

# VieMusic: Vietnamese Music Genres Classification using Mel-spectrogram

Kiều Quý Hùng

Khoa Khoa học và Kỹ thuật thông tin

Trường Đại học Công nghệ Thông tin -

ĐHQG Thành phố Hồ Chí Minh

Thành phố Hồ Chí Minh, Việt Nam

22520505@gm.uit.edu.vn

Nguyễn Duy Liêm

Khoa Khoa học và Kỹ thuật thông tin

Trường Đại học Công nghệ Thông tin -

ĐHQG Thành phố Hồ Chí Minh

Thành phố Hồ Chí Minh, Việt Nam

22520752@gm.uit.edu.vn

Mạc Nguyên Phúc

Khoa Khoa học và Kỹ thuật thông tin

Trường Đại học Công nghệ Thông tin -

ĐHQG Thành phố Hồ Chí Minh

Thành phố Hồ Chí Minh, Việt Nam

22521123@gm.uit.edu.vn

**Tóm tắt nội dung**—Báo cáo này trình bày một hệ thống phân loại hình ảnh Mel-spectrogram được chuyển đổi từ dữ liệu âm thanh của ba thể loại nhạc truyền thống Việt Nam. Hệ thống sử dụng các mô hình Mạng Nơ-ron Tích chập (CNN) để dự đoán thể loại nhạc. Chúng tôi xây dựng ba mô hình CNN với các lớp như Conv2d, max pooling, Dropout, Batch Normalization nhằm đánh giá ảnh hưởng của từng lớp đến hiệu suất mô hình. Đồng thời, chúng tôi thử nghiệm với một số mô hình pre-trained như VGG16, Efficient Net, ConvNext và ResNet50 để so sánh hiệu quả. Ngoài ra, chúng tôi rút trích đặc trưng HOG và huấn luyện mô hình KNN và SVM.

**Index Terms**—Deep Learning, Machine Learning, Vietnamese Music Genres, Classification, Mel-spectrogram, pre-trained model

## I. GIỚI THIỆU

Trong những năm gần đây, các mô hình học máy dựa trên hình ảnh đã chứng minh sức mạnh vượt trội trong việc giải quyết các bài toán thị giác máy tính, đặc biệt là nhận dạng vật thể, phân đoạn hình ảnh và các ứng dụng thực tế phức tạp. Khả năng biểu diễn thông tin một cách trực quan và tự động học các đặc trưng phức tạp đã giúp các mô hình này đạt được độ chính xác cao hơn so với các phương pháp truyền thống như HMM. Đặc biệt, mạng nơ-ron tích chập (CNN) đã thể hiện khả năng bắt giữ các mối quan hệ không gian và đặc trưng cấp cao của dữ liệu hình ảnh. Ngược lại, trong xử lý âm thanh, các phương pháp trích xuất đặc trưng thủ công thường đòi hỏi kiến thức chuyên sâu, phụ thuộc vào kinh nghiệm của người thiết kế, và gặp khó khăn trong việc mở rộng ứng dụng cho nhiều bài toán khác nhau. Tuy nhiên, việc chuyển đổi tín hiệu âm thanh thành dạng hình ảnh, như Spectrogram hoặc Mel-spectrogram, đã mở ra một hướng đi mới. Spectrogram trực quan hóa các đặc trưng tần số và thời gian của tín hiệu âm thanh, cho phép các mô hình xử lý hình ảnh tiên tiến tự động học các đặc trưng phức tạp mà không cần sự can thiệp của con người. Báo cáo này đề xuất một hệ thống phân loại thông minh, kết hợp giữa xử lý tín hiệu âm thanh và các kỹ thuật học sâu. Đầu tiên, tín hiệu âm thanh được chuyển đổi thành Spectrogram thông qua Short-Time Fourier Transform (STFT) và Mel-spectrogram, sau đó lưu dưới dạng ảnh. Các ảnh này được đưa vào các mô hình học sâu pre-trained và hai mô hình CNN tự xây dựng để phân loại. Ngoài ra, hệ thống cũng áp dụng trích xuất đặc trưng HOG từ ảnh và sử dụng

các mô hình học máy như KNN và SVM để đánh giá hiệu quả phân loại. Đặc trưng HOG nổi bật nhờ khả năng mô tả hình dạng và cấu trúc của đối tượng thông qua phân bố hướng gradient, giúp tăng độ chính xác trong các bài toán nhận diện và phân loại. Kết quả của các phương pháp này được so sánh nhằm tìm ra mô hình tối ưu nhất, đồng thời tích hợp hệ thống vào một ứng dụng web để nhận diện âm thanh, mở rộng tiềm năng ứng dụng trong thực tế.

