PHÂN LOẠI ÂM THANH NHẠC CỤ DỰA TRÊN ĐẶC TRƯNG PHỔ VÀ MẠNG NƠ-RON TÍCH CHẬP

Phạm Thị Thu Trang[1*], Dương Tấn Nghĩa[2]
[*]Tác giả liên hệ, Email: ptttrang@hou.edu.vn

Ngày tòa soạn nhận được bài báo: 02/06/2025
Ngày phản biện đánh giá: 08/09/2025
Ngày bài báo được duyệt đăng: 15/10/2025

DOI: 10.59266/houjs.2025.729

Tóm tắt: *Trong những năm gần đây, bài toán phân loại âm thanh đã thu hút sự quan tâm đáng kể nhờ tiềm năng ứng dụng rộng rãi trong lĩnh vực kỹ thuật và việc trích xuất đặc trưng âm thanh đóng vai trò then chốt trong quá trình phân loại này. Bài báo này trình bày hệ thống phân loại âm thanh nhạc cụ sử dụng kết hợp các phương pháp trích chọn đặc trưng âm thanh phổ biến và các mô hình học sâu. Tín hiệu âm thanh được chuyển đổi thành các biểu diễn phổ thời-tần số như Short-Time Fourier Transform (STFT), Mel-spectrogram, hệ số MFCC và MFCC kèm hệ số delta. Các đặc trưng này được đưa vào ba mô hình CNN: SimpleCNN, MobileNetV2 và VGG16. Thí nghiệm trên dữ liệu Solo MUSIC cho thấy Mel-spectrogram và MFCC+delta cho kết quả chính xác và F1-score cao nhất. MobileNetV2 đáp ứng tốt yêu cầu tính toán nhẹ nhưng vẫn giữ được hiệu quả.*

Từ khóa: *Trí tuệ nhân tạo, Học sâu, Phân loại âm thanh, Trích chọn đặc trưng, Mạng nơ-ron tích chập*

I. Đặt vấn đề

Trong những năm gần đây, hệ thống thông minh dựa trên âm thanh đã thu hút sự quan tâm lớn, đặc biệt là bài toán phân loại âm thanh nhạc cụ – chủ đề quan trọng trong xử lý tín hiệu âm thanh và âm nhạc số. Bài toán này có nhiều ứng dụng thiết thực như nhận diện nhạc cụ trong bản nhạc, hỗ trợ giảng dạy âm nhạc, giám sát âm thanh và tổng hợp dữ liệu cho các mô hình âm thanh.

Phân loại âm thanh hướng đến việc xác định và chia các tín hiệu đầu vào thành những nhóm âm thanh đã định nghĩa. Nguồn đầu vào thường là các tệp (.wav) được ghi lại từ nhạc cụ, tiếng môi trường hoặc các loại âm khác. Một trong những thách thức lớn là sự trùng lặp về đặc trưng

[1] Trường Đại học Mở Hà Nội
[2] Đại học Bách khoa Hà Nội

giữa các loại âm thanh, sự thay đổi do môi trường thu âm và thiết bị ghi âm.

Giải pháp phổ biến là chuyển đổi âm thanh sang miền đặc trưng trước khi đưa vào mô hình phân loại. Trong nhiều nghiên cứu gần đây (Li et al., 2001; Boddapati et al., 2017; Cowling et al., 2003; Bountourakis et al., 2015; Bountourakis et al., 2019), quy trình chung thường gồm hai bước: phân tích tín hiệu và trích xuất đặc trưng. Bước trích xuất giúp giảm nhiễu và chỉ giữ lại các thông tin có giá trị cho mô hình. Các đặc trưng thường chia thành hai nhóm: dạng sóng (Gong et al., 2023; Naman et al., 2025) và phổ (Wu et al., 2018; Pons et al., 2017; Choi et al., 2016). Nhóm đầu xử lý dữ liệu âm thanh ở dạng tín hiệu 1 chiều, trong khi nhóm phổ dùng biến đổi Fourier để tạo đại diện tần số–thời gian.

Li et al. (2001) cho thấy đặc trưng LPC và MFCC có hiệu quả cao trong phân loại âm thanh. Dhanalakshmi et al. (2009) phân loại sáu loại âm thanh sử dụng tổ hợp các đặc trưng như LPC, MFCC, LPCC. Sau bước trích chọn đặc trưng, mô hình học máy hoặc học sâu sẽ đảm nhận nhiệm vụ phân loại.

