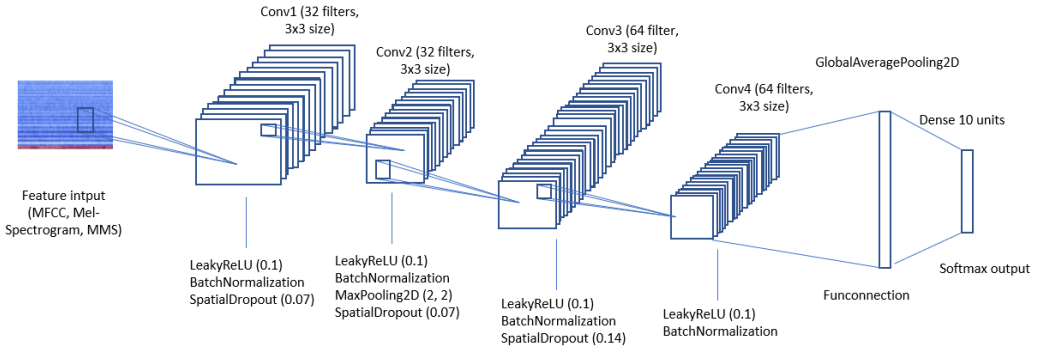
(2) Lớp thứ hai được dùng giống như những thiết lập ở lớp thứ nhất, trong đó 32 nhân tích chập với trường tiếp nhận 3 x 3 và bước rộng 2 x 2. Chuẩn hóa lô và hàm kích hoạt ReLU được thực hiện rất tốt. Sự khác biệt là lớp thứ hai áp dụng gộp (max-pooling) để giảm chiều của ánh xạ đặc trưng.

(3) Lớp thứ ba sử dụng 64 nhân tích chập với một trường tiếp nhận 3 x 3 và bước rộng cũng là 2 x 2, trong đó chuẩn hóa lô được sử dụng. Theo sau bởi hàm kích hoạt ReLU.

(4) Lớp thứ tư 64 nhân tích chập với trường tiếp nhận 3 x 3 và bước rộng 2 x 2. Chuẩn hóa lô được thực hiện và hàm kích hoạt là ReLU.

(5) Lớp thứ năm là lớp kết nối đầy đủ với 1024 đơn vị ẩn và hàm kích hoạt là Sigmoid. Các lớp Dense được thay thế bằng một lớp GlobalAveragePooling để tính toán trung bình đầu ra của mỗi ánh xạ đặc trưng trong lớp trước đó, làm giảm đáng kể kích thước.

(6) Đầu ra là 10 đơn vị theo bộ dữ liệu, theo sau bởi hàm kích hoạt softmax.



Hình 6. Kiến trúc chi tiết mô hình CNNN

3.3. Thực nghiệm và Kết quả

3.3.1. Thực nghiệm

3.3.1.1. Tiền xử lý và tách dữ liệu

Tập dữ liệu UrbanSound8K có sự khác biệt về số lượng kênh, tốc độ lấy mẫu và độ sâu bit, do đó cần được chuẩn hóa để đảm bảo tính nhất quán. Chúng tôi sử dụng hàm *load()* từ Librosa để chuyển tất cả tệp âm thanh về tốc độ lấy mẫu 22.05 kHz, chuẩn hóa biên độ tín hiệu trong khoảng [−1,1] và chuyển đổi về định dạng mono. Quá trình này giúp đơn giản hóa việc xử lý và trích xuất đặc trưng mà không làm mất thông tin quan trọng. Bảng 2 trình bày chi tiết các thông số tiền xử lý.

