DOI: 10.15625/vap.2023.0016

VIPRIME: ỨNG DỤNG NHẬN DẠNG NHẠC CỔ TRUYỀN VIỆT NAM VỚI MẠNG HỌC SÂU TÍCH CHẬP

Trần Thanh Huy, Trần Minh Đạt, Huỳnh Gia Khương, Mã Trường Thành, Phạm Nguyên Khang, Đỗ Thanh Nghị

Trường Công nghệ thông tin và Truyền thông, Đại học Cần Thơ

{mtthanh,hgkhuong,pnkhang,dtnghi}ctu.edu.vn, tranhuymrp@gmail.com

TÓM TẮT: *Văn hoá phi vật thể vừa là một biểu trưng của một quốc gia vừa mang bản sắc của một dân tộc. Việc bảo tồn văn hoá phi vật thể luôn được nhiều nhà nghiên cứu quan tâm và đề xuất những giải pháp hiệu quả. Từ góc nhìn này, chúng tôi đã đặt sự chú ý đến các nhạc cổ truyền của Việt Nam, như ca trù, đờn ca tài tử, hát cung đình, chèo,... Nhìn chung, những loại hình văn hóa phi vật thể này đang dần bị "lãng quên" và các thế hệ trẻ đang ít quan tâm và chú ý đến chúng. Nhận thấy sức mạnh của trí tuệ nhân tạo, chúng tôi đã đưa hướng nhìn đến khía cạnh này để giúp bảo tồn các văn hoá Việt. Cụ thể, một mô hình để phân lớp kết hợp với giải thuật máy học truyền thống để nhận dạng nhạc cổ truyền Việt Nam đã được đề xuất. Ý tưởng chính là sử dụng mạng học sâu tích chập để nhận dạng thể loại của dòng nhạc và sẽ gợi ý một số tác phẩm nổi tiếng liên quan đến dòng nhạc đó, đồng thời hệ thống cũng cung cấp các thông tin liên quan. Một giải thuật để lựa chọn kết quả nhận dạng sẽ được đề xuất trong nghiên cứu này. Từ thực nghiệm, kết quả phân lớp của mô hình học sâu đạt hiệu quả mong đợi với độ chính xác trên 95%. Cuối cùng, hướng tiếp cận này sẽ được triển khai trên nền tảng Web và công bố mã nguồn cài đặt.*

Từ khóa: *Văn hóa phi vật thể, Trí tuệ nhân tạo, Mạng học sâu, Âm thanh, Giải thuật máy học.*

I. GIỚI THIỆU

Di sản văn hóa được xem như một niềm tự hào đối với mỗi quốc gia trên thế giới. Nó thể hiện được bản sắc dân tộc và biểu trưng cho từng vùng miền. Mỗi di sản văn hóa thường mang lại những giá trị to lớn trong đời sống của chúng ta về vật chất lẫn tinh thần. Nếu như di sản văn hóa vật thể là những thứ mà chúng ta có thể tận mắt chứng kiến cũng như khám phá vẻ đẹp của chúng thì ngược lại di sản văn hóa phi vật thể mang lại một điều gì đó đọng lại trong ký ức mỗi người bởi chính vẻ đẹp tâm hồn mà chúng mang lại. Mặc dù *không hiện hữu* nhưng các di sản văn hóa phi vật thể lại len lỏi trong *đời sống tinh thần* hàng ngày của mỗi người. Tại Việt Nam, có nhiều di sản văn hóa được Tổ chức Giáo dục, Khoa học và Văn hóa Liên hợp quốc (UNESCO) công nhận là di sản văn hóa thế giới cần được gìn giữ và bảo tồn. Theo [2], giới thiệu một số di sản văn hóa phi vật thể tại Việt Nam, trong đó có một số loại nhạc truyền thống từ xưa đến nay mang đậm bản sắc dân tộc như: ca trù, đờn ca tài tử, dân ca quan họ Bắc Ninh… Ở đây, chúng tôi gọi những dòng nhạc này là "*nhạc cổ truyền Việt Nam*". Thật vậy, những dòng nhạc cổ truyền này thể hiện được nét rất riêng của từng vùng miền ở Việt Nam mà đôi lúc chúng còn được xem như tinh hoa văn hóa của mỗi dân tộc anh em trong Cộng đồng các dân tộc Việt Nam *(64 dân tộc)*. Tuy nhiên, trong thời đại ngày nay, việc toàn cầu hóa dẫn đến làn sóng xâm nhập từ những thể loại nhạc ngoại quốc, xu hướng âm nhạc này lại trở thành thị hiếu của giới trẻ nên chúng trở nên rất phổ biến. Cũng chính vì thế mà các dòng nhạc cổ truyền lại dần bị "lãng quên", xuất phát từ việc này mà bài báo [10] được xuất bản như một lời động viên, nhắc nhở việc giáo dục thế hệ trẻ nói riêng cũng như mỗi người trong chúng ta nói chung biết cách giữ gìn cũng như trân trọng các di sản văn hóa phi vật thể của dân tộc, để giữ được tinh thần hòa nhập nhưng không hòa tan.

Nhìn chung, để có thể có ý thức gìn giữ cũng như bảo tồn di sản văn hóa thì việc nhận biết chúng là điều rất cần thiết trong mỗi người chúng ta. Đối với di sản văn hóa phi vật thể, chúng có thể được thể hiện qua những văn bản, tạp chí, hình ảnh chụp, video và cả âm thanh. Với việc điều tra những nghiên cứu liên quan, một số đề tài nhận dạng các di sản văn hóa phi vật thể với hình ảnh đạt được nhiều kết quả tốt và mong đợi. Chẳng hạn, các nghiên cứu [3, 5] xây dựng mô hình nhận dạng và phân loại các loại di sản văn hóa phi vật thể trên tập dữ liệu hình ảnh. Tuy nhiên, âm thanh từ các loại nhạc cổ truyền cũng chính là một nét đặc sắc của văn hóa phi vật thể mà chúng ta có thể dễ dàng cảm nhận được. Những làn điệu truyền thống của chèo, ca trù… sẽ không thể nhìn thấy trực quan mà chúng ta sẽ lắng nghe từng hơi thở và sức sống mà người nghệ sĩ mang lại. Theo nghiên cứu, điều tra, và khảo sát, các đề tài nghiên cứu về việc vận dụng trí tuệ nhân tạo trong việc phân loại các dòng nhạc cổ truyền Việt Nam vẫn chưa có nhiều công trình nghiên cứu liên quan. Chính vì thế, trong nghiên cứu này, chúng tôi đề xuất một số đóng góp chính như sau:

