TẠP CHÍ KHOA HỌC ĐẠI HỌC HẠ LONG
Scientific Journal of Ha Long University
http://uhl.edu.vn/

ỨNG DỤNG MẠNG NƠ-RON TẾ BÀO VỚI LUẬT HỌC PERCEPTRON HỒI QUY NHẬN DẠNG CẢM XÚC QUA GIỌNG NÓI

Đinh Bùi Thu Linh[1], Nguyễn Quang Hoan[1*], Đoàn Hồng Quang[2]

[1]Trường Đại học Thuỷ lợi

[2]Viện Ứng dụng Công nghệ, Bộ Khoa học và Công nghệ

**Email: quanghoanptit@gmail.com*

Ngày nhận bài: 27/06/2025 *Ngày nhận bài sửa sau phản biện: 28/08/2025* *Ngày chấp nhận đăng: 04/09/2025*

TÓM TẮT

Bài báo đề xuất phương pháp nhận dạng cảm xúc qua giọng nói sử dụng mạng nơ-ron tế bào (CeNNs) với luật học Perceptron hồi quy (RPLA: Regression Perceptron Learning Algorithm), một giải pháp mới. Mô hình phân loại các cảm xúc thành hai nhóm: tích cực và tiêu cực từ tín hiệu âm thanh. Thử nghiệm được tiến hành trên bộ dữ liệu hợp nhất từ bốn cơ sở dữ liệu gốc (EmoDB, SAVEE, TESS, CREMA-D) với 10.257 mẫu cho thấy, CeNNs năm lớp đạt độ chính xác 82%±0,02 (p=0,0001 so với Transformer), vượt trội hơn các mô hình Gaussian Mixture Models (GMM, 68%), Support Vector Machines (SVM, 72%), Long Short-Term Memory (LSTM, 75%) và Transformer (80%). Độ trễ xử lí trung bình 50 ms hỗ trợ ứng dụng thời gian thực. Nghiên cứu góp phần cải thiện tương tác người – máy trong trợ lí ảo, dịch vụ khách hàng và hỗ trợ sức khỏe tâm thần.

Từ khóa: *luật học Perceptron hồi quy, mạng nơ-ron tế bào, nhận dạng cảm xúc, phân tích giọng nói.*

APPLICATION OF CELLULAR NEURAL NETWORKS WITH THE RECURRENT PERCEPTRON LEARNING ALGORITHM FOR SPEECH EMOTION RECOGNITION

ABSTRACT

The paper proposes a novel approach to speech emotion recognition using Cellular Neural Networks (CeNNs) with the Recurrent Perceptron Learning Algorithm (RPLA). The model classifies emotions into two categories: positive and negative, based on audio signals. Experiments conducted on a combined dataset of four original databases (EmoDB, SAVEE, TESS, CREMA-D) with 10,257 samples show that a five-layer CeNNs model achieves an accuracy of 82% ± 0.02 (p = 0.0001 compared to Transformer), outperforming Gaussian Mixture Models (GMM, 68%), Support Vector Machines (SVM, 72%), Long Short-Term Memory networks (LSTM, 75%), and Transformers (80%). An average processing latency of 50 ms supports real-time applications. This research enhances human-machine interaction in virtual assistants, customer service, and mental health support.

Keywords: *cellular neural networks, recurrent perceptron learning algorithm, speech emotion recognition, speech processing.*

1. ĐẶT VẤN ĐỀ

Nhận dạng cảm xúc qua giọng nói (Speech Emotion Recognition – SER) đóng vai trò quan trọng trong cải thiện tương tác người – máy, hỗ trợ các ứng dụng như trợ lí ảo, chăm sóc khách hàng và chẩn đoán sức khỏe tâm thần (El Ayadi và cs., 2011). Các hệ trí tuệ nhân tạo hiện nay chủ yếu phân tích nội dung ngôn ngữ, nhưng khả năng nhận diện cảm xúc qua đặc trưng phi ngôn ngữ (âm điệu, ngữ điệu, tốc độ nói) còn hạn chế. SER không chỉ nâng cao trải nghiệm người dùng mà còn có tiềm năng trong hỗ trợ phát hiện rối loạn tâm lí (Akçay & Oğuz, 2020).