Bảng 2. Các thông số thuộc tính âm thanh

	Original - audio	**Librosa - audio**
Sample_rate	44100	22050
Bit-Depth	-23628 to 27507	-0.50266445 to 0.74983937
Channels	2	1

3.3.1.2. Rút trích đặc trưng

Chúng tôi xây dựng các phương thức *get_mel_spectrogram()* và *get_mfcc()* trong lớp *helper.py*, sử dụng hàm *melspectrogram()* và *mfcc()* từ Librosa để trích xuất đặc trưng Mel-Spectrogram và MFCC từ chuỗi dữ liệu âm thanh. Hàm *mfcc()* tạo 40 MFCC trên 174 frame, với các frame ngắn được bổ sung bộ đệm bằng 0 để đảm bảo độ dài nhất quán. Các đặc trưng sau đó được lưu trữ dưới dạng DataFrame bằng Pandas, thu được 8.732 đặc trưng MFCC. Quy trình tương tự được áp dụng cho Mel-Spectrogram, tạo 8.732 đặc trưng. Tiếp theo, chúng tôi sử dụng *np.vstack()* từ thư viện Numpy để kết hợp MFCC và Mel-Spectrogram, tạo đặc trưng MMS. Cuối cùng, tập đặc trưng (X) và nhãn (y) được chuyển đổi thành mảng Numpy và mã hóa bằng thư viện *sklearn*. Bảng 3 trình bày chi tiết kết quả trích xuất đặc trưng.

Bảng 3. Kết quả rút trích đặc trưng

Đặc trưng	**MFCC**	**Mel Spec-trogram**	**MMS**
Tổng số	8732	8732	8732

Chúng tôi chia bộ dữ liệu UrbanSound8K thành tập huấn luyện (Train) và tập kiểm tra (Test), trong đó Test chiếm 20% và được chọn ngẫu nhiên. Đồng thời, 20% tập Train được sử dụng làm tập Validation để đánh giá mô hình. Việc phân tách dữ liệu được thực hiện bằng hàm *train_test_split()* từ sklearn. Bảng 4 trình bày số lượng mẫu của từng tập.

Bảng 4. Số lượng mẫu trên tập dữ liệu

Dataset	Train	Validate	Test
8732	6986	1397	1746

3.3.1.3. Huấn luyện mô hình

Chúng tôi xây dựng mô hình CNN đơn giản bằng Keras/TensorFlow và SciKit-Learn trên Python, gồm bốn lớp Conv2D và một lớp Dense đầu ra.

- Conv2D: Bộ lọc 2×2, số kênh (16, 32, 64, 128), hàm kích hoạt ReLU.
- Đầu vào: (40, 174, 1) với 40 đặc trưng (MFCC, Mel-Spectrogram, MMS), 174 frame, và 1 kênh mono.
- Pooling:
  - MaxPooling2D giúp giảm chiều dữ liệu, tránh overfitting.
  - GlobalAveragePooling2D trích xuất thông tin quan trọng trước lớp đầu ra.
- Dense đầu ra: 10 nút, dùng Softmax để dự đoán xác suất lớp.

Chi tiết tham số mô hình được trình bày trong Bảng 5.

Bảng 5. Tham số mô hình CNN

Layers	Filters	Filter size	l2	ReLU	Batch Normalization	Spatial Dropout2D	Max pooling2D	Softmax
Conv1	32	3x3	0.0005	0.1	x	0.07		
Conv2	32	3x3	0.0005	0.1	x	0.07	2x2	
Conv3	64	3x3	0.0005	0.1	x	0.14		
Conv4	64	3x3	0.0005	0.1	x			
GlobalAverage Pooling2D								
Softmax output								x

Chúng tôi thiết lập optimizer và hàm mất mát để biên dịch mô hình. Trong đó, phương thức ADAM được sử dụng với các tham số mặc định do khả năng tối ưu hóa hiệu quả trong quá trình huấn luyện. Trong quá trình huấn luyện, một phần tập train được sử dụng làm tập validation. Cả ba mô hình CNN được huấn luyện với 360 epochs, batch size là 128, trên cùng một cấu hình phần cứng: Dell Workstation Precision M4800, CPU Intel Core i7-4810MQ (2.80GHz), RAM 12GB, GPU NVIDIA Quadro K1100M, ổ cứng SSD 500GB.