- Một tập dữ liệu âm thanh (10 giây và 20 giây) và được chuyển đổi sang dạng quang phổ về các nhạc cổ truyền Việt Nam. Tập dữ liệu này được thu thập từ các đoạn video trên Youtube với không có yếu tố bản quyền. Ở đây, chúng ta chuyển đổi và cắt những video dài thành những đoạn âm thanh 10 giây và 20 giây. Dựa trên ý kiến của chuyên gia và đặc trưng các thể loại hát mà đã không thu thập với 5 giây hoặc nhiều hơn 20 giây;
- Một mô hình học sâu kết hợp với giải thuật máy học (được gọi là CNN-SVM [11]) để phân loại 08 loại nhạc. Đồng thời tích hợp chúng vào một ứng dụng mang tên Vietnamprime (hay được viết tắt ViPrime) để cho người dùng dễ dàng sử dụng trong việc phân biệt các loại nhạc cổ truyền Việt Nam;
- Ngoài ra, để lựa chọn kết quả phân lớp, một giải thuật đã được đề xuất trong bài báo này;
- Cuối cùng, những thông tin bổ sung và các đoạn video liên quan đến bài hát đó sẽ được cung cấp.

Trong bài báo này, thay vì nhận dạng trực tiếp với đặc trưng MFCC [12] và các giải thuật máy học, chúng ta sẽ chuyển góc nhìn nhận dạng âm thanh về bài toán thị giác máy tính. Cụ thể, những đoạn âm thanh sẽ được thu thập và chuyển về dạng quang phổ (lưu trữ dưới dạng hình ảnh). Tại đây, bài toán nhận dạng bằng thị giác máy tính với các mạng học sâu tích chập sẽ được vận dụng để dự đoán kết quả. Tuy nhiên, nghiên cứu này sẽ không nhận dạng trực tiếp bằng CNN mà sẽ sử dụng CNN để trích đặc trưng và sử dụng giải thuật máy học để xây dựng mô hình nhận dạng. Lý do thực hiện hướng tiếp cận này là vì đặc trưng của hình quang phổ có sự lặp lại về thông tin và sự khác biệt về tần số âm thanh khi biểu diễn trên quang phổ sẽ có nhiều điểm khác nhau. Hơn nữa, trường hợp âm thanh nhiễu sẽ được mô hình máy học xử lý tốt. Một điểm đáng chú ý trong nghiên cứu này là cách lựa chọn kết quả nhận dạng. Vì dữ liệu đầu vào là một đoạn âm thanh dài trong khi mô hình nhận dạng sẽ được huấn luyện với đoạn âm thanh 10 giây và 20 giây. Do vậy, chúng ta sẽ cắt đoạn âm thanh thành các đoạn âm thanh nhỏ để phù hợp mô hình huấn luyện. Tuy nhiên, đôi khi trong cả đoạn âm thanh dài sẽ có những phân đoạn không có "tiếng hát - âm thanh" thì kết quả nhận dạng sẽ không chính xác nếu chúng ta lựa chọn một đoạn âm thanh ngẫu nhiên. Vì vậy, một giải thuật đơn giản để lựa chọn kết quả là cần thiết và mong đợi.

Bài báo này thì được cấu trúc như sau: những cơ sở lý thuyết mà sẽ vận dụng trong bài báo sẽ được trình bày trong Mục II; kế tiếp, Mục III sẽ giới thiệu mô hình đề xuất của nghiên cứu này và được trình bày chi tiết về các thành phần thực hiện; bài báo sẽ trình bày thực nghiệm của hệ thống và một thảo luận về kết quả của mô hình trên nền Web trong Mục IV; cuối cùng, một kết luận và hướng phát triển của bài báo sẽ cung cấp trong Mục V.

II. CƠ SỞ LÝ THUYẾT

A. Nhạc cổ truyền Việt Nam

Nhạc cổ truyền Việt Nam [2, 15] là một phần quan trọng trong di sản văn hóa của đất nước. Đây là thể loại nhạc truyền thống và cổ điển đã tồn tại từ lâu đời, thể hiện tinh thần, tâm hồn và đặc trưng văn hóa của người dân Việt Nam qua các thế kỷ. Nhạc cổ truyền Việt Nam gồm nhiều thể loại và dòng nhạc khác nhau, như trong Hình 1, cụ thể: (1) Nhạc cung đình: Được thể hiện tại các triều đại phong kiến, nhạc cung đình thường có tính chất trang trọng, tinh tế và thể hiện sự phấn đấu của các nhạc sĩ trong việc gìn giữ và phát triển văn hóa cung đình; (2) Nhạc ru: Nhạc này thường được sử dụng trong các lễ cúng, lễ hội hoặc để giúp người nghe tạo cảm xúc yên bình, thư thái; (3) Nhạc dân ca: Thể hiện tâm hồn và cuộc sống của người dân thông qua các bài hát thường được truyền đồng qua thế hệ; (4) Nhạc tài tử: Thường biểu diễn tại các sân khấu và các lễ hội dân gian, nhạc tài tử thường kết hợp nhiều loại nhạc cụ và tiết tấu nhịp điệu độc đáo; (5) Nhạc cải lương và hát bội: Là các thể loại kịch truyền thống, thường sử dụng âm nhạc để diễn tả và tạo tình huống cho các vở kịch; và những loại khác. Nhìn chung, nhạc cổ truyền Việt Nam không chỉ là một phần không thể thiếu của cuộc sống và văn hóa của người dân, mà còn là biểu tượng của sự đa dạng và sự độc đáo của ngôn ngữ âm nhạc Việt Nam. Thật vậy, khi nhắc đến nhạc cổ truyền Việt Nam, nhiều loại hình âm nhạc do người Việt sáng tác đã ra đời từ rất sớm và được lưu truyền từ đời này sang đời khác, và cho đến nay các thế hệ vẫn sáng tác, thưởng thức.