Mặc dù đã có nhiều nghiên cứu liên quan đến việc áp dụng các đặc trưng và mô hình học sâu trong phân loại âm thanh, vẫn còn thiếu các nghiên cứu so sánh trực tiếp giữa các phương pháp trích xuất đặc trưng dưới dạng hình ảnh khi áp dụng lên mô hình CNN, đặc biệt với dữ liệu âm thanh nhạc cụ. Hiệu quả phân loại phụ thuộc đáng kể vào cách trích chọn đặc trưng và kiến trúc mô hình sử dụng.

Nghiên cứu này tập trung so sánh bốn phương pháp trích xuất phổ biến gồm: STFT, Mel Spectrogram, MFCC và MFCC+Delta, áp dụng lần lượt cho ba mô hình CNN: SimpleCNN, MobileNetV2 và VGG16. Mục tiêu là xác định sự kết hợp nào giữa phương pháp trích xuất và mô hình mạng cho hiệu suất cao nhất khi phân loại âm thanh nhạc cụ.

Phần tiếp theo của bài báo được trình bày như sau: Phần II tổng quan các cơ sở lý thuyết; Phần III mô tả chi tiết các mô hình sử dụng và quy trình huấn luyện; Phần IV trình bày kết quả thực nghiệm và phân tích hiệu suất; Phần V kết luận và định hướng nghiên cứu tương lai.

II. Cơ sở lý thuyết

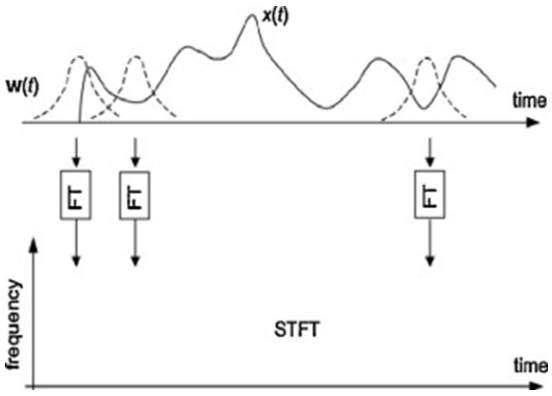
2.1. Các phương pháp biến đổi và trích chọn đặc trưng âm thanh nhạc cụ

2.1.1. Phương pháp STFT và Mel Spectrogram

Short-Time Fourier Transform (STFT) là một kỹ thuật phân tích tín hiệu trong miền thời gian–tần số, cho phép ta quan sát cách năng lượng tín hiệu phân bố theo tần số thay đổi theo thời gian. Về cơ bản, STFT thực hiện việc chia tín hiệu dài thành các đoạn con ngắn, sau đó áp dụng phép biến đổi Fourier rời rạc (DFT) lên từng đoạn để thu được phổ tần của từng khung thời gian.

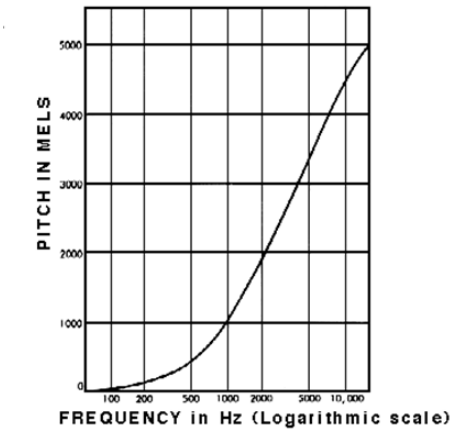
Trước khi thực hiện DFT, mỗi khung $x_m[n]=x[n+mR]$ được nhân với một hàm cửa sổ $w[n]$. Thường sẽ sử dụng hàm cửa sổ Hamming. Về mặt hình thức, STFT có các đặc tính cục bộ do sự tồn tại của hàm cửa sổ $w(t)$, do đó nó vừa là hàm thời gian vừa là hàm tần số. Trong một khoảng thời

gian nhất định, STFT của tín hiệu có thể được coi là phổ tại thời điểm đó, tức là phổ cục bộ.



Hình 1: STFT

Năm 1937, Stevens, Volkmann và Newmann đề xuất một đơn vị cao độ trong đó sự khác biệt tương tự bằng với người nghe đặt tên là Mel Scale. Điều khiển Mel Scale trở thành nền tảng cơ bản trong các ứng dụng Machine Learning (ML) đối với âm thanh là vì nó bắt chước nhận thức của con người về âm thanh. Chúng ta thực hiện phép toán trên tần số để biến nó thành thang đo Mel Scale.