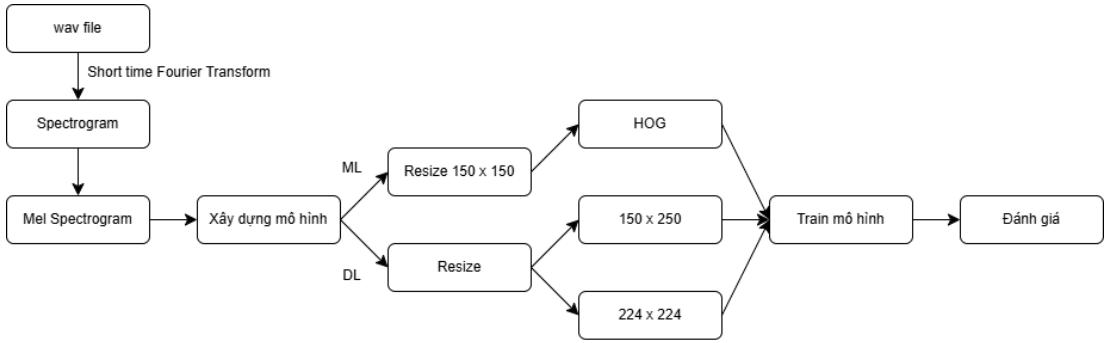
## II. THIẾT LẬP VẤN ĐỀ

Bài toán này nhằm xây dựng một hệ thống phân loại tự động các đoạn nhạc truyền thống Việt Nam (ca trù, chèo, cải lương) với độ dài 30 giây. Hệ thống sẽ nhận đầu vào là một file âm thanh và trả về kết quả là thể loại nhạc mà đoạn âm thanh đó thuộc về. Để giải quyết bài toán này, chúng ta sẽ sử dụng kỹ thuật học máy có giám sát. Chúng tôi sử dụng với bộ dữ liệu Vietnam Traditional Music (5 genres) trên Kaggle, được thu thập từ hơn 4,1 giờ ghi âm cho mỗi thể loại, bao gồm 2499 tệp âm thanh thuộc năm thể loại: cải lương, ca trù, chầu văn, chèo và hát xẩm. Mỗi thể loại bao gồm dạng tệp '.wav' với độ dài 30 giây. Tuy nhiên, chúng tôi chỉ sử dụng 3 thể loại là cải lương, ca trù và chèo. Thể loại ca trù gồm 499 tệp, 2 thể loại còn lại gồm 500 tệp. Với bộ dữ liệu này, nhiệm vụ chính của chúng tôi là phân loại các bản ghi khác nhau vào các nhãn được xác định trước như đã nêu. Để đánh giá hiệu suất và so sánh mô hình cho 3 thể loại khác nhau, chúng tôi sử dụng F1-score với phương pháp macro-average.

## III. QUY TRÌNH THỰC HIỆN

Quy trình xử lý và xây dựng mô hình trong đồ án được thể hiện qua pipeline ở Hình 1, bao gồm các bước chính như xử lý tín hiệu âm thanh, trích xuất đặc trưng, xây dựng mô hình học máy (ML) và học sâu (DL), và cuối cùng là đánh giá hiệu suất mô hình.