Hình 1. Các thể loại nhạc cổ truyền Việt Nam

Có nhiều thể loại nhạc cổ truyền khác nhau và đặc trưng cho mỗi vùng miền trên đất nước Việt Nam. Xét về các loại hình âm nhạc cổ truyền thì có thể kể đến nhã nhạc cung đình, chèo, tuồng, hát xẩm, đờn ca tài tử, dân ca, ca trù… Chúng có thể thể hiện qua các hình thức đơn ca, song ca hoặc biểu diễn theo nhóm. Người biểu diễn có thể kết hợp với các loại nhạc cụ truyền thống vô cùng đặc sắc như đàn bầu, đàn cò, đàn nguyệt,… đã góp phần tạo nên nét đặc trưng riêng cho âm nhạc cổ truyền Việt Nam.

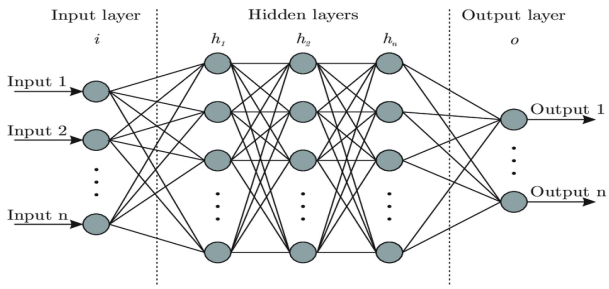
Trong nghiên cứu này, chúng tôi tập trung vào 08 loại nhạc truyền thống phổ biến và nổi tiếng tại Việt Nam với đa dạng nhiều vùng miền khác nhau, bao gồm: Ca trù, chèo, hát chầu văn, hò, nhạc cung đình Huế, nhạc tài tử, quan họ và xẩm. Mặc dù còn rất nhiều loại dòng nhạc khác nhưng bài báo này sẽ tập trung vào các loại hình âm nhạc này vì một trong số chúng đã được UNESCO công nhận và có nhiều giá trị văn hóa trong kho tàng âm nhạc truyền thống Việt Nam. Hơn nữa, những loại hình âm nhạc được chọn này trải dài trên đất nước hình chữ "S" với sự phong phú về tập quán và bản sắc dân tộc mà cần được lưu truyền và giữ gìn.

Để bảo tồn những loại nhạc này, chúng tôi sẽ tận dụng trí tuệ nhân tạo mà sẽ được giới thiệu trong mục kế tiếp.

B. Mạng học sâu tích chập

Mạng học sâu tích chập (Convolutional Neural Network - CNN) [13, 14] là một kiến trúc mạng nơ-ron nhân tạo được thiết kế đặc biệt cho việc xử lý và phân tích dữ liệu hình ảnh và âm thanh. CNN là một trong những công cụ quan trọng trong lĩnh vực thị giác máy tính và xử lý ngôn ngữ tự nhiên.

Kiến trúc CNN được thiết kế dựa trên các lớp tích chập và các lớp kết nối đầy đủ (fully connected). Các lớp tích chập giúp mạng học cách nhận diện các đặc trưng cục bộ trong dữ liệu, trong khi các lớp kết nối đầy đủ giúp tổng hợp thông tin để thực hiện phân loại hoặc dự đoán. Các đặc điểm chính của CNN như trong Hình 2, bao gồm: (1) *Lớp tích chập*: Các lớp này thực hiện phép tích chập trên dữ liệu đầu vào để tìm ra các đặc trưng cục bộ như cạnh, góc, hoặc mẫu trong hình ảnh. Các lớp này giúp giảm thiểu số lượng tham số cần huấn luyện và tạo ra tính năng tự động rút trích đặc trưng. (2) *Lớp gộp* (pooling): Các lớp gộp giúp giảm kích thước dữ liệu và làm giảm độ phức tạp của mạng. Lớp gộp thường thực hiện việc chọn ra giá trị quan trọng nhất trong các vùng của đặc trưng đã được rút trích. (3) *Lớp kết nối đầy đủ*: Sau khi qua các lớp tích chập và gộp, thông tin được đưa vào các lớp kết nối đầy đủ để thực hiện phân loại hoặc dự đoán. Nhìn chung, CNN đã đạt được nhiều thành tựu ấn tượng trong việc xử lý hình ảnh, bao gồm nhận dạng đối tượng, nhận biết khuôn mặt, phân loại hình ảnh và nhiều ứng dụng khác liên quan đến thị giác máy tính.

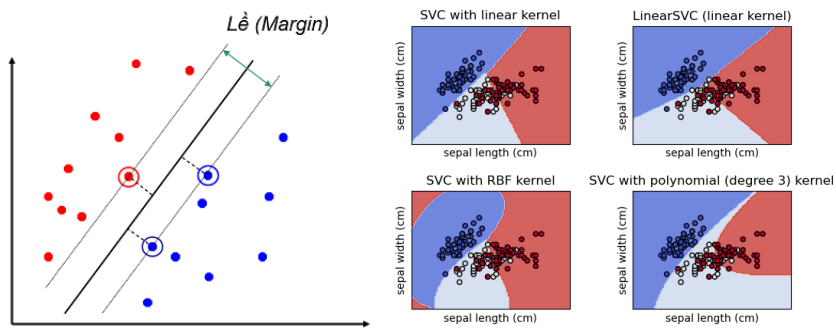


Hình 2. Mô hình mạng học sâu tích chập

Ở đây, chúng ta sẽ cung cấp một số thông tin chi tiết về mô hình này cũng như cách sử dụng cho mô hình trong bài báo này: (1) Tại lớp trích lọc đặc trưng ảnh, ngoài tầng đầu vào (Input Layer), người ta thường triển khai một số tầng tiêu biểu khác như tầng tích chập (Convolutional Layer) dùng để trích đặc trưng ảnh đầu vào thông qua các bộ lọc với phép toán tích chập, tầng ReLU để đưa ảnh một mức ngưỡng nhằm loại bỏ các giá trị âm không cần thiết mà có thể ảnh hưởng đến việc tính toán của các tầng sau đó, tầng Pooling với chức năng chính là giảm chiều không gian của đầu vào để giảm độ phức tạp tính toán trên mô hình. Lưu ý rằng, nghiên cứu này sẽ tận dụng những tầng này để trích đặc trưng ảnh quang phổ. Kế tiếp, tại lớp phân loại người ta thường quan tâm đến mạng liên kết đầy đủ (Fully Connected Layer - FC). Tại tầng này thì mỗi một nơron sẽ được liên kết với mọi nơron của tầng FC khác. Tầng FC sẽ nhận các ảnh từ các tầng trước (có thể là Conv Layer hoặc Pooling Layer) nhưng phải được làm phẳng (Flatten) thành vector trước khi đưa vào. Sau các tầng liên kết đầy đủ thì sẽ đến tầng đầu ra, tại tầng đầu ra này người ta có thể sử dụng các hàm kích hoạt phù hợp với bài toán đặt ra. Các hàm kích hoạt thường gặp như Softmax function hoặc Sigmoid function. Tuy nhiên, hướng tiếp cận của bài báo này sẽ thay Softmax bởi giải thuật máy học SVM.