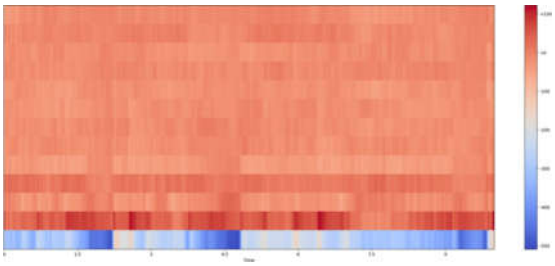


Hình 2: Thang đo Mel Scale

Mel Spectrogram là một biểu đồ tần số giống như Spectrogram tiêu chuẩn, nhưng trục tần số y được thay thế bằng giá trị Mel Scale của nó. Điều này cho phép chúng ta biểu diễn mức độ hiện diện của các tần số theo cách cảm nhận của chúng ta về âm thanh, giúp đạt được độ phân giải tốt hơn ở các tần số thấp.

2.1.2. Phương pháp Mel-frequency Cepstral Coefficients (MFCC)

MFCC là phương pháp trích xuất đặc trưng tần số mô phỏng cách tai người cảm nhận âm thanh. Quá trình gồm: chia tín hiệu thành các khung nhỏ (thường dùng cửa sổ Hanning), áp dụng DFT để thu phổ công suất, sau đó chiếu phổ này lên các bộ lọc tam giác theo thang Mel—tuyến tính dưới 1 kHz và logarit phía trên—phản ánh độ nhạy tần số của con người. Kế tiếp, phổ năng lượng được log hóa và biến đổi DCT để thu được các hệ số cepstral, thể hiện đặc trưng hàm lọc (formant) của âm thanh. Thông thường, chỉ giữ 12 hệ số cepstral đầu tiên, cộng thêm hệ số thứ 13 đại diện cho năng lượng khung, giúp phân biệt các mức cường độ. MFCC nhờ vậy tạo ra véc-tơ đặc trưng cô đọng, hiệu quả cho nhận dạng giọng nói và phân loại âm thanh.



Hình 3: Ảnh phổ MFCC

Tuy nhiên, âm thanh không hoàn toàn tĩnh, mà thay đổi theo thời gian. Để bắt được đặc trưng động, người ta bổ sung thêm các hệ số delta (Δ) và delta-delta (ΔΔ) cho mỗi khung. Delta phản ánh tốc độ thay đổi giữa các khung, còn delta-delta là gia tốc. Với N = 2 (số khung lùi/tiến), mỗi khung có tổng cộng 39 đặc

trưng: 12 cepstral, 1 năng lượng, 12 Δ, 12 ΔΔ, 1 Δ năng lượng và 1 ΔΔ năng lượng. Điều này giúp mô hình học sâu hiểu rõ cả phổ tĩnh lẫn chuyển động ngắn hạn—rất quan trọng cho dữ liệu phi tĩnh như giọng nói hay nhạc cụ. Trong nghiên cứu này, ta gọi MFCC gốc gồm 13 hệ số là MFCC, còn phiên bản đầy đủ 39 hệ số (gồm cả Δ và ΔΔ) là MFCC Feature.

2.2. Các mô hình phân loại âm thanh nhạc cụ

2.2.1. Mạng nơ-ron tích chập

Mạng neural tích chập (CNN) sử dụng dữ liệu ba chiều để phân loại hình ảnh và nhận dạng đối tượng. Là một phần của học máy và trung tâm của các thuật toán học sâu (Deep Learning), mạng neural gồm các lớp nút: đầu vào, ẩn và đầu ra. Các nút kết nối với nhau, có trọng số (weight) và ngưỡng (threshold). Nếu đầu ra vượt ngưỡng, dữ liệu được truyền đến lớp tiếp theo. Các mạng neural khác nhau được sử dụng cho các loại dữ liệu khác nhau: mạng hồi quy cho ngôn ngữ tự nhiên và nhận dạng giọng nói, CNNs cho phân loại và thị giác máy tính. Trước CNN, việc nhận diện đối tượng đòi hỏi trích xuất đặc trưng thủ công. CNN hiện cung cấp phương pháp mở rộng hơn, sử dụng đại số tuyến tính, nhưng đòi hỏi nhiều tính toán và GPU để huấn luyện mô hình.