3.3.2. Kết quả và thảo luận

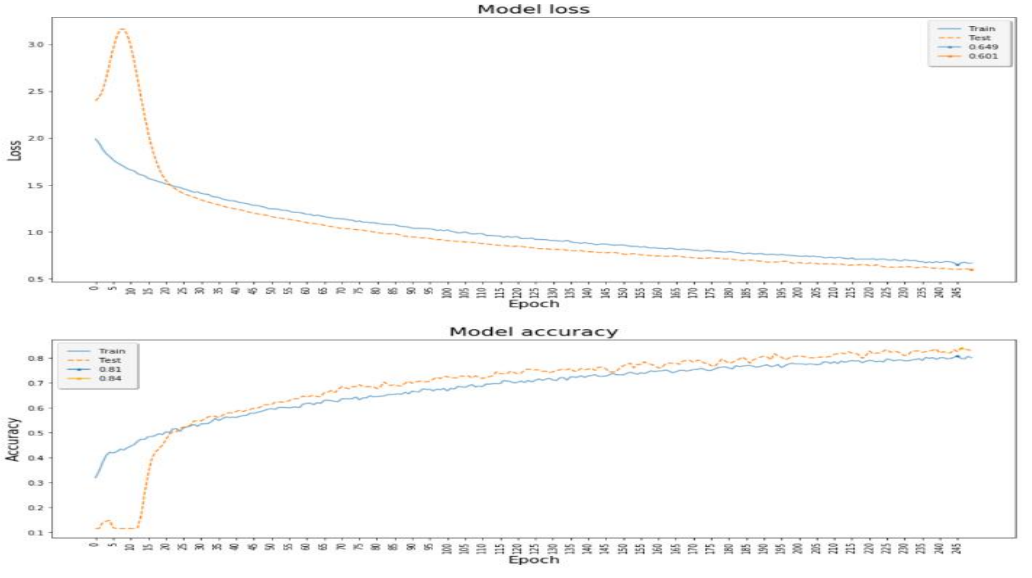
Bài báo đánh giá hiệu suất của các mô hình CNN dựa trên giá trị LOSS thấp nhất trong quá trình huấn luyện và ACCURACY trên tập kiểm tra (Test). Kết quả được trình bày trong Bảng 6.

Bảng 6. Kết quả đánh giá 3 mô hình CNN

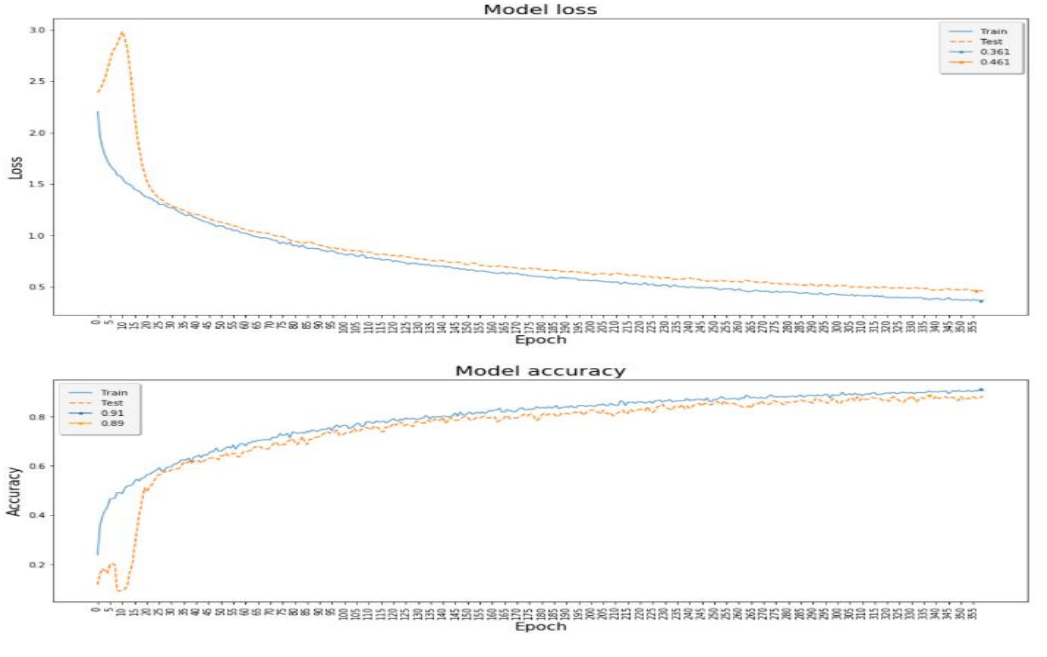
Dataset	Train			Test		
Model	**CNN-MFCC**	**CNN-Mel Spec**	**CNN-MMS**	**CNN-MFCC**	**CNN-Mel Spec**	**CNN-MMS**
LOSS	0.5764	0.2951	0.2832	0.6339	0.3221	0.3473
ACCURACY	84%	93%	94%	81.71%	91.26%	90.62%