C. Máy học hỗ trợ vector

SVM (Support Vector Machine) [16] là một thuật toán trong lĩnh vực máy học, thuộc loại học có giám sát, được sử dụng chủ yếu cho các tác vụ phân loại và hồi quy. SVM là một phương pháp mạnh mẽ để tìm ra một ranh giới phân tách tốt nhất giữa các lớp dữ liệu. Ý tưởng chính của SVM là tìm ra một đường phân chia (đối với bài toán phân loại hai lớp) hoặc siêu phẳng (đối với bài toán phân loại nhiều lớp) sao cho khoảng cách từ các điểm dữ liệu gần nhất đến đường phân chia (siêu phẳng) là lớn nhất. Các điểm dữ liệu gần nhất này được gọi là các vectơ hỗ trợ (support vectors).



Hình 3. Mô hình phân lớp SVM và hàm SVC (của scikit-learn) với chọn hàm nhân

SVM có thể áp dụng các hàm kernel để ánh xạ dữ liệu từ không gian ban đầu sang không gian cao chiều hơn, giúp tạo ra các đường phân chia phức tạp hơn và phù hợp với dữ liệu không tuyến tính. Các hàm kernel phổ biến bao gồm hàm tuyến tính, hàm đa thức, hàm Gaussian (RBF - Radial Basis Function) và nhiều loại khác.

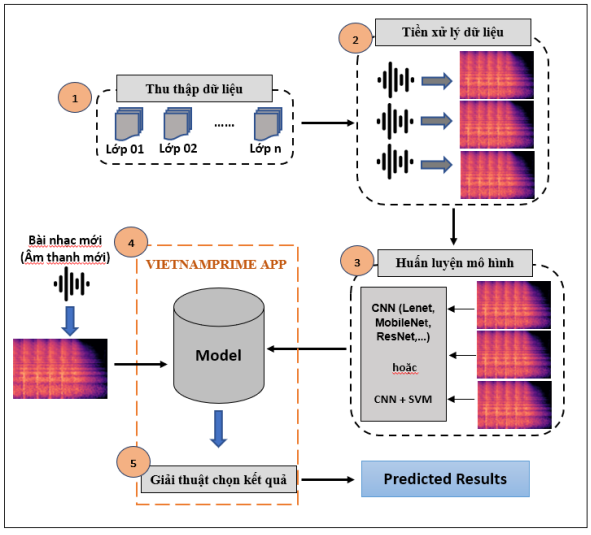
SVM có ứng dụng rộng rãi trong nhiều lĩnh vực, bao gồm nhận dạng chữ viết tay, nhận biết đối tượng trong hình ảnh, phân loại văn bản, và nhiều tác vụ phân loại và hồi quy khác. Thuật toán SVM thường đạt hiệu suất tốt trong việc xử lý các bộ dữ liệu có sự phân tách rõ ràng giữa các lớp. Do vậy, bài toán nhận dạng âm thanh với đặc trưng từ mạng học sâu sẽ phù hợp với giải thuật SVM.

III. ĐỀ XUẤT PHƯƠNG PHÁP

Để nhận dạng dòng nhạc cổ truyền Việt Nam, mô hình tổng quan của hướng tiếp cận này sẽ được giới thiệu trong mục này. Hơn nữa, giải thuật lựa chọn sẽ được trình bày trong chi tiết.

A. Mô hình ViPrime

Mô hình khung (framework) để dự đoán dòng nhạc cổ truyền Việt Nam đã được đề xuất, với tên gọi Vietnam Prime (viết tắt, ViPrime), bao gồm 05 bước chính như trong Hình 5.



Hình 4. Kiến trúc hệ thống đề xuất

Một mô tả tiết về mô hình ViPrime như sau:

Bước 1. **Thu thập dữ liệu**: Tiến hành thu thập các đoạn âm thanh về nhạc cổ truyền bao gồm tất cả 8 thể loại. Các video về những loại nhạc cổ truyền này được thu thập từ Youtube. Do đó, các đoạn video dài ngắn khác nhau, thời gian trung bình cho các video này khoảng 0,5 – 2,78 giờ. Mỗi dữ liệu liên quan đến từng loại nhạc sẽ được lưu trữ trong một thư mục riêng, nó tương ứng với một lớp trong quá trình xây dựng mô hình phân lớp. Lưu ý rằng, chúng ta sẽ thực hiện xóa nhiễu và cắt âm thanh thành những đoạn 10 giây và 20 giây cho quá trình phân lớp. Lý do chúng tôi lựa chọn 10 giây và 20 giây sẽ được trình bày ở Mục IV.A;

Bước 2. **Tiền xử lý dữ liệu:** Hướng tiếp cận trong bài báo này là chuyển các dữ liệu âm thanh về dạng ảnh quang phổ (phổ tần số của tín hiệu hay Spectrogram) để tạo bộ dữ liệu đầu vào cho các mô hình huấn luyện. Để thực hiện việc này, gói thư viện Librosa sẽ được sử dụng để thực hiện thao tác chuyển đổi âm thanh thì hình ảnh. Nhấn mạnh rằng, bài báo sẽ không thực hiện trực tiếp nhận dạng âm thanh bằng các đặc trưng âm thanh, chẳng hạn MFCC, mà chuyển sang giải quyết bài toán của thị giác máy tính. Ở bước này, các âm thanh đã được lọc nhiễu với việc lấy trung bình các tần số gần (sử dụng các hàm của librosa).