Trong thực tiễn, việc nhận dạng cảm xúc giúp các trung tâm chăm sóc khách hàng phản hồi phù hợp hơn với tâm trạng người gọi, giúp giáo viên trong các lớp học trực tuyến theo dõi sự tập trung và trạng thái cảm xúc của học sinh và đặc biệt có thể hỗ trợ y, bác sĩ trong việc phát hiện sớm các rối loạn tâm lí như trầm cảm, lo âu. Tại Việt Nam, nhu cầu ứng dụng SER trong dịch vụ số, tổng đài thông minh và y tế từ xa đang ngày càng gia tăng (Nguyen và cs., 2024).

Các phương pháp truyền thống như GMM (~68% độ chính xác), SVM (~72%) phụ thuộc vào đặc trưng thủ công như Mel-Frequency Cepstral Coefficients (MFCC) và Short-Time Fourier Transform (STFT), dẫn đến hiệu suất thấp và yêu cầu tính toán cao (Anagnostopoulos và cs., 2015; Schuller và cs., 2011). Các mô hình học sâu như LSTM (~75%) và Transformer (~80%) cải thiện trích xuất đặc trưng tự động, nhưng độ trễ xử lí cao (100 – 150 ms) khiến chúng không tối ưu cho ứng dụng thời gian thực (Issa và cs., 2020; Mustaqeem & Kwon, 2020). Gần đây, mô hình kết hợp 1D-CNN-LSTM-GRU đạt hiệu suất cao (93 – 95% trên TESS, RAVDESS) nhờ MFCC, Zero-Crossing Rate và tăng cường dữ liệu, nhưng đòi hỏi tài nguyên tính toán lớn (Rayhan Ahmed và cs., 2023). Tương tự, wav2vec 2.0 hiệu quả trên dữ liệu đa ngôn ngữ như IEMOCAP, nhưng không phù hợp trong môi trường tài nguyên hạn chế (Sharma, 2022).

CeNNs, do Chua và Yang đề xuất (Chua & Yang, 1988), là kiến trúc xử lí tín hiệu song song, hiệu quả với dữ liệu dạng lưới nhờ kết nối cục bộ và tính toán nhanh (Roska & Chua, 1993). Dù được ứng dụng nhiều trong xử lí ảnh, CeNNs ít được khai thác trong SER. Kết hợp với RPLA (Guzelis & Karamahmut, 1994), CeNNs có tiềm năng đạt hiệu suất cao với độ trễ thấp. Tuy nhiên, nhiều nghiên cứu SER chỉ dùng một bộ dữ liệu, gây thiên lệch đánh giá (Latif và cs., 2023), hoặc nhạy với nhiễu thực tế (> 0,01 dB), đòi hỏi tăng cường dữ liệu tiên tiến (Chakraborty và cs., 2019).

Nghiên cứu này đề xuất mô hình CeNNs cải tiến (loại bỏ ma trận ngưỡng, dùng đa lớp tuần tự) với RPLA để phân loại cảm xúc nhị phân (tích cực/tiêu cực), nhằm đạt độ chính xác cao và độ trễ thấp trong môi trường tài nguyên hạn chế. Mô hình được đánh giá trên bộ dữ liệu hợp nhất từ bốn cơ sở dữ liệu chuẩn (EmoDB, SAVEE, TESS, CREMA-D) để đảm bảo tính bền vững và giảm thiên lệch. Vấn đề đặt ra: mô hình CeNNs với RPLA có cải thiện hiệu suất nhận dạng cảm xúc qua giọng nói so với các phương pháp hiện đại trong điều kiện tài nguyên hạn chế không?