2.2.2. Mạng MobieNetV2

Hiện nay, nhiều mô hình Deep Learning được thiết kế với hàng triệu tham số và cấu trúc rất phức tạp để đạt độ chính xác cao, nhờ sự phát triển nhanh của phần cứng như GPU và TPU. Tuy nhiên, các thiết bị này rất đắt đỏ và không phải là lựa chọn khả thi cho mọi nghiên cứu. Khi triển khai mô hình trên các thiết bị phần cứng hạn chế (edge devices) như vi điều khiển hoặc máy tính nhúng, ta cần các mô hình nhỏ gọn, nhanh và tiết kiệm tài nguyên. MobileNet là một trong số đó.

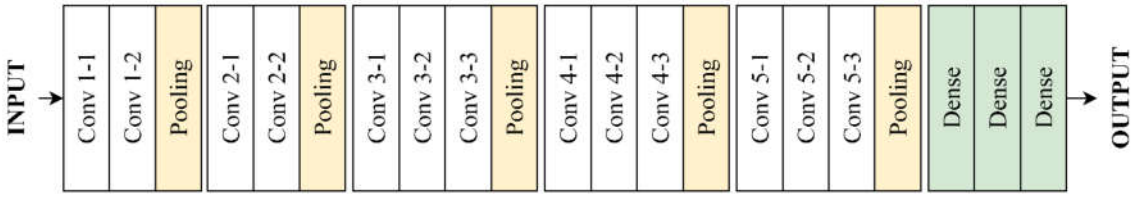
MobileNetV2 là một kiến trúc CNN được thiết kế để đạt hiệu quả cao trong các tác vụ thị giác máy tính, đồng thời tối ưu hóa cho thiết bị di động và nhúng. Được Google giới thiệu năm 2018, MobileNetV2 là bản nâng cấp từ MobileNetV1, với hai cải tiến chính: Linear bottlenecks và Inverted Residual Blocks.

MobileNetV2 đã chứng tỏ là một giải pháp mạnh mẽ trong việc phân loại âm thanh, nhờ vào thiết kế kiến trúc linh hoạt và khả năng tối ưu hóa tài nguyên. Mặc dù ban đầu được phát triển để giải quyết các bài toán thị giác máy tính như phân loại hình ảnh, MobileNetV2 có thể dễ dàng thích ứng để xử lý dữ liệu âm thanh, đặc biệt là khi âm thanh được chuyển đổi thành các biểu diễn hình ảnh như Spectrograms hoặc các dạng phổ khác. Những biểu diễn này chuyển đổi tín hiệu âm thanh thành dạng trực quan, cho phép MobileNetV2 khai thác những khả năng mạnh mẽ của mình trong việc nhận diện và phân loại các đặc trưng phức tạp.

2.2.3. Mô hình VGG16

VGG16 là một cấu trúc mạng nơ-ron tích chập (CNN) được đề xuất bởi Visual Geometry Group (VGG) tại Đại học Oxford. Mô hình này nổi bật với độ sâu bao gồm 16 lớp, trong đó có 13 lớp tích chập và 3 lớp kết nối đầy đủ. VGG16 được biết đến với sự đơn giản

và hiệu quả, cũng như khả năng đạt hiệu suất mạnh mẽ trong nhiều nhiệm vụ thị giác máy tính, bao gồm phân loại hình ảnh và nhận diện đối tượng. Cấu trúc của mô hình bao gồm một chuỗi các lớp tích chập tiếp theo là các lớp max-pooling với độ sâu ngày càng tăng cho phép mô hình học được các đặc trưng hình ảnh theo cấp bậc phức tạp, dẫn đến dự đoán chính xác và đáng tin cậy. Dù có sự đơn giản so với các cấu trúc hiện đại hơn, VGG16 vẫn là sự lựa chọn phổ biến cho nhiều ứng dụng học sâu nhờ vào tính linh hoạt và hiệu suất xuất sắc của nó.