Bước 3. **Huấn luyện mô hình:** Sau khi có được dữ liệu huấn luyện, việc tiến hành thử nghiệm trên các mô hình học sâu như MobileNet, InceptionV3, Lenet, Resnet50 sẽ được thực hiện. Hơn nữa, bài báo này đã thực nghiệm với việc kết hợp giữa mô hình mạng học sâu với giải thuật máy học vector hỗ trợ (SVM) *(bằng phương pháp Transfer Learning trong việc trích xuất đặc trưng dữ liệu qua mô hình học sâu).* Tại đây, bài báo sẽ cung cấp một so sánh giữa các mô hình và các bộ tham số khác nhau nhằm tìm ra mô hình tốt nhất cài đặt ứng dụng.

Bước 4. **Xây dựng ứng dụng VietnamPrime:** Dựa trên mô hình đạt được ở bước 03, chúng tôi sẽ lưu trữ lại mô hình và đồng thời cài đặt ứng dụng trên nền tảng Web với gói thư viện Flask. Bước này cũng được biết như kiểm tra mô hình và triển khai hướng tiếp cận vào hệ thống/ứng dụng thực tế.

Bước 5. **Giải thuật lựa chọn kết quả:** Vì dữ liệu đầu vào của người dùng sẽ đa dạng và không thể là 10 giây hoặc 20 giây như mô hình đã huấn luyện. Hơn nữa, nghiên cứu này cũng đã thực nghiệm kiểm tra với âm thanh dài cho mô hình đã được huấn luyện. Kết quả cho thấy độ chính xác không cao. Do vậy, thay vì thực hiện nhận dạng trực tiếp với âm thanh dài, một giải thuật đơn giản, mang tên S2S, đã

được đề xuất để lựa chọn kết quả. Cụ thể, giả sử có một đoạn âm thanh dài 1 phút (60 giây). Chúng tôi sẽ cắt ngẫu nhiên 4-6 đoạn âm thanh (mỗi đoạn 10 giây). Sau đó, mỗi đoạn âm thanh sẽ được đưa qua mô hình CNN để nhận dạng. Cuối cùng mô hình sẽ dựa trên giải thuật S2S (nền tảng chính là luật bình chọn số đông) để đưa ra kết quả nhận dạng. Giải thuật sẽ được trình bày trong mục con kế tiếp. Giải thuật sẽ được trình bày trong mục con kế tiếp.

B. Giải thuật S2S

Ý tưởng của giải thuật là chia nhỏ để chọn (Split to Select, viết tắt S2S). Cụ thể, từ đoạn âm thanh dài (ký hiệu τ), chúng ta cắt ra thành nhiều phần (sử dụng hàm $\omega(\tau)$) với việc chọn lấy ngẫu nhiên 10 giây không trùng nhau. Kế tiếp, sẽ thực hiện thu thập các kết quả nhận dạng từ các phần được cắt ra với mô hình máy học (ký hiệu Δ)) đã lựa chọn (từ bước 3). Ví dụ, giả sử chúng ta có 01 đoạn âm thanh 03 phút, khi đó, chúng ta sẽ cắt ra thành 18 đoạn 10 giây. Và sử dụng mô hình máy học để nhận dạng 18 đoạn đó để thu kết quả. Kết quả sẽ bao gồm "tên của dòng nhạc" (ký hiệu μ_{Name}) và "độ chính xác" (μ_{Acc}). Ở đây, chúng tôi sẽ thiết lập một biến α để làm ngưỡng lựa chọn (mặc định là 90). Dựa trên kết quả thu được với độ chính xác cao, chúng tôi sẽ lưu vào một mảng (ρ) và sử dụng luật bình chọn số đông để đưa ra quyết định cuối cùng.

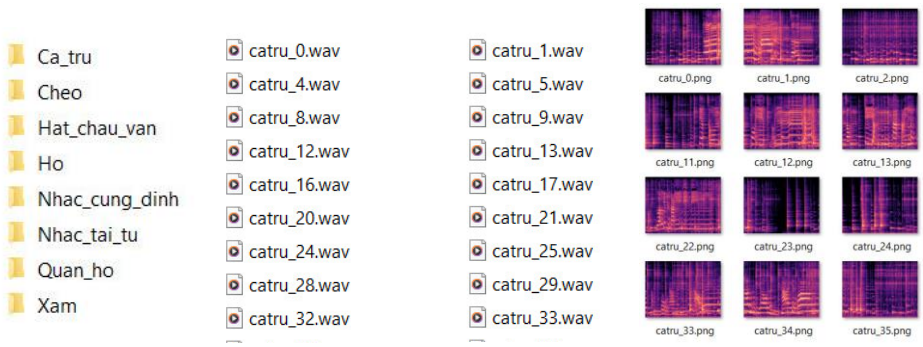
Giải thuật: SplitToSelect (Viết tắt, S2S)
Đầu vào: Âm thanh dài (ký hiệu τ), mô hình nhận dạng (ký hiệu Δ), số giây cắt (g)
Đầu ra: Kết quả nhận dạng

$\alpha \leftarrow 90, \rho \leftarrow \emptyset, \sigma \leftarrow \emptyset,$
Function $S2S(\tau, \Delta, g)$:
  $\vartheta \leftarrow \omega(\tau, g)$ // $\omega(\text{âm thanh dài, số giây})$: Hàm cắt âm thanh ngẫu nhiên
  for each $\varepsilon \in \vartheta$:
    $\mu_{Name}, \mu_{Acc} \leftarrow \Delta(\varepsilon)$
    if $\mu_{Acc} \geq \alpha$:
      $\rho \leftarrow \rho \cup \mu_{Name}$
  $K \leftarrow \delta(\rho)$ // $\delta(\rho)$: Hàm thống kê các kết quả, ví dụ: $a: 3, b: 7$
  $Re = Majority\ (K)$
  return Re
end