Đầu tiên, các tệp âm thanh định dạng ".wav" được tải lên với tần số lấy mẫu gốc là 22050 Hz. Tín hiệu sau đó được chuyển đổi sang miền tần số bằng cách áp dụng Short-Time Fourier Transform (STFT) với số tín hiệu cho mỗi cửa sổ là 512 và bước nhảy là 1024. Với các tham số này, mỗi lần thực hiện STFT, chỉ 512 mẫu được tính toán, trong khi 512 mẫu tiếp theo bị bỏ qua. Quá trình này lặp lại cho đến khi toàn



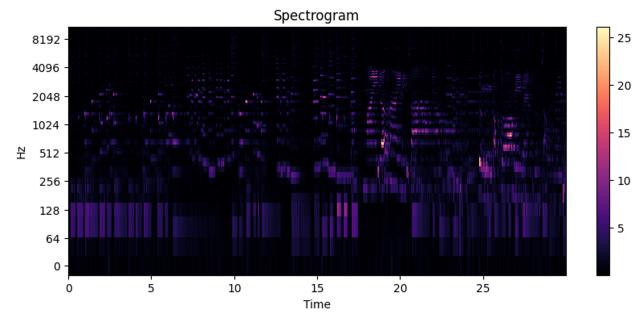
Hình 1. Pipeline

bộ tín hiệu âm thanh được quét. Kết quả thu được là hình ảnh dạng Spectrogram, biểu diễn trực quan của tín hiệu âm thanh dưới dạng tần số theo thời gian. Trong Spectrogram, trục hoành thể hiện thời gian, trục tung biểu diễn tần số, và biên độ tín hiệu được thể hiện qua màu sắc (Hình 2).

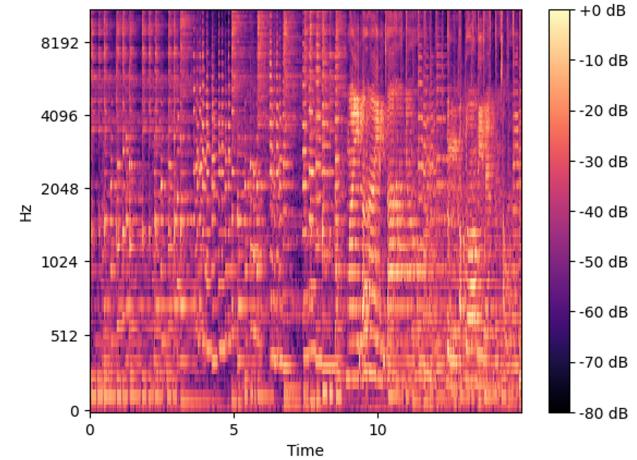
Tiếp theo, từ phổ năng lượng của tín hiệu, chúng tôi tính toán Mel-Spectrogram bằng cách áp dụng bộ lọc Mel với 128 dải tần và cận trên của tần số (fmax) là 8000 Hz. Mel-Spectrogram (Hình ??) là một biểu diễn phổ tần số dựa trên thang đo Mel, mô phỏng cách tai người cảm nhận âm thanh. Khác với Spectrogram truyền thống, Mel-Spectrogram chia tần số thành các dải tương ứng với những vùng tần số mà tai người nhạy cảm nhất, giúp trích xuất các đặc trưng âm thanh ý nghĩa hơn. Số lượng dải tần càng cao thì độ phân giải càng lớn, nhưng nếu quá cao sẽ làm tăng kích thước hình ảnh mà không cải thiện đáng kể hiệu quả trích xuất đặc trưng. Chúng tôi sử dụng 128 dải tần để cân bằng giữa độ phân giải và giới hạn tài nguyên. Tần số trên mức 8000 Hz bị loại bỏ do giới hạn này, và phổ biến độ được chuyển sang phổ năng lượng để phù hợp với thang Mel. Cuối cùng, kết quả được chuyển đổi sang thang đo decibel để tăng khả năng biểu diễn.

Mel-Spectrograms được lưu dưới dạng ảnh PNG sử dụng thư viện Matplotlib, với thang màu magma. Ảnh được lưu ở kích thước 369 x 496 pixel với định dạng RGBA (RGB cùng chiều alpha biểu thị độ trong suốt). Sau đó, các đặc trưng HOG (Histogram of Oriented Gradients) được trích xuất từ các hình ảnh này để sử dụng cho các mô hình học máy. Gradient của hình ảnh được chia thành 9 hướng, mỗi hướng đại diện cho 20 độ. Mỗi cell gồm 8x8 pixel, mỗi block có kích thước 2x2 cell. Các vector đặc trưng trong block được chuẩn hóa theo phương pháp L2-Hys, giúp điều chỉnh năng lượng vector về mức 1, giảm ảnh hưởng của thay đổi độ sáng hoặc độ tương phản. Phương pháp Hys giới hạn giá trị histogram trong một ngưỡng cố định để tăng tính ổn định và giảm nhiễu.