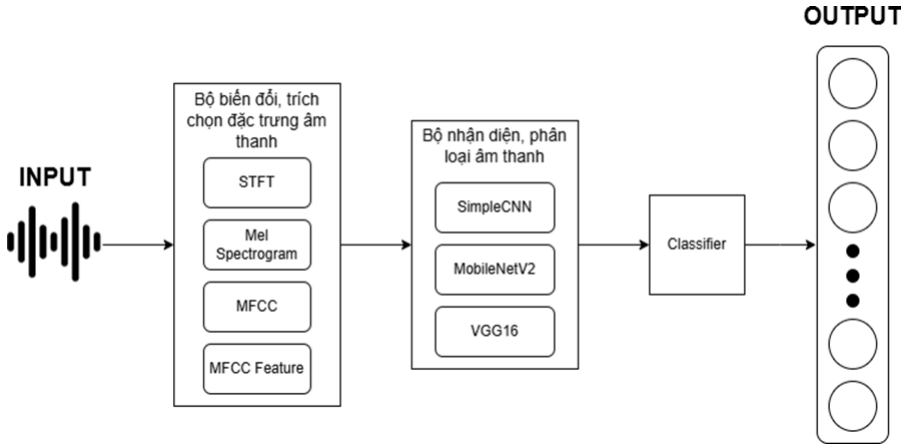
Hình 4: Cấu trúc VGG16

VGG16 nổi bật với sự đơn giản và cấu trúc đồng nhất, giúp dễ dàng hiểu và triển khai. Cấu hình của VGG16 thường bao gồm 16 lớp, trong đó có 13 lớp convolutional và 3 lớp fully connected. Các lớp này được tổ chức thành các khối, mỗi khối chứa nhiều lớp tích chập tiếp theo là một lớp max-pooling để giảm kích thước.

III. Phương pháp nghiên cứu

3.1. Mô hình đề xuất

Nghiên cứu này sẽ thực hiện phân loại âm thanh các nhạc cụ bằng 4 phương pháp trích chọn đặc trưng trên 3 mô hình CNN khác nhau. Từ đó, có thể đánh giá một cách khách quan nhất vai trò, ảnh hưởng của 4 phương pháp biến đổi và trích chọn đặc trưng này. Đầu vào của mô hình là các file âm thanh thô được thu từ nhiều loại nhạc cụ khác nhau. Để xử lý và phân loại các âm thanh này, đầu tiên, các file âm thanh sẽ được áp dụng bốn phương pháp trích xuất đặc trưng âm thanh khác nhau: STFT, Mel Spectrogram, MFCC và MFCC Feature. Mỗi phương pháp trích xuất đặc trưng này sẽ chuyển đổi dữ liệu âm thanh thô thành bốn tập dữ liệu hình ảnh riêng biệt, mỗi hình ảnh trong tập đều có kích thước , tương ứng với các đặc trưng của bốn phép trích chọn đặc trưng.



Hình 5: Mô hình đề xuất

Sau khi trích xuất đặc trưng hoàn tất, bốn tập dữ liệu hình ảnh tương ứng với các phương pháp sẽ được đưa vào ba mô hình CNN khác nhau: SimpleCNN, MobileNetV2 và VGG16. Mỗi mô hình sẽ xử lý đầu vào để phân loại âm thanh nhạc cụ tương ứng. Việc kết hợp bốn phương pháp trích xuất đặc trưng với ba kiến trúc CNN cho phép so sánh hiệu suất từng phương pháp trên từng mô hình. Kết quả thu được sẽ cho biết âm thanh được phân loại vào đúng nhóm nhạc cụ nào dựa trên đặc trưng học được. Mục tiêu là tìm ra sự kết hợp tối ưu giữa kỹ thuật trích đặc trưng và mô hình phân loại nhằm nâng cao độ chính xác và hiệu suất toàn hệ thống. Mô hình đầu tiên, SimpleCNN, sử dụng các lớp tích chập cơ bản để làm chuẩn tham chiếu. Tiếp theo là MobileNetV2, được thiết kế tối ưu cho thiết bị tài nguyên hạn chế với kiến trúc nhẹ, tốc độ nhanh, nhưng có thể kém hơn về độ chính xác so với các mô hình lớn. Cuối cùng là VGG16, mô hình lớn với nhiều tham số, đòi hỏi tài nguyên tính toán cao và thời gian huấn luyện lâu, nhưng mang lại hiệu quả cao.

3.2. Tập dữ liệu

Bảng 1: Chi tiết bộ dữ liệu MUSIC Solo

Nhạc cụ	Tập huấn luyện	Tập kiểm tra	Tổng
Accordion	603	151	754
Acoustic Guitar	733	183	916
Cello	664	166	830
Clarinet	274	68	342
Erhu	436	109	545
Flute	357	89	446
Saxophone	209	52	261
Trumpet	179	45	224
Tuba	331	83	414
Violin	489	123	612
Xylophone	403	101	504
Tổng 11 nhạc cụ	**4678**	**1170**	**5848**