IV. THỰC NGHIỆM VÀ THẢO LUẬN

A. Dữ liệu và cấu hình máy tính

Nghiên cứu này đã thực hiện huấn luyện với 08 lớp[1], bao gồm ca trù, chèo, hát chầu văn, hò, nhạc cung đình Huế, nhạc tài tử, quan họ và xẩm. Cho mỗi lớp chúng ta sẽ có 500 mẫu được thu thập từ các video dài. Với mỗi mẫu sẽ được sinh ra một ảnh spectrogram tương ứng (ảnh quang phổ). Tổng số lượng ảnh để huấn luyện là 4,000. Ở đây, chúng tôi đã thực nghiệm với tỷ lệ 8:2 (cụ thể, 80% để huấn luyện và 20% để kiểm tra). Trong tập huấn luyện, dữ liệu lại được chia theo tỷ lệ 9:1 lần nữa (cụ thể, 90% cho huấn luyện và 10% cho validation). Thực nghiệm này đã sử dụng earlystop để dừng lại cho các mô hình huấn luyện bằng CNN. Hướng tiếp cận này đã thực hiện với hai tập dữ liệu, gồm 10 giây và 20 giây. Một vài lý do chúng tôi lựa chọn hai tập dữ liệu này là (1) theo ý kiến của chuyên gia thì các dòng nhạc sẽ khác nhau về độ dài của "âm ngân". Nếu chúng ta chỉ lấy 5 giây thì sẽ không đủ độ dài để thể hiện dòng nhạc đó, chẳng hạn "Nhạc tài tử" với một câu ngân giọng rất dài (đôi khi 10-20 giây) hoặc đôi khi có "câu hò" trên 10 giây. Khi đó nếu chúng ta chỉ thu thập 05 giây sẽ rất khó để dự đoán sự khác nhau giữa chúng; (2) sự lại đi lặp lại của một số dòng nhạc sẽ tạo nên sự trùng lặp trong quá trình dự đoán; (3) với đoạn âm thanh dài sẽ tốn nhiều thời gian để trích xuất thông tin và huấn luyện; (4) nếu lấy trên 30 giây thì sẽ không phù hợp với những trường hợp thực tế. Nếu người dùng muốn dự đoán trước tiếp thì gần như không thực tế để người dùng chờ 30 giây để thu âm thanh. Nó sẽ là phù hợp với 10 giây đến 20 giây cho một đoạn thu âm thanh nhanh *(nếu khách du lịch muốn nhận dạng trực tiếp)*.



Hình 5. Hình dữ liệu thu thập được lưu trữ vào các thư mục và chuyển từ âm thanh sang ảnh quang phổ

[1] https://github.com/thanhhuytran919/data_spectrogram_vietnamese_music_instruments

Đối với môi trường cài đặt hệ thống, các cấu hình và thư viện được cài đặt như trong Bảng 1. Hệ thống được cài đặt bằng mã lệnh Python để xây dựng mô hình máy học và triển khai trên nền Web. Ứng dụng Web được triển khai với Flask framework. Đây là một framework phát triển ứng dụng web phía máy chủ được xây dựng bằng ngôn ngữ lập trình Python. Hệ thống lựa chọn Flask để triển khai vì chúng được thiết kế đơn giản, linh hoạt và dễ dàng để phát triển các ứng dụng web từ những dự án nhỏ đến những ứng dụng phức tạp hơn. Để chuyển đổi từ âm thanh sang ảnh dạng quang phổ, nghiên cứu hiện tại đang cài đặt và thực nghiệm với thư viện Librosa[2] với phiên bản 0.10.1.

Bảng 1. Cấu hình máy tính và thư viện sử dụng

Cấu hình	**Phiên bản / Kích thước**
Hệ điều hành	Windows 10
CPU	AMD Ryzen 5-2500U 2.0 GHz up to 3.6 GHz
RAM	16 GB
Card đồ họa	NVIDIA GEFORCE GTX 1050 4GB GDDR5 + Radeon Vega 8 graphics
Cuda	11.2.0 – 460.89
TensorFlow, Keras [19]	Phiên bản 2.7.0

Lưu ý rằng tập dữ liệu thu thập và ứng dụng đã được công bố trên Github[3].

B. Đánh giá mô hình

Đối với việc huấn luyện mô hình, nghiên cứu này đã thực nghiệm 05 mô hình mạng học sâu với các kiến trúc khác nhau. Cụ thể, bài báo kiểm tra với MobilenetV2, Inception V3, Lenet-5, Resnet50, và CNN-SVM. Đối với giải thuật SVM, bộ tham số của hàm nhân RBF đã được thực nghiệm như sau: $C = 100000$ và $gamma = 0.001$. Hơn nữa, sau khi sử dụng thư viện librosa để chuyển đổi âm thanh (tập tin có phần mở rộng là wav) thành ảnh quang phổ, ảnh đầu vào sẽ được resize lại với kích thước 128×128. Lưu ý rằng, trong quá trình thực nghiệm chúng tôi đã kiểm tra với nhiều kích thước đầu vào khác nhau, với kích thước 128×128 đã đạt được độ chính xác cao nhất. Trong nghiên cứu này, chúng tôi đã kiểm tra với đặc trưng MFCC và sử dụng SVM để so sánh với hướng tiếp cận của bài báo.

Mặt khác, các mô hình đều được kiểm tra với 10 lần thực nghiệm kiểm tra và lấy giá trị trung bình của chúng để viết vào Bảng 2. Tất cả các mô hình đã được thiết lập với epochs 50, tuy nhiên với số epochs khoảng 20-25 thì earlystop đã dừng lại. Một điểm quan trọng là nghiên cứu đã thực hiện kiểm tra với tập dữ liệu 10 giây và 20 giây. Các kết quả đánh giá độ chính xác của các mô hình trình bày trong Bảng 2.

Bảng 2. Bảng kết quả đánh giá các mô hình huấn luyện

Data	**Model**	**F1**		**Recall**		**Precision**		**Accuracy**	
		BS24	**BS32**	**BS24**	**BS32**	**BS24**	**BS32**	**BS24**	**BS32**
10 seconds	**MobilenetV2**	92.13%	93.37%	92.12%	93.38%	92.29%	93.53%	92.12%	93.38%
	InceptionV3	77.33%	79.29%	77.75%	79.38%	78.67%	80.04%	77.75%	79.38%
	Lenet-5	**94.41%**	**94.23%**	**94.50%**	**94.25%**	**94.48%**	**94.28%**	**94.50%**	**94.25%**
	CNN + SVM	**95.36%**	**95.26%**	**95.38%**	**95.38%**	**95.38%**	**95.38%**	**95.38%**	**95.38%**
	ResNet50	56.93%	51.15%	59.50%	53.12%	66.39%	75.19%	59.50%	53.12%
	MFCC+SVM	83.53%		81.64%		85.24%		86.54%	
20 seconds	**MobilenetV2**	85.33%	88.63%	85.25%	88.62%	86.01%	89.13%	85.25%	88.62%
	InceptionV3	78.17%	89.27%	78.38%	89.25%	83.54%	90.09%	78.38%	89.25%
	Lenet-5	**91.62%**	**90.74%**	**91.62%**	**90.75%**	**91.65%**	**91.14%**	**91.62%**	**90.75%**
	CNN + SVM	**87.67%**	**92.26%**	**87.75%**	**92.25%**	**87.75%**	**92.25%**	**87.75%**	**92.25%**
	ResNet50	46.20%	47.27%	49.12%	50.62%	55.54%	52.42%	49.12%	50.62%
	MFCC+SVM	81.52%		80.15%		83.43%		83.78%	