Kết quả cho thấy CNN với Mel Spectrogram đạt 91.26% độ chính xác trên tập kiểm tra, cao hơn 9.55% so với CNN-MFCC và 0.64% so với CNN-MMS. Điều này chứng tỏ Mel Spectrogram là đặc trưng hiệu quả hơn trong phân lớp âm thanh môi trường. Bên cạnh đó, CNN-MMS cũng đạt hiệu suất cao, mở ra hướng nghiên cứu tiềm năng về kết hợp đặc trưng.

Quá trình huấn luyện của cả ba mô hình diễn ra ổn định, nhờ sử dụng batch size lớn và dropout, giúp giảm overfitting và giữ lỗi kiểm thử ở mức thấp. Đồ thị huấn luyện và kiểm thử được minh họa trong Hình 7, Hình 8 và Hình 9.

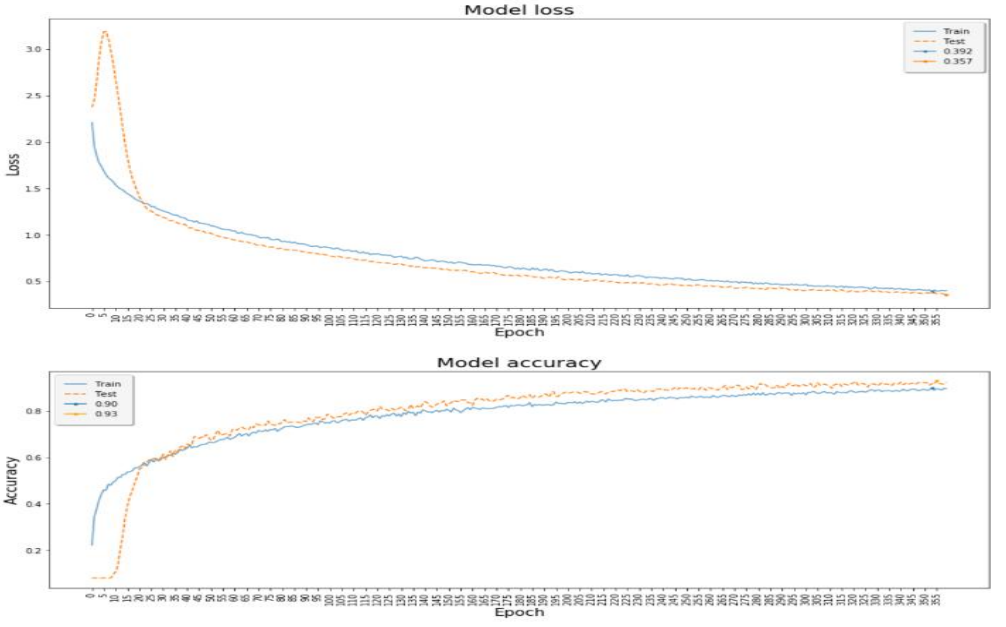


Hình 7. Đồ thị huấn luyện và kiểm thử mô hình CNN với MFCC



Hình 8. Đồ thị huấn luyện và kiểm thử mô hình CNN với Mel Spectrogram

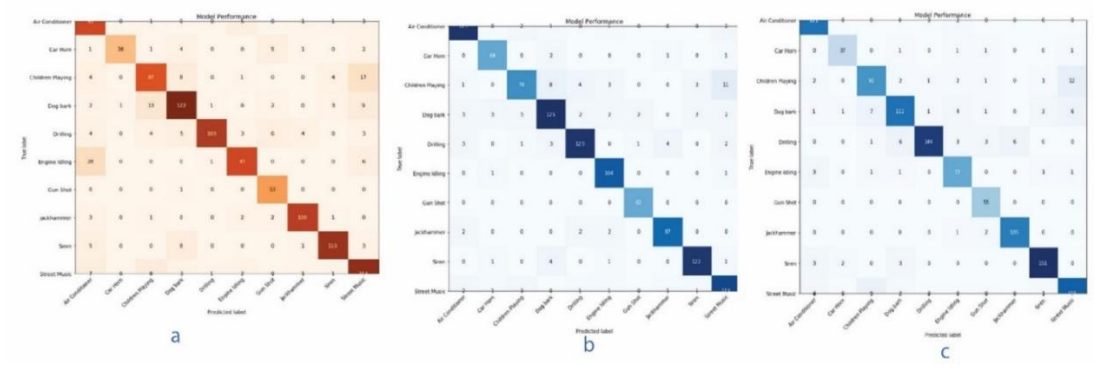
Chúng tôi đánh giá hiệu suất của ba mô hình CNN dựa trên khả năng phân lớp các lớp âm thanh trên tập kiểm tra (Test), vốn chưa được sử dụng trong quá trình huấn luyện. Kết quả được thể hiện qua các ma trận nhầm lẫn trong Hình 10.



Hình 9. Đồ thị huấn luyện và kiểm thử mô hình CNN với MMS

Phân tích ma trận nhầm lẫn cho thấy một số lớp có đặc điểm âm thanh tương đồng, gây ra nhầm lẫn khi phân lớp. Cụ thể:

- **"Street Music" và "Children Playing"** có nền âm thanh tương tự như tiếng người nói và âm thanh môi trường đô thị.