Với hướng tiếp cận deep learning, chúng tôi đề xuất một số mô hình đã được huấn luyện trước đó cùng với hai mô hình Mạng Nơ-ron Tích chập (Convolutional Neural Network - CNN) tự xây dựng, với kiến trúc được trình bày trong bảng dưới đây cùng với các pre-trained model trong CNN như EfficientNet, VGG16, .... Cụ thể, mô hình CNN mà chúng tôi tự xây dựng về cơ bản bao gồm các lớp tích chập (Convolution), hàm kích hoạt Rectified Linear Unit (ReLU), Max pooling,



Hình 2. Spectrogram



Hình 3. Mel Spectrogram

Flatten, Dropout, Fully Connected (FC), và Softmax. Hàm Softmax được sử dụng tại lớp FC cuối cùng để dự đoán xác suất thuộc về các danh mục được phân loại.

Chúng tôi cũng áp dụng kỹ thuật transfer learning để xây dựng mô hình phân loại hình ảnh. Bằng cách sử dụng các lớp convolutional đã được huấn luyện trước của VGG16, mô hình mới chỉ cần học các đặc trưng đặc thù của tập dữ liệu mới, đơn giản và dễ triển khai, có khả năng trích xuất đặc trưng tốt từ ảnh Mel-spectrogram.

Để nâng cao khả năng của mô hình phân loại hình ảnh, kiến

Bảng I  
MODEL 1 ARCHITECTURE

Layers	Output
Conv[5x5] @ 32 - ReLU - MP [2x2]	(110, 110, 32)
Conv[3x3] @ 32 - ReLU - MP [2x2]	(54, 54, 32)
Conv[3x3] @ 64 - ReLU - MP [2x2]	(26, 26, 64)
Flatten-Dr(0.2)	43264
FC	128
FC	64
FC-Softmax	3

Bảng II  
MODEL 2 ARCHITECTURE

Layers	Output
Conv[3x3] @ 16 - ReLU - MP [2x2]	(99, 249, 16)
Conv[3x3] @ 32 - ReLU - MP [2x2]	(48, 123, 32)
Conv[3x3] @ 64 - ReLU - MP [2x2]	(23, 60, 64)
Flatten-Dr(0.2)	88320
FC - Dr(0.2) - BN	256
FC - Dr(0.2) - BN	128
FC-Softmax	3

trúc pre-trained EfficientNet. EfficientNetB0 được biết đến với khả năng cân bằng giữa hiệu suất và chi phí tính toán, phù hợp với các bài toán có tài nguyên hạn chế nhưng vẫn đảm bảo độ chính xác cao. Tương tự như VGG16, các lớp convolutional được huấn luyện trước của EfficientNet được tận dụng để trích xuất đặc trưng từ ảnh Mel-spectrogram. Các lớp này đặc biệt hiệu quả trong việc khai thác các đặc trưng phong phú và phân cấp từ ảnh spectrogram nhờ thiết kế tối ưu của mô hình.

Bên cạnh các kiến trúc VGG16 và EfficientNet, mô hình ConVNeXt cũng được sử dụng để khai thác các đặc trưng sâu từ ảnh Mel-spectrogram. ConvNeXt là một kiến trúc hiện đại được phát triển dựa trên CNN truyền thống, nhưng được cải tiến về mặt thiết kế để phù hợp với các tiêu chuẩn hiện đại như khả năng tổng quát hóa và tối ưu hiệu suất. Nhờ vào thiết kế này, ConVNeXt có khả năng học các đặc trưng tốt hơn từ ảnh Mel-spectrogram, góp phần cải thiện độ chính xác trong bài toán phân loại thể loại âm nhạc.

Pre-trained model ResNet (Residual Network) khác biệt so với các mô hình pre-trained khác như VGG16, EfficientNet và ConVNeXt chủ yếu ở việc sử dụng các khối residual (skip connections) giúp giảm thiểu vấn đề gradient vanishing trong các mạng sâu, cho phép mô hình học hiệu quả hơn từ dữ liệu phức tạp như Mel-spectrogram. Lớp Global Average Pooling không chỉ giảm số lượng tham số mà còn giữ lại những đặc trưng quan trọng nhất, trong khi các lớp fully connected giúp mô hình học các quan hệ đặc trưng chi tiết hơn.