Từ bảng kết quả thử nghiệm trên, chúng ta nhận thấy việc sử dụng mô hình Lenet-5 và mô hình (CNN-SVM) kết hợp giữa mô hình học sâu và giải thuật máy học truyền thống mang lại kết quả rất khả quan. Thực nghiệm trong nghiên cứu này đã kiểm tra với CNN-SVM với nhiều kiến trúc khác nhau. Nhận thấy với kiến trúc Lenet-5 thì phù hợp với tập dữ liệu đã được thu thập với độ chính xác cao nhất. Do vậy, trong bài báo này chỉ ghi nhận kết quả của mô hình CNN (Lenet) + SVM. Hơn nữa, vì dữ liệu quang phổ sẽ khác nhau nếu giọng khác nhau (cùng một chữ hoặc một câu) nên việc thực nghiệm sẽ thực hiện một số batch size (8, 16, 24, 32, 64). Tuy nhiên trong bài báo này chỉ đề cập đến những batchsize có độ chính xác tốt nhất. Nhìn chung, Batchsize=24 với 10 giây cho độ chính xác cao nhất.

Bài báo sẽ cung cấp một thảo luận về kết quả nghiên cứu và giới thiệu một ứng dụng đã được cài đặt trên nền tảng Web cho hướng tiếp cận này trong mục kế tiếp.

[2] https://librosa.org/
[3] https://github.com/thanhhuytran919/vietnamprime_app

C. Thảo luận và triển khai hệ thống

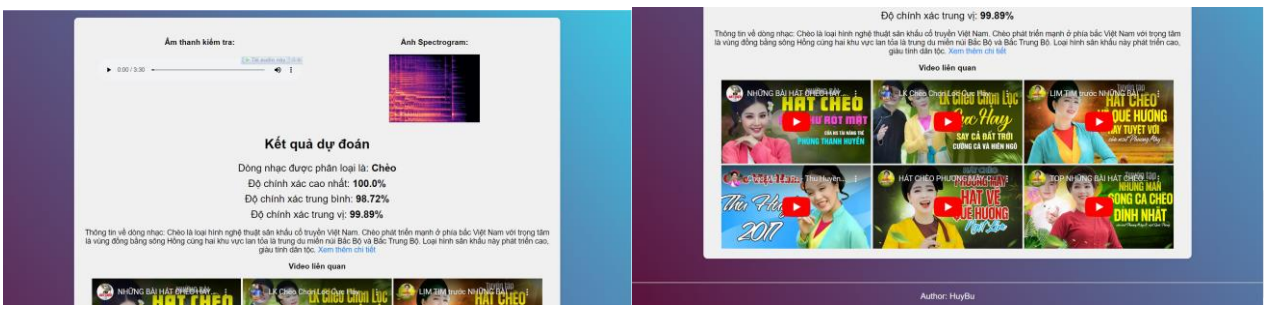
Quan sát kết quả huấn luyện từ mô hình, chúng tôi chọn mô hình tốt nhất để tiến hành xây dựng ứng dụng, mang tên *VietnamPrime (ViPrime)*. Chúng tôi đã cài đặt một giao diện trực quan với các thao tác đơn giản. Một số chức năng chính như sau:

- Người dùng chọn một tập tin âm thanh từ giao diện như Hình 7. Kế tiếp, người dùng chọn "Dự đoán" để tiến hành trích xuất đặc trưng và phân lớp dòng nhạc.
- Khi xác định được dòng nhạc, ứng dụng còn nhiều thông tin chi tiết như: loại nhạc truyền thống, độ chính xác dự đoán cũng như có kèm theo quang phổ của âm thanh đang kiểm tra, và thông tin của dòng nhạc đó. Hơn nữa, các video liên quan đến dòng nhạc được dự đoán sẽ được gợi ý để người dùng có thể nghe với các nhạc phẩm nổi tiếng (xem Hình 8).

Hình 6. Giao diện Vietnamprime

Đối với việc dự đoán kết quả từ đoạn âm thanh đầu vào, chúng tôi đề xuất phương pháp kiểm tra: Do mỗi đoạn âm thanh có thể sẽ có những đoạn *không có tiếng* hoặc những đoạn *có lẫn tạp âm* nên chúng tôi đề xuất như ở Bước 5, và rồi tiến hành dự đoán trên các đoạn âm thanh này và kết quả cuối cùng sẽ được đưa ra dựa trên bình chọn số đông (Giải thuật S2S). Hơn nữa, chúng tôi cũng cung cấp kết quả có độ chính xác cao nhất trong danh sách các kết quả từ âm thanh dài và cũng đã hiển thị trên ứng dụng về độ chính xác trung bình và trung vị. Một đều quan tâm rằng, dựa trên độ chính xác trung bình và trung vị để chúng tôi trả lời đoạn âm thanh đó có phải là dòng nhạc trong bộ phân lớp của chúng tôi hay không.

Nhìn chung kết quả dự đoán của bài báo này đạt được kết quả mong đợi. Tuy nhiên một số trường hợp kết quả chưa bao quát, do vậy hệ thống cần thực hiện thêm dữ liệu nguồn dữ liệu để có kết quả tốt hơn. Mặt khác, hệ thống hiện tại chỉ cho phép người dùng đăng tải đoạn âm thanh dài tối đa 5 phút. Do vậy, hệ thống cần cải tiến về giải thuật lựa chọn (S2S) và kể cả giải thuật đầu vào cắt âm thanh. Dự định của chúng tôi là xây dựng một mô hình máy học để tìm các đoạn âm thanh có nhiều sự khác biệt để cắt nhằm tang khả năng nhận dạng.