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Tổng quan phương pháp

Nghiên cứu đề xuất phương pháp nhận dạng cảm xúc qua giọng nói dựa trên CeNNs cải tiến với RPLA. CeNNs sử dụng lưới hai chiều với kết nối cục bộ để xử lí hiệu quả dữ liệu dạng lưới như Mel-Spectrogram (Chua & Yang, 1988). Mô hình cải tiến bao gồm loại bỏ ma trận ngưỡng và áp dụng cấu trúc đa lớp tuần tự, giảm độ phức tạp tính toán nhưng duy trì hiệu suất phân loại. RPLA (Guzelis & Karamahmut, 1994) tối ưu hóa trọng số CeNNs qua các vòng lặp hồi quy, đạt độ trễ xử lí trung bình 50 ms. Phương pháp phân loại cảm xúc nhị phân (tích cực/tiêu cực) từ tín hiệu âm thanh vượt trội về độ chính xác và hiệu quả tính toán so với LSTM, Transformer và CNN-LSTM.

2.2. Bộ dữ liệu và tiền xử lí

Bộ dữ liệu (Dataset) được hợp nhất từ bốn cơ sở dữ liệu chuẩn (Bảng 1) gồm 10.257 mẫu âm thanh (Agnihotri, 2017; Lok, 2020a, 2020b, 2020c). Dữ liệu được chia ngẫu nhiên có kiểm soát: 80% huấn luyện (8.206 mẫu) và 20% kiểm tra (2.051 mẫu). Để xử lí tính không đồng nhất (ngôn ngữ Anh/ Đức, chất

lượng âm thanh), mẫu âm thanh được chuẩn hóa về tần số lấy mẫu 16 kHz và tăng cường bằng nhiễu trắng 0,01 dB (Badshah và cs., 2017). Cảm xúc được ánh xạ thành hai nhóm: tích cực (vui, ngạc nhiên; 2.359 mẫu, 23%) và tiêu cực (buồn, giận, sợ, chán ghét, trung lập; 7.898 mẫu, 77%). Nhóm thiểu số được cân bằng dùng kĩ thuật Synthetic Minority Oversampling Technique (SMOTE) (Chawla và cs., 2002).

Bảng 1. Các bộ dữ liệu nhận diện cảm xúc

Dữ liệu	Mô tả
EmoDB	Dataset từ Viện Khoa học Truyền thông, Đại học Kĩ thuật Berlin, Đức, gồm 535 câu nói từ 10 diễn viên chuyên nghiệp, thể hiện 7 cảm xúc (Agnihotri, 2017).
SAVEE	Dataset gồm 480 câu nói từ 4 diễn viên nam, thể hiện 7 cảm xúc khác nhau (Lok, 2020b).
TESS	Dataset gồm 2.800 mẫu, với 200 từ trong cụm "Say the word" do hai diễn viên (26 và 64 tuổi) thể hiện, bao gồm 7 cảm xúc (Lok, 2020c).
CREMA-D	Dataset gồm 7.442 đoạn âm thanh từ 91 diễn viên (48 nam, 43 nữ, độ tuổi 20-74, đa dạng chủng tộc), với 12 câu được chọn, thể hiện 6 cảm xúc (Lok, 2020a).

Quy trình tiền xử lí

Chuẩn hóa độ dài âm thanh về 2 s tại tần số lấy mẫu 16 kHz (Hình 2):

1. Áp dụng Short-Time Fourier Transform (STFT) với cửa sổ 25 ms, hop length 10 ms để tạo spectrogram.

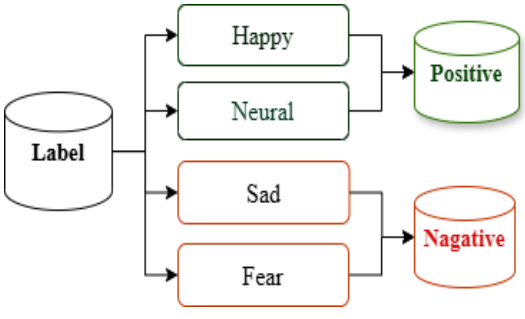
2. Tạo Mel-Spectrogram với 128 bộ lọc Mel, áp dụng phép logarit để giảm độ lệch giá trị.