- **"Children Playing" và "Dog Bark"** đôi khi bị nhầm lẫn do một số mẫu của "Children Playing" chứa âm thanh chó sủa.
- **"Drilling" và "Jackhammer"** có tần số và biên độ tương tự nhau, gây khó khăn cho mô hình trong việc phân biệt.



Hình 10: (a) Ma trận nhầm lẫn của mô hình CNN – MFCC, (b) Ma trận nhầm lẫn của mô hình CNN – Mel Spectrogram, (c) Ma trận nhầm lẫn của mô hình CNN – MMS

Bảng 7, Bảng 8 và Bảng 9 trình bày độ chính xác của từng lớp âm thanh đối với ba mô hình CNN.

Bảng 7. Độ chính xác phân lớp của mô hình CNN với MFCC

ID	CLASS	ACCURACY
6	Gun Shot	96.00%
1	Car Horn	94.54%
8	Siren	93.38%
0	Air Conditioner	92.38%
7	Jackha m mer	87.40%
5	Engine Idling	81.52%
3	Dog bark	80.41%
4	Drilling	79.72%
9	Street Music	78.57%
2	Children Playing	69.82%

Bảng 8. Độ chính xác phân lớp của mô hình CNN với Mel Spectrogram

ID	CLASS	ACCU RACY
6	Gun Shot	97.61%
4	Drilling	95.21%
0	Air Conditioner	95.19%
7	Jackhammer	94.03%
8	Siren	92.70%
5	Engine Idling	91.39%
1	Car Horn	90.36%
3	Dog bark	87.09%
9	Street Music	85.86%
2	Children Playing	83.80%

Bảng 9. Độ chính xác phân lớp của mô hình CNN-MMS

ID	CLASS	ACCU RACY
6	Gun Shot	100.00%
0	Air Conditioner	98.23%
7	Jackhammer	97.22%
8	Siren	94.96%
5	Engine Idling	91.66%
1	Car Horn	90.24%
9	Street Music	88.40%
4	Drilling	88.34%
3	Dog bark	82.35%
2	Children Playing	81.25%