Hình 7. Giao diện kết quả từ VietnamPrime

V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi đã xây dựng được mô hình nhận dạng được âm nhạc cổ truyền Việt Nam dựa trên bộ dữ liệu thu thập được từ thực tế. Việc áp dụng được các mô hình học sâu vào trong xây dựng mô hình cùng với việc kết hợp giải thuật máy học truyền thống đã mang lại kết quả khá tốt và đạt độ chính xác tốt nhất xấp xỉ 95%. Với các mô hình này, chúng tôi cũng tiến hành xây dựng một website để thử nghiệm thông qua việc phân đoạn được các khúc nhạc cổ truyền Việt Nam giúp cho người dùng có thể dễ dàng nhận biết được các loại nhạc khác nhau. Điều đó giúp cho mọi người có một sự tiếp cận dễ hơn với các loại hình âm nhạc đang có nguy cơ bị lãng quên này. Trong tương lai, chúng tôi dự định mở rộng thêm các dòng nhạc cùng với đó là thêm các lớp bù của các âm thanh không phải là dòng nhạc truyền thống (chẳng hạn, pop, rock…) để hệ thống đưa ra quyết định phải nhạc cổ truyền hay không. Dĩ nhiên, với kết quả này sẽ không hoàn toàn tốt cho việc nhận dạng nếu người dùng đưa một dòng nhạc của nước ngoài vào hệ thống mà có giai điệu tương tự. Do vậy, chúng tôi sẽ phát triển hệ thống bằng mô hình phân tầng trong những nghiên cứu kế tiếp. Chúng

tôi cũng sẽ phát triển thêm các ứng dụng thông minh nhằm có khả năng tương tác dễ dàng với người dùng, đặc biệt là bộ phận giới trẻ ngày nay. Với mong muốn cuối cùng là cố gắng duy trì và phát triển được các dòng nhạc được xem như tinh hoa văn hóa của nước nhà.

TÀI LIỆU THAM KHẢO

[1] Anjali, A. Kumar and N. Birla, Voice Command Recognition System based on MFCC and DTW, International Journal of Engineering Science and Technology, 2 (12), 2010.

[2] Central Theoretical Council. (2021). 14 Representative Intangible Cultural Heritage of Humanity in Vietnam: http://hdll.vn/vi/nghien-cuu-trao-doi/14-di-san-van-hoa-phi-vat-the-dai-dien-cua-nhan-loaitai-viet-nam.html

[3] Cosovic, Marijana & Jankovic Babic, Radmila. (2020). CNN Classification of the Cultural Heritage Images. 10.1109/INFOTEH48170.2020.9066300.

[4] Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods. Cambridge University, New York (2000)

[5] Do, TN., Pham, TP., Nguyen, HH., Pham, NK. (2021). Visual Classification of Intangible Cultural Heritage Images in the Mekong Delta. In: Belhi, A., Bouras, A., Al-Ali, A.K., Sadka, A.H. (eds) Data Analytics for Cultural Heritage. Springer, Cham. https://doi.org/10.1007/978-3-030-66777-1_4

[6] Do, T., Pham, T., Pham, N., Nguyen, H., Tabia, K., Benferhat, S.: Stacking of SVMS for classifying intangible cultural heritage images. In: Advanced Computational Methods for Knowledge Engineering. Proceedings of the 6th International Conference on Computer Science, Applied Mathematics and Applications ICCSAMA 2019. Advances in Intelligent Systems and Computing, vol. 1121, pp. 186–196. Springer, Berlin (2019)

[7] Durairaj, Prabakaran & Sriuppili, S.. (2021). Speech Processing: MFCC Based Feature Extraction Techniques- An Investigation. Journal of Physics: Conference Series. 1717. 012009. 10.1088/1742-6596/1717/1/012009.

[8] Fiorucci, M., Khoroshiltseva, M., Pontil, M., Traviglia, A., Bue, A.D., James, S.: Machine learning for cultural heritage: a survey. Pattern Recognit. Lett. 133, 102–108 (2020).

[9] Hossan, Md & Memon, Sheeraz & Gregory, Mark. (2011). A novel approach for MFCC feature extraction. 1 - 5. 10.1109/ICSPCS.2010.5709752.

[10] Tran Minh Duc, " The necessity for education of intangible cultural heritage in schools for Vietnam's sustainable development in the current context", American Research Journal of Humanities & Social Science (ARJHSS), Volume-05, Issue-09, pp-50-56, 2022.

[11] Chaganti, Sai Yeshwanth, et al. "Image Classification using SVM and CNN", 2020 International conference on computer science, engineering and applications (ICCSEA). IEEE, 2020.

[12] Han, Wei, et al. "An efficient MFCC extraction method in speech recognition," 2006 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2006.

[13] Hussain, Mahbub, Jordan J. Bird, and Diego R. Faria. "A study on CNN transfer learning for image classification." Advances in Computational Intelligence Systems: Contributions Presented at the 18th UK Workshop on Computational Intelligence, September 5-7, 2018, Nottingham, UK. Springer International Publishing, 2019.

[14] Li, Qing, et al. "Medical image classification with convolutional neural network," 2014 13th international conference on control automation robotics & vision (ICARCV). IEEE, 2014.

[15] Janse, Olov RT. "On the origins of traditional Vietnamese music," Asian Perspectives 6.1/2 (1962): 145-162.

[16] Wang, Lipo, ed. Support vector machines: theory and applications. Vol. 177. Springer Science & Business Media, 2005.

VIPRIME: TRADITIONAL VIETNAMESE MUSIC RECOGNITION APPLICATION WITH CONVOLUTIONAL NEURAL NETWORK

Tran Thanh Huy, Tran Minh Dat, Huynh Gia Khuong, Ma Truong Thanh, Pham Nguyen Khang, Do Thanh Nghi

ABSTRACT: *Intangible culture symbolizes both a nation's and its people's essence. Preserving intangible cultural heritage attracts significant attention from researchers who propose effective solutions. With this perspective, we've focused on Vietnam's traditional music genres like ca trù, đờn ca tài tử, hát cung đình, and chèo. These forms of intangible culture are gradually fading, with younger generations showing less interest. Recognizing the power of artificial intelligence, we've shifted our focus to help preserve Vietnamese culture. Specifically, we've proposed a model combining classification and traditional machine learning to identify traditional Vietnamese music. The core idea involves using a convolutional deep learning network to classify music genres and suggest renowned works related to those genres while also providing relevant information. An algorithm for selecting recognition results will be proposed in this study. Experimental results show that the classification performance of the deep learning model achieves the expected effectiveness with an accuracy rate above 95%. Ultimately, this approach will be implemented on a web platform.*