Mục tiêu của nghiên cứu này là phát triển, cải thiện hiệu suất của bài toán phân tách nguồn âm thanh, vì vậy tác giả sử dụng bộ dữ liệu MUSIC Solo (Montesinos et al., 2020) đang được sử dụng phổ biến trong bài toán phân tách nguồn âm thanh để thực hiện thử nghiệm. Bộ dữ liệu này bao gồm các đoạn âm thanh được thu âm từ các nhạc cụ chơi đơn (solo), tức là mỗi đoạn âm thanh chỉ có một nhạc cụ duy nhất. Các tệp âm thanh trong bộ data có định dạng (.wav) với tốc độ lấy mẫu là 11025Hz. Bộ dữ liệu có 5848 file âm thanh solo của 11 loại nhạc cụ bao gồm: Accordion, Acoustic Guitar, Cello, Clarinet, Erhu, Flute, Saxophone, Trumpet, Tuba, Violin, Xylophone. Các tệp âm thanh này được chia theo tỷ lệ 80:20 thành các tập huấn luyện và kiểm tra tương ứng.

3.3. Phương pháp đánh giá

3.3.1. Độ chính xác từng lớp (Precision)

Precision *đo lường tỷ lệ các mẫu được dự đoán thuộc một lớp cụ thể là chính xác (tức là thực sự thuộc lớp đó).*

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

Trong đó:

• **TP** *(True Positive): Số mẫu thuộc lớp đó và được dự đoán đúng.*

• **FP** *(False Positive): Số mẫu không thuộc lớp đó nhưng bị dự đoán nhầm vào lớp đó.*

3.3.2. Độ nhạy (Recall)

Recall *đo lường khả năng phát hiện đúng các mẫu thực sự thuộc về một lớp cụ thể.*

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

Trong đó:

• **FN** *(False Negative): Số mẫu thuộc lớp đó nhưng bị dự đoán nhầm sang lớp khác.*

3.3.3. F1-Score

F1-Score *là trung bình điều hòa giữa Precision và Recall, giúp cân bằng giữa hai chỉ số này, đặc biệt hữu ích khi dữ liệu không cân bằng.*

$$\text{F1 - Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

IV. Kết quả thực nghiệm

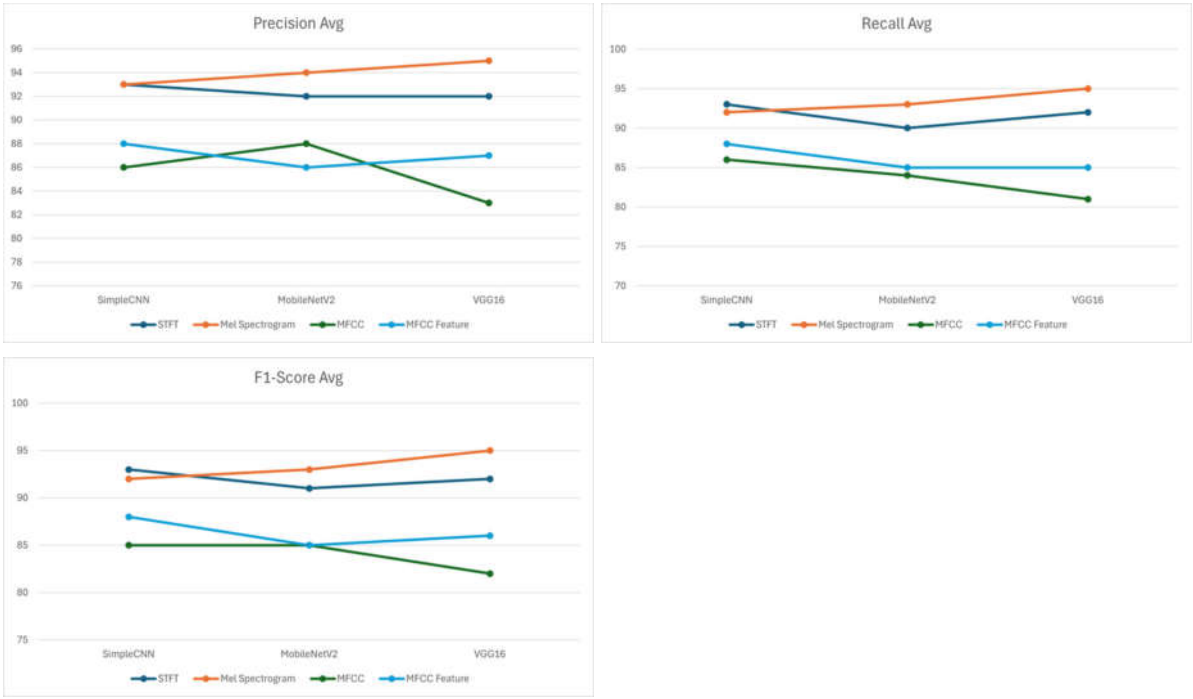
4.1. Kết quả so sánh theo các phương pháp biến đổi và trích chọn đặc trưng