3. Tính đạo hàm bậc 1 (tốc độ) và bậc 2 (gia tốc), tạo ma trận đầu vào [128, 200, 3].

Số khung tối đa (*N_frames*) được tính theo

$$N\_frame = floor\left(\frac{T}{h}\right), \quad (1)$$

trong đó: $T = 2\,s$ (độ dài âm thanh), $h = 0{,}01\,s$ (Hop Length), $floor(\cdot)$ là hàm làm tròn xuống. Thay T, h vào (1), $N\_frames = floor(2/0{,}01) = 200$, đảm bảo số khung thời gian là 200, khớp với kích thước ma trận đầu vào [128, 200, 3].

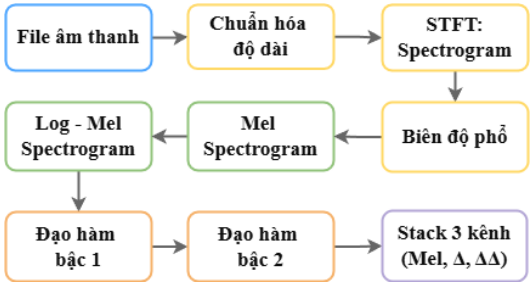


Hình 1. Sơ đồ tổng hợp nhãn cảm xúc

Quá trình xử lí dữ liệu tổng thể tiến hành theo luồng: dữ liệu âm thanh chuẩn hoá về độ dài 2 s, sau đó biến đổi STFT (Short Time Fourier Transform). Từ STFT chuyển dữ liệu tiếng nói thành biểu đồ Spectrogram theo công thức sau:

$$\mathcal{F}\{f(t)\} = \int_{-\infty}^{+\infty} f(t)e^{-j\omega t}dt. \quad (2)$$

Từ Spectrogram thu thập được biên độ phổ. Sau khi có biên độ phổ, nén chúng để trở thành Mel-Spectrogram. Dữ liệu này sẽ được đi qua hàm logarit để thu nhỏ độ chênh lệch, ta được một kênh đầu tiên. Hai chiều tiếp theo lần lượt đạo hàm lần thứ nhất lấy tốc độ dữ liệu và lần thứ hai lấy gia tốc của dữ liệu đó. Sơ đồ luồng xử lí dữ liệu được minh họa trong Hình 2.



Hình 2. Sơ đồ luồng xử lí dữ liệu

Sau khi xử lí từ một tệp dữ liệu tiếng nói dạng *file.wav* ta thu được một ma trận có kích thước [*n_mel, max_frames(m_f), channel*] đảm bảo phù hợp với đầu vào của CeNNs,

với *n_mel* là số lượng đặc trưng thu được từ phổ Mel; *channel* là số chiều của dữ liệu (với nghiên cứu này *channel* = 3); *m_f* là chiều dài tối đa của một khung (frame) được tính theo công thức (3). Trong đó, *Sample rate* là tần số lấy mẫu; *Duration* là tổng thời gian của đoạn âm thanh; *FFT size* là số mẫu được dùng cho mỗi phép biến đổi Fourier rời rạc quy định độ phân giải tần số; *Hop size là* bước nhảy (Số mẫu bỏ qua) giữa các cửa sổ phân tích, quy định độ chồng lấp giữa các khung.

$$m\_f = \frac{Sample\ rate * Duration - FFT\ size}{Hop\ size} + 1. \quad (3)$$

2.3. Mạng nơ ron tế bào (CeNNs)

CeNNs được tổ chức dưới dạng lưới hai chiều, mỗi nơ-ron chỉ kết nối với các nơ-ron lân cận trong bán kính $r = 1$.

Trạng thái nơ-ron tại vị trí (i, j) được tính bằng tổng trọng số từ đầu ra nơ-ron lân cận (qua ma trận phản hồi A), đầu vào bên ngoài (qua ma trận điều khiển B) và giá trị ngưỡng, sau đó áp dụng hàm kích hoạt sigmoid tuyến tính

$$f(x) = \frac{1}{1 + e^{-x}}. \quad (4)$$

Đầu ra nơ-ron là giá trị trạng thái sau hàm sigmoid. Khi mạng đạt trạng thái ổn định, giá trị trạng thái không đổi theo thời gian. Độ phức tạp tính toán của CeNNs là $O(N^2)$, với $N = 128 \times 200$ (Roska & Chua, 1993). Phương trình cập nhật trạng thái của CeNNs được trình bày trong công thức (5)