Trong quá trình thực nghiệm, các mô hình học sâu đều được biên dịch với các siêu tham số được tối ưu nhằm đảm bảo hiệu quả huấn luyện. Cụ thể, mô hình sử dụng bộ tối ưu hóa Adam và hàm mất mát Categorical Cross-Entropy, với độ đo hiệu năng được lựa chọn là accuracy. Để đảm bảo khả năng dừng huấn luyện sớm khi mô hình đạt hiệu suất tốt nhất, callback EarlyStopping đã được tích hợp. Các siêu tham số cho callback bao gồm: theo dõi giá trị loss để đánh giá hiệu suất, thiết lập patience là 5, nghĩa là nếu giá trị mất mát không cải thiện sau 5 epoch liên tiếp thì quá trình huấn luyện sẽ tự động dừng

Bảng III  
PRE-TRAINED MODEL VGG16 ARCHITECTURE

Layers	Output
VGG16	(4, 7,512)
GlobalAveragePooling2D	(512)
FC - Dr(0.3)	128
FC-Softmax	3

Bảng IV  
PRE-TRAINED MODEL EFFICIENTNETB0 ARCHITECTURE

Layers	Output
EfficientNetB0	(5, 8, 1280)
GlobalAveragePooling2D	(1280)
FC - Dr(0.3)	128
FC-Softmax	3

Bảng V  
PRE-TRAINED MODEL CONVNEXT-TINY ARCHITECTURE

Layers	Output
ConvNeXt-Tiny	(4, 7, 768)
GlobalAveragePooling2D	(768)
FC - Dr(0.3)	128
FC-Softmax	3

Bảng VI  
PRE-TRAINED MODEL RESNET50 ARCHITECTURE

Layers	Output
ResNet50	(5, 8, 2048)
GlobalAveragePooling2D	(2048)
FC - Dr(0.3)	128
FC-Softmax	3

(patience = 5). Đồng thời, tham số restore\_best\_weights được đặt là True nhằm khôi phục các trọng số tốt nhất đạt được trong suốt quá trình huấn luyện. Với các thiết lập này, các mô hình học sâu sẽ được kỳ vọng đạt hiệu quả tối ưu và tránh hiện tượng overfitting. Ngoài ra, chúng tôi huấn luyện các mô hình với số lượng epoch là 20, batch\_size là 32. Tập dữ liệu được chia theo tỉ lệ 8:1:1 lần lượt cho tập train, tập dev và tập test. Để tránh sự mất cân bằng dữ liệu giữa các thể loại cải lương, ca trù, chèo khi chia các tập, chúng tôi sử dụng stratified sampling.

Đối với các phương pháp học máy cơ bản, chúng tôi đã sử dụng hai mô hình: Support Vector Machine (SVM) và K-Nearest Neighbors (KNN) vì với bộ dữ liệu 3 nhãn với 1499 mẫu, các mô hình máy học truyền thống như thường hoạt động tốt, đặc biệt khi đặc trưng đầu vào được xử lý hiệu quả. SVM là một kỹ thuật phân loại mạnh mẽ, hoạt động bằng cách tìm kiếm siêu phẳng tốt nhất để phân tách dữ liệu thành các nhóm khác nhau. Nó đặc biệt hiệu quả trong các không gian đa chiều và sử dụng kỹ thuật kernel để xử lý các biên giới phi tuyến tính. Trong khi đó, KNN là một thuật toán đơn giản, phân loại các mẫu mới dựa trên sự bình chọn ở k lảng giềng gần nhất trong không gian đặc trưng. Việc tìm kiếm siêu tham số cho các mô hình này đã được thực hiện bằng GridSearchCV. Đối với mô hình SVM, bộ siêu tham số tối ưu được xác định là sử dụng linear kernel và tham số C có giá trị 0.1. Trong khi đó, đối với mô hình KNN, cấu hình tối ưu bao gồm việc chọn số

láng giềng k bằng 3, sử dụng độ đo khoảng cách Manhattan, và áp dụng uniform weights. Chúng tôi chia tập train và test cho phương pháp này theo tỉ lệ 8:2.