Nhìn vào Bảng 2, có thể nhận thấy rằng đối với bài toán phân loại âm thanh âm nhạc, cụ thể là phân loại âm thanh nhạc cụ sử dụng mô hình CNN thì các phương pháp phân tích Spectrogram cho kết quả tốt hơn các phương pháp phân tích hệ số Cepstral. Các chỉ số hiệu suất của phương pháp phân tích Spectrogram đạt từ 0.904 – 0.948 với cả 3 mô hình, trong khi các phương pháp phân tích hệ số Cepstral chỉ đạt từ 0.815 – 0.880. Giả thuyết đưa ra có thể do các phương pháp phân tích Spectrogram bảo tồn thông tin về cả tần số và thời gian, điều này rất hữu ích khi sử dụng các mô hình học sâu như CNN. Trong khi đó các phương pháp phân tích hệ số Cepstral không bảo tồn thông tin thời gian tốt như Spectrogram, nên khó để nắm bắt những thay đổi nhỏ trong tín hiệu. Ngoài ra các phương pháp này còn có thể bỏ qua một số chi tiết quan trọng trong tín hiệu âm thanh so với Spectrogram. Với phương pháp phân tích hệ số Cepstral, phương pháp MFCC tính Δ và ΔΔ để có thêm các đặc trưng về sự khác biệt giữa hai khung liên tiếp cho kết quả tốt hơn MFCC với 13 hệ số thông thường. Với phương pháp phân tích Spectrogram, phương pháp Mel Spectrogram cũng cho kết quả tốt hơn so với STFT thông thường nhờ bộ lọc tần số Mel. Như vậy, trong 4 phương pháp phân tích và trích chọn đặc trưng ở trong nghiên cứu *này, phương pháp Mel Spectrogram cho kết quả tốt nhất.*

Bảng 2: Kết quả so sánh theo các phương pháp biến đổi và trích chọn đặc trưng

	CNN			MobieNetV2			VGG16		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
STFT	0.925	0.916	0.924	0.919	0.904	0.909	0.917	0.915	0.916
Mel Spectrogram	0.928	0.929	0.927	0.935	0.927	0.933	0.946	0.951	0.948
MFCC	0.858	0.858	0.855	0.878	0.845	0.855	0.833	0.815	0.815
MFCC Feature	0.881	0.880	0.880	0.860	0.852	0.855	0.874	0.854	0.860

4.2. Kết quả so sánh theo các mô hình phân loại



Hình 6: Độ chính xác theo lớp, độ nhạy, F1- Score trung bình của các mô hình

Nhìn vào biểu đồ so sánh kết quả các mô hình, có thể nhận thấy rằng:

• Mô hình VGG16 với phương pháp Mel Spectrogram cho kết quả tốt nhất với tất cả phép đo đều đạt giá trị 95%. Tuy nhiên mô hình này tốn khá nhiều tài nguyên, thời gian huấn luyện khá dài và lâu nhất trong cả 3 mô hình, đồng thời với phương pháp trích xuất có ít đặc trưng được thể hiện hơn như MFCC thì mô hình lại cho kết quả thấp nhất.

• Mô hình MobileNetV2 có ưu điểm nhẹ hơn, có thời gian huấn luyện nhanh nhất trong cả 3 mô hình, tốn ít tài nguyên nguyên hơn, nhưng kết quả lại thấp nhất đối với phương pháp STFT mà MFCC Feature.

• Mô hình SimpleCNN có thời gian huấn luyện ở mức trung bình nhưng lại cho kết quả khá tốt với 3 phương pháp STFT, MFCC và MFCC Feature. Tuy nhiên ở phương pháp Spectrogram mô hình này lại cho kết quả thấp nhất.

V. Kết luận

Trong nghiên cứu này này, tác giả đã đề xuất một mô hình cho bài toán phân loại âm thanh nhạc cụ. Qua thử nghiệm với các phương pháp, các mô hình phân loại âm thanh và các phương pháp đánh giá khác nhau, có thể nhận thấy rằng phương pháp Mel Spectrogram với mô hình VGG16 cho kết quả hiệu suất cao nhất lên tới 95%. Với kết quả này có thể thấy rằng Mel Spectrogram là phương pháp vô cùng hiệu trong các kỹ thuật phân loại âm thanh nhạc cụ được thử nghiệm. Kết quả này không chỉ cho thấy khả năng mạnh mẽ của Mel Spectrogram trong việc trích xuất các đặc trưng âm thanh quan trọng, mà còn chứng minh hiệu quả của mô hình VGG16 trong việc phân loại âm thanh với độ chính xác cao. Kết quả này

có thể được xem như là bước đệm quan trọng để triển khai lên mô hình phân tách nguồn âm thanh, mở ra khả năng nghiên cứu sâu hơn về cách tối ưu hóa và áp dụng các phương pháp phân loại âm thanh trong các tình huống và điều kiện khác nhau.