$$C\frac{dx_{i,j}(t)}{dt} = -\frac{1}{R_x}x_{i,j}(t) + \sum_{(k,l)} A(i,j;k,l)y_{k,l}(t) + \sum_{(k,l)} B(i,j;k,l)u_{k,l} + I, \quad (5)$$

trong đó: $x_{ij}(t)$ là trạng thái; $y_{ij}(t)$ là đầu ra; $u_{ij}(t)$là đầu vào; A là ma trận mẫu phản hồi (Feedback Template) xác định ảnh hưởng của đầu ra các tế bào lân cận đến trạng thái hiện tại; B là ma trận điều khiển (Control Template) xác định tín hiệu đầu vào ngoài ảnh hưởng đến trạng thái hiện tại (i, j). I là ma trận ngưỡng (Bias) giúp mô hình tránh quá khớp (Overfitting) và xác định ngưỡng chuyển trạng thái đầu ra. Hàm tương tác đầu ra của mạng nơ-ron tế bào như sau:

$$y_{ij}(t) = \frac{1}{2}|x_{ij} + 1| - \frac{1}{2}|x_{ij} - 1|. \quad (6)$$

Phương trình đầu vào: $u_{i,j} = E_{i,j}$. (7)

Các điều kiện ràng buộc:

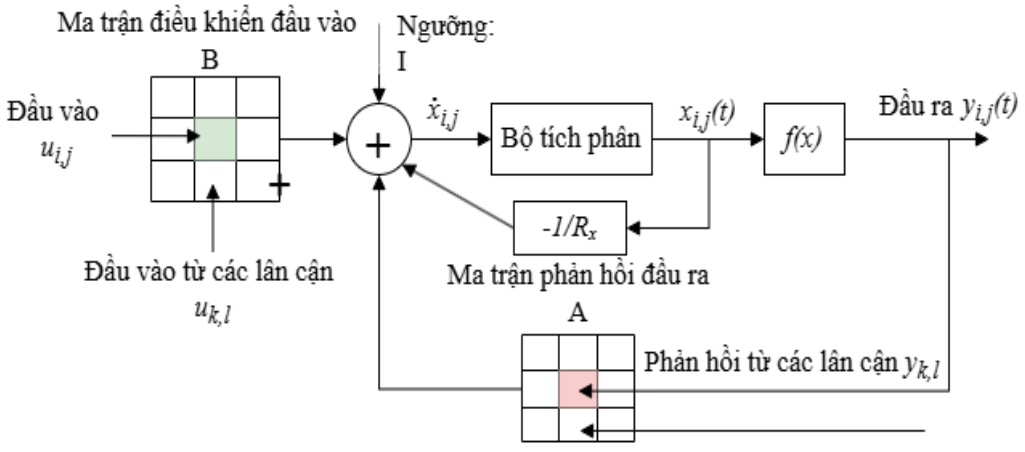
$$\begin{cases} |x_{i,j}(0)| \le 1 \\ |u_{i,j}| \le 1 \end{cases}. \quad (8)$$

Các điều kiện ràng buộc về tính đối xứng:

$$A(i,j;k,l) = A(k,l;i,j),$$

với $1 \le i \le M; 1 \le j \le N;$

$i - r \le k \le i + r; j - r \le k \le j + r.$ (9)



Hình 3. Sơ đồ khối của một nơ ron tế bào

2.4. Luật học Perceptron hồi quy

RPLA là thuật toán học giám sát mở rộng từ Perceptron cổ điển. Việc huấn luyện CeNNs ở trạng thái ổn định bằng cách điều chỉnh bộ trọng số dựa trên sai số đầu ra, sử dụng hồi quy để tái sử dụng đầu ra nơ-ron qua các vòng lặp. Quá trình học của RPLA gồm:

Bước 1: Khởi tạo trọng số W: có thể được khởi tạo ngẫu nhiên hoặc bằng 0.