#### IV. THỰC NGHIỆM

##### A. Evaluation Metrics

1) *F1 score*: Trung bình điều hòa của một độ chính xác (precision) và độ nhạy (recall). Nó cung cấp sự cân bằng giữa độ chính xác và độ nhạy. F1 score chỉ cao khi cả độ nhạy và độ chính xác đều cao.

2) *Confusion Matrix*: Ma trận nhầm lẫn là một bảng dùng để mô tả hiệu suất của mô hình phân loại. Nó hiển thị số lượng các dự đoán đúng dương tính, đúng âm tính, sai dương tính và sai âm tính, cung cấp cái nhìn sâu sắc về các loại lỗi mà mô hình đang mắc phải.

##### B. Experimental Setup

Trong nghiên cứu này, chúng tôi sử dụng môi trường Google Colab để thực hiện quá trình huấn luyện và đánh giá các mô hình. Google Colab cung cấp tài nguyên tính toán mạnh mẽ, đặc biệt là GPU và TPU, miễn phí, giúp tăng tốc độ xử lý và giảm thời gian huấn luyện các mô hình học sâu. Môi trường này hỗ trợ Python cùng với các thư viện quan trọng như TensorFlow, PyTorch, scikit-learn và NumPy, đáp ứng tốt các yêu cầu cho việc xây dựng và thử nghiệm mô hình. Ngoài ra, Google Colab còn hỗ trợ tích hợp với Google Drive, giúp dễ dàng lưu trữ và truy cập dữ liệu, cũng như chia sẻ kết quả và mã nguồn. Tính năng notebook tương tác của Colab cho phép theo dõi và kiểm tra kết quả trong thời gian thực, tạo điều kiện thuận lợi cho việc tinh chỉnh siêu tham số và tối ưu hóa mô hình.

#### V. KẾT QUẢ

Bảng VII

COMPARISON OF MODELS ON DATASETS RESIZED TO DIFFERENT DIMENSIONS

Dataset resized to (150, 250)			
Model	F1-score	Precision	Recall
Model0	0.9867	0.9867	0.9867
Model1	0.9403	0.9428	0.94
ConvNeXt	0.8743	0.8816	0.8733
EfficientNet	0.1667	0.1111	0.3333
ResNet50	0.437	0.369	0.54
VGG16	0.9067	0.9069	0.9067

Dataset resized to (224, 224)			
Model	F1-score	Precision	Recall
Model0	0.8473	0.8784	0.8467
Model1	0.5409	0.4951	0.6467
ConvNeXt	0.4701	0.5243	0.5067
EfficientNet	0.1667	0.1111	0.3333
ResNet50	0.1667	0.1111	0.3333
VGG16	0.1667	0.1111	0.3333

Bảng VII thể hiện kết quả của các mô hình trên tập test với 2 kích thước khác nhau. Trên bộ dữ liệu có kích thước (150,250), Model0 đạt hiệu suất vượt trội với F1-score đạt 98.67%, Precision và Recall đều rất cao (98.67%). ConvNeXt và VGG16 cũng đạt kết quả tốt với F1-score lần lượt là 87.43% và 90.67%, cho thấy khả năng phân loại ổn định. Model1 có hiệu suất tương đối tốt với F1-score 94.03%, vượt qua