Tài liệu tham khảo:

[1]. Boddapati, V., Petef, A., Rasmusson, J., & Lundberg, L. (2017). Classifying environmental sounds using image recognition networks. *Procedia Computer Science, 112*, 2048–2056. https://doi.org/10.1016/j.procs.2017.08.250

[2]. Bountourakis, V., Vrysis, L., & Papanikolaou, G. (2015). Machine learning algorithms for environmental sound recognition: Towards soundscape semantics. *In Proceedings of the Audio Mostly Conference*.

[3]. Bountourakis, V., Doukakis, D., Stathis, K., & Papanikolaou, G. (2019). An enhanced temporal feature integration method for environmental sound recognition. *Acoustics, 1*(2), 410–422. https://doi.org/10.3390/acoustics1020023

[4]. Choi, K., Fazekas, G., & Sandler, M. (2016). Automatic tagging using deep convolutional neural networks. *arXiv preprint arXiv:1606.00298*.

[5]. Cowling, M., & Sitte, R. (2003). Comparison of techniques for environmental sound recognition. *Pattern Recognition Letters, 24*(15), 2895–2907. https://doi.org/10.1016/S0167-8655(03)00147-8

[6]. Gong, Y., Chung, Y. A., & Glass, J. (2021). Ast: Audio spectrogram transformer. arXiv preprint arXiv:2104.01778.

[7]. Li, D., Sethi, I. K., Dimitrova, N., & McGee, T. (2001). Classification of general audio data for content-based retrieval. *Pattern Recognition Letters, 22*(5), 533–544. https://doi.org/10.1016/S0167-8655(00)00119-7

[8]. Lu, L., Li, S. Z., & Zhang, H.-J. (2001). Content-based audio segmentation using support vector machines. *In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)* (pp. 749–752). https://doi.org/10.1109/ICME.2001.1237830

[9]. Montesinos, J. F., Slizovskaia, O., & Haro, G. (2020). Solos: A dataset for audio-visual music analysis. *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 1–6. https://doi.org/10.1109/MMSP48831.2020.9287124

[10]. Naman, A., & Zhang, G. (2025, April). FAST: Fast Audio Spectrogram Transformer. In ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

[11]. Pons, J., & Serra, X. (2017). Designing efficient architectures for modeling temporal features with convolutional neural networks. *In Proceedings of the IEEE ICASSP* (pp. 2472–2476).

[12]. P. Dhanalakshmi, S. Palanivel, and V. Ramalingam. Classification of audio signals using svm and rbfnn. *In Expert Systems with Applications, vol. 36, no. 3, Part 2, pp. 6069–6075, 2009, ISSN: 0957-4174.*

[13]. Wu, Y., Mao, H., & Yi, Z. (2018). Audio classification using attention-augmented convolutional neural network. *Knowledge-Based Systems, 161*, 90–100. https://doi.org/10.1016/j.knosys.2018.07.033

MUSICAL INSTRUMENT CLASSIFICATION USING SPECTRAL FEATURES AND CONVOLUTIONAL NEURAL NETWORKS

Pham Thi Thu Trang[3], Duong Tan Nghia[4]

Abstract: *In recent years, the problem of sound classification has attracted considerable attention due to its wide range of applications in engineering. Feature extraction plays a crucial role in the classification process. This paper presents a musical instrument sound classification system that combines popular spectral feature extraction methods with deep learning models. Audio signals are transformed into time-frequency representations such as Short-Time Fourier Transform (STFT), Mel-spectrogram, Mel-Frequency Cepstral Coefficients (MFCC), and MFCC with delta coefficients. These features are fed into three convolutional neural network models: SimpleCNN, MobileNetV2, and VGG16. Experiments conducted on the Solo MUSIC dataset show that Mel-spectrogram and MFCC+delta achieve the highest accuracy and F1-score. Among the models, MobileNetV2 strikes a good balance between computational efficiency and performance..*

Keywords: *artificial Intelligence, deep learning, sound classification, feature extraction, convolutional neural Network*

[3] Hanoi Open University

[4] Hanoi University of Science and Technology