Kết quả phân lớp theo từng lớp cho thấy sự khác biệt đáng kể về hiệu suất giữa các mô hình. Cụ thể, CNN-MFCC đạt 96% độ chính xác đối với lớp "Gun Shot", nhưng chỉ đạt 69.82% với lớp "Children Playing". Trong khi đó, CNN-Mel Spectrogram cải thiện hiệu suất với 97.61% cho "Gun Shot" và 83.80% cho "Children Playing". Đáng chú ý, CNN-MMS đạt 100% độ chính xác đối với "Gun Shot" và cải thiện đáng kể độ chính xác trên hầu hết các lớp. Những kết quả này khẳng định rằng CNN-Mel Spectrogram và CNN-MMS hoạt động vượt trội hơn CNN-MFCC, đặc biệt đối với các lớp có dữ liệu mất cân bằng như "Gun Shot".

Bảng 10, Bảng 11 và Bảng 12 trình bày Precision, Recall và F1-score của ba mô hình. Kết quả cho thấy CNN-Mel Spectrogram và CNN-MMS có hiệu suất vượt trội hơn CNN-MFCC.

Bảng 10. Điểm Precision, Recall và F1 scores của mô hình CNN-MFCC

Class	Preci-sion	Recall	F1-score	Sup-port
Air Conditioner	0.78	0.92	0.85	105
Car Horn	0.73	0.95	0.83	55
Children Playing	0.74	0.7	0.72	116
Dog bark	0.82	0.8	0.81	143
Drilling	0.89	0.8	0.84	148
Engine Idling	0.83	0.82	0.82	92
Gun Shot	0.94	0.96	0.95	47
Jackhammer	0.97	0.87	0.92	127
Siren	0.93	0.93	0.93	136
Street Music	0.77	0.79	0.78	140
accuracy	-	-	0.84	1109
macro avg	0.84	0.85	0.84	1109
weighted avg	0.84	0.84	0.84	1109

Bảng 11. Điểm Presicion, Recall và F1 scores của mô hình CNN-Mel Spectrogram

Class	Preci-sion	Recall	F1-score	Support
Air conditioner	0.87	0.95	0.91	208
Car Horn	0.90	0.90	0.90	83
Children Playing	0.88	0.84	0.86	210
Dog bark	0.88	0.87	0.87	186
Drilling	0.93	0.95	0.94	188
Engine Idling	0.97	0.91	0.94	186
Gun Shot	0.92	0.98	0.95	84
Jackhammer	0.96	0.94	0.95	218
siren	0.98	0.93	0.95	192
Street Music	0.84	0.86	0.85	191
accuracy	-	-	0.91	1746
mac.co avg	0.91	0.91	0.91	1746
weighted avg	0.91	0.91	0.91	1746

Kết quả đánh giá F1-score cho thấy CNN-MMS đạt F1-score 0.91, phản ánh khả năng phân lớp ổn định trên tất cả các lớp. Tương tự, CNN-Mel Spectrogram cũng đạt F1-score 0.91, chứng minh hiệu suất tương đương với CNN-MMS. Trong khi đó, CNN-MFCC có F1-score thấp hơn (0.84), cho thấy mô hình này kém chính xác hơn khi xử lý các lớp có đặc trưng phức tạp.

Bảng 12. Điểm Precision, Recall và F1 scores của mô hình CNN-MMS

Class	Preci-sion	recall	fl-score	support
Air Conditioner	0.90	0.98	0.94	113
Car Horn	0.93	0.90	0.91	41
Children Playing	0.83	0.81	0.82	112
Dog bark	0.88	0.82	0.85	136
Drilling	0.99	0.88	0.93	163
Engine Idiling	0.86	0.92	0.89	84
Gun Shot	0.87	1.00	0.93	55
Jackhammer	0.95	0.97	0.96	108
Siren	0.96	0.95	0.96	159
Street Music	0.86	0.88	0.87	138
accuracy	-	-	0.91	1109
macro avg	0.90	0.91	0.91	1109
weighted avg	0.91	0.91	0.91	11 09

Bảng 13 trình bày so sánh độ chính xác của mô hình CNN-MMS với các phương pháp trước đó. Kết quả cho thấy CNN-MMS đạt độ chính xác cao nhất (91%), vượt trội so với các phương pháp trước đấy. Điều này khẳng định hiệu quả của việc kết hợp đặc trưng MFCC và Mel Spectrogram trong mô hình CNN.