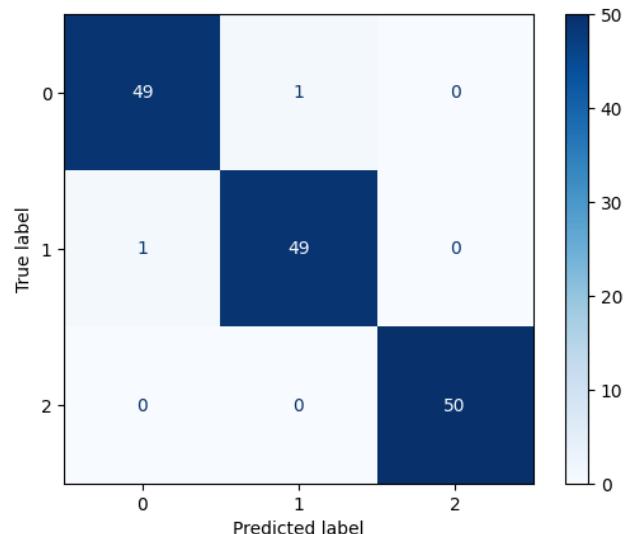
ConvNeXt. Tuy nhiên, EfficientNet và ResNet50 thể hiện hiệu suất rất thấp, với F1-score lần lượt chỉ đạt 16.67% và 43.70%, điều này cho thấy hai mô hình này không phù hợp với kích thước đầu vào này. Và khi áp dụng kích thước (224,224), hiệu suất của các mô hình nhìn chung đều giảm. Model0 vẫn giữ vị trí dẫn đầu với F1-score đạt 84.73%, nhưng đã giảm đáng kể so với khi sử dụng kích thước (150,250). ConvNeXt và Model1 thể hiện tính ổn định hơn với F1-score lần lượt là 47.01% và 54.09%. Trong khi đó, VGG16, EfficientNet, và ResNet50 có hiệu suất giảm mạnh, với F1-score của các mô hình này dao động từ 11.11% đến 16.67%, cho thấy sự suy giảm hiệu quả đáng kể khi kích thước dữ liệu thay đổi. Nhìn chung, Model0 và Model1 là hai mô hình hoạt động tốt nhất trên cả hai kích thước dữ liệu, trong khi EfficientNet và ResNet50 không phù hợp với cả hai loại kích thước đầu vào.

Bảng VIII  
RESULTS OF MACHINE LEARNING MODELS

Model	F1_score	Precision	Recall
SVM	0.88	0.88	0.88
KNN	0.81	0.83	0.82

SVM đạt F1-score 88%, thể hiện hiệu suất ổn định với sự cân bằng tốt giữa Precision và Recall. KNN đạt F1-score 81%, cho thấy khả năng phân loại ở mức khá, nhưng kém hiệu quả hơn so với SVM. Nhìn chung, SVM vượt trội và phù hợp hơn với bài toán phân loại Mel-spectrogram, dù cả hai mô hình đều đơn giản và hiệu suất thấp hơn các mô hình học sâu tốt nhất như Model0 và Model1.

Confusion matrix dưới đây thể hiện hiệu suất của mô hình có kết quả cao nhất trong việc phân loại các thể loại nhạc cải lương, ca trù và chèo. Có thể thấy, việc nhầm lẫn giữa 2 thể loại cải lương và ca trù là do ở một số tệp, sự khác biệt giữa 2 thể loại nhạc này khó có thể nhận thấy.



Hình 4. Ma trận nhầm lẫn của mô hình tốt nhất (Model1 với kích thước ảnh 150 x 250)

## VI. ƯU ĐIỂM VÀ HẠN CHẾ

### A. Ưu điểm

Phân loại thể loại nhạc dựa trên mel-spectrogram được chúng tôi thực hiện bằng hai hướng tiếp cận: học sâu và học máy truyền thông, mỗi hướng đều có những ưu điểm riêng. Với học sâu, mạng CNN tự động trích xuất và học các đặc trưng phức tạp từ mel-spectrogram, cho phép nhận diện tốt các đặc điểm âm học riêng biệt của từng thể loại, đồng thời đạt độ chính xác cao. Trong khi đó, học máy truyền thông sử dụng các đặc trưng HOG (Histogram of Oriented Gradients) được trích xuất từ mel-spectrogram, sau đó áp dụng các mô hình SVM và KNN. Hướng này đơn giản hơn và phù hợp khi dữ liệu hạn chế, đồng thời cho phép kiểm soát tốt hơn các đặc trưng đầu vào. Sự khác biệt này tạo nên tính linh hoạt trong lựa chọn phương pháp, tùy thuộc vào yêu cầu cụ thể về độ phức tạp, hiệu năng, và quy mô dữ liệu.

### B. Hạn chế

Việc sử dụng Google Colab không trả phí đã đặt ra những hạn chế về tài nguyên, cụ thể là dung lượng RAM và GPU. Điều này khiến chúng tôi phải lựa chọn hình ảnh có chất lượng thấp hơn so với tiêu chuẩn mong muốn. Trong nghiên cứu, chúng tôi sử dụng Short-Time Fourier Transform (STFT) với số tín hiệu cho mỗi cửa sổ là 512, số bước nhảy là 1024, tham số n\_mels là 128 và cận trên của tần số (fmax) được đặt ở mức 8000 Hz.

Tuy nhiên, việc số bước nhảy vượt quá số tín hiệu (1024 > 512) đã dẫn đến mất mát một nửa lượng dữ liệu, do mỗi cửa sổ chỉ chia thành một nửa cửa sổ trước. Bên cạnh đó, việc đặt n\_mels là 128 có thể chưa đủ để tạo ra hình ảnh spectrogram chất lượng cao, làm giảm khả năng mô hình trích xuất đặc trưng hiệu quả. Hơn nữa, tham số fmax = 8000 Hz chưa hoàn toàn tuân thủ định lý Nyquist, theo đó tần số lấy mẫu phải ít nhất gấp đôi tần số lớn nhất có trong tín hiệu để tái tạo chính xác. Để phù hợp với tín hiệu lấy mẫu ở 22050 Hz, giá trị fmax hợp lý hơn sẽ là 11025 Hz.

Tuy nhiên, các điều chỉnh như tăng số tín hiệu, giảm số bước nhảy, tăng n\_mels hoặc sử dụng giá trị fmax phù hợp đều làm tăng kích thước dữ liệu, vượt quá dung lượng tính toán cho phép của Google Colab miễn phí.

Ngoài ra, chúng tôi giả định rằng việc lựa chọn kích thước hình ảnh (resize hoặc không resize) hiện tại có thể chưa tối ưu. Việc lưu trữ hình ảnh với kích thước lớn hơn có thể cải thiện độ rõ nét, qua đó tăng cường hiệu quả của các phương pháp trích xuất đặc trưng như HOG hoặc các mô hình học sâu. Tuy nhiên, do hạn chế tài nguyên và chưa thử nghiệm các kích thước resize khác, chúng tôi chưa thể xác định cụ thể mức độ ảnh hưởng của việc thay đổi kích thước đến hiệu suất mô hình.

## VII. CONCLUSION AND FUTURE WORKS

Hệ thống phân loại Mel-spectrogram từ dữ liệu âm thanh ba thể loại nhạc truyền thống Việt Nam đã được thực hiện thông qua hai hướng tiếp cận: học sâu và học máy. Kết quả thí nghiệm cho thấy, trong nhóm học sâu, Model0 và Model1 đạt hiệu suất vượt trội, đặc biệt với kích thước đầu vào 150

x 250, khi F1-score lần lượt đạt 98.67% và 94.03%. Đối với nhóm học máy, SVM và KNN cũng cho thấy hiệu quả đáng khích lệ, với F1-score đạt lần lượt 88% và 81%. Tuy nhiên, hiệu suất của các mô hình giảm đáng kể khi kích thước đầu vào thay đổi sang 224 x 224, nhấn mạnh tầm quan trọng của việc chọn lựa và tiền xử lý dữ liệu phù hợp. Nhìn chung, hệ thống đã cho thấy tiềm năng trong việc phân loại Mel-spectrogram, đồng thời mở ra hướng nghiên cứu tiếp theo, bao gồm việc tinh chỉnh mô hình, mở rộng bộ dữ liệu, áp dụng các kỹ thuật tăng cường dữ liệu và trích xuất các đặc trưng để nâng cao hiệu quả phân loại.

## TÀI LIỆU

- [1] Scaringella, Nicolas, Giorgio Zoia, and Daniel Mlynek. "Automatic genre classification of music content: a survey." IEEE Signal Processing Magazine 23.2 (2006): 133-141.
- [2] Nguyen, Huy Do Nhat, et al. "A Comparative Study of DenseNets for Vietnamese Traditional Music Genre Classification." 2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE). IEEE, 2024.
- [3] Van Toan Pham, Ngoc Tran Ngo Quang, and Ta Minh Thanh. "Deep learning approach for singer voice classification of Vietnamese popular music." Proceedings of the 10th International Symposium on Information and Communication Technology. 2019.
- [4] <https://github.com/LTPhat/Vietnamese-Traditional-Music-Classification>
- [5] Dataset link: <https://www.kaggle.com/datasets/homata123/vntm-for-building-model-5-genres>