Bảng 13. Kết quả so sánh độ chính xác phân lớp

Model	Features	Accuracy
Piczak	LM	71.7%
CRNN	RawWaveforms	78.3%
TDSN	Mel Spec	77%
VGG Net	Mels + Gammatone Spec	83.7%
CNN	MMS	91%

Kết luận

Nghiên cứu này đã tập trung vào bài toán phân lớp âm thanh môi trường bằng cách ứng dụng mô hình học sâu, cụ thể là mạng nơ-ron tích chập (CNN). Chúng tôi đã xây dựng kiến trúc CNN và tiến hành huấn luyện trên tập dữ liệu UrbanSound8K, bao gồm 10 loại âm thanh môi trường thường gặp trong đô thị như: *Air Conditioner, Car Horn, Children Playing, Dog bark, Drilling, Engine Idling, Gun Shot, Jackhammer, Siren, Street Music*. Dữ liệu đầu vào được xử lý thành ba dạng đặc trưng âm thanh: MFCC, Mel

Spectrogram và đặc trưng kết hợp MMS để đánh giá mức độ ảnh hưởng của từng phương pháp trích xuất đặc trưng đến hiệu suất của mô hình.

Kết quả thực nghiệm cho thấy rằng cả ba mô hình CNN sử dụng các đặc trưng MFCC, Mel Spectrogram và MMS đều đạt hiệu suất cao trong bài toán phân lớp âm thanh môi trường. Trong đó, mô hình CNN sử dụng đặc trưng Mel Spectrogram đạt độ chính xác 91%, cao hơn 7% so với mô hình CNN-MFCC. Đồng thời, mô hình CNN-MMS được đề xuất cũng đạt độ chính xác 91%, chứng minh tính hiệu quả của phương pháp kết hợp đặc trưng trong việc tăng cường hiệu suất nhận dạng. Những kết quả này không chỉ khẳng định ưu thế của mô hình CNN trong bài toán phân lớp âm thanh môi trường mà còn mở ra hướng nghiên cứu mới trong việc kết hợp nhiều đặc trưng để tối ưu hóa hiệu suất nhận dạng âm thanh.

TÀI LIỆU THAM KHẢO

Ballan, L., Bazzica, A., Bertini, M., Del Bimbo, A., & Serra, G. (2009). Deep networks for audio event classification in soccer videos. *In 2009 IEEE International Conference on Multimedia and Expo*, pp. 474-477.

Chu, S. N. (2009). Environmental sound recognition with timefrequency audio features. *Institute of Electrical and Electronics Engineers Inc*.

Graves, A., Mohamed, A.-r., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *In 2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645-6649.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., & Kingsbury,, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *The shared views of four research groups. IEEE Signal processing magazine*, 29(6), 82-97.

Khamparia, A., G., N., Khanna, A., Pandey, B., & Tiwari, P. (2019). Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network. *IEEE Access*, 7717-7727.

Li, S., Yao, Y., Hu, J., Liu, G., Yao, X., & Hu, J. (2018). An Ensemble Stacked Convolutional Neural Network Model for Environmental Event Sound Recognition. *Applied Sciences*, 8(7), 1152.

Sang, J., Park, S., & Lee, J. (2018). Convolutional recurrent neural networks for urban sound classification using raw waveforms. *In 2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 2444-2448.

Su, Y., Zhang, K., Wang, J., & Madani, K. (2019). Environment sound classification using a two-stream CNN based on decision-level fusion. *Sensors*, 19(7), 1733.