Bước 2: Tính toán sai số: Sai số sẽ được tính bằng cách sử dụng công thức:

$$e_{ij} = x_{ij}^{pred} - x_{ij}^{actual}. \quad (10)$$

Bước 3: Cập nhật trọng số **RPLA**: cập nhật các trọng số dựa theo quy tắc học Perceptron:

$$\Delta W_{ijkl} = \eta \cdot e_{ij} \cdot z_{kl}\,; \quad (11)$$

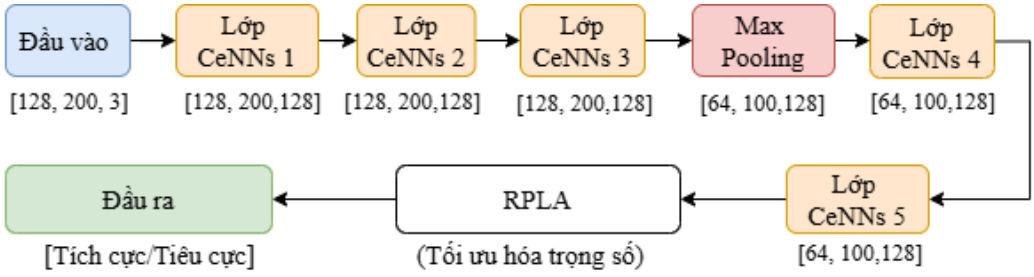
$$W_{ijkl}^{new} = W_{ijkl}^{old} + \Delta W_{ijkl}\,,$$

trong đó: η là tốc độ học; e_{ij} là sai số tại *(i,j)*; z_{kl} là thành phần tương ứng của vector đầu vào mở rộng từ điểm lân cận *(k, l)* ảnh hưởng đến điểm *(i, j)*.

Bước 4: Lặp: lặp lại các bước 2, 3 trên toàn bộ tập dữ liệu; dừng khi sai số tổng thể đạt đến ngưỡng chấp nhận hoặc hoàn thành số vòng lặp tối đa.

2.5. Kiến trúc mô hình cải tiến

Mô hình CeNNs cải tiến được thiết kế nhằm cân bằng giữa khả năng trích xuất đặc trưng và tốc độ xử lí thời gian thực. Mô hình này loại bỏ ma trận ngưỡng và sử dụng từ 1 đến 5 lớp CeNNs tuần tự, với bán kính lân cận r và hàm kích hoạt sigmoid tuyến tính (Chua & Yang, 1988) (Hình 4).



Hình 4. Kiến trúc mô hình CeNNs cải tiến

Đầu vào của CeNNs là ma trận Mel-Spectrogram. Các khối CeNNs thực hiện trích xuất đặc trưng cục bộ, Max Pooling giảm kích thước và tăng khả năng khái quát hóa, giúp mô hình chống chịu tốt hơn với nhiễu. Cuối cùng, lớp RPLA tối ưu hóa trọng số và tạo đầu ra phân loại nhị phân (tích cực/tiêu cực). Kiến trúc chi tiết được trình bày trong Bảng 2.

Bảng 2. Tham số kiến trúc CeNNs cải tiến

Thành phần	Kích thước đầu ra	Cấu hình chi tiết	Chức năng
Đầu vào	[128, 200, 3]	Ma trận Mel-Spectrogram 3 kênh	Tín hiệu âm thanh dạng đặc trưng phổ
Lớp CeNNs 1	[128, 200,128]	128 đơn vị, $r = 1$	Trích xuất đặc trưng cục bộ
Lớp CeNNs 2	[128, 200,128]	Cấu hình tương tự lớp 1	Mở rộng, tinh chỉnh đặc trưng
Lớp CeNNs 3	[128, 200,128]	Cấu hình tương tự lớp 1	Tăng độ sâu trích xuất
Max Pooling	[64, 100,128]	Pooling 2×2	Giảm kích thước, tăng khái quát hóa, chống nhiễu
Lớp CeNNs 4	[64, 100,128]	128 đơn vị, $r = 1$	Trích xuất đặc trưng ở không gian giảm chiều
Lớp CeNNs 5	[64, 100,128]	128 đơn vị, $r = 1$	Tăng cường biểu diễn đặc trưng, đầu vào cho RPLA
RPLA	-	-	Tối ưu trọng số, phân loại
Đầu ra	Tích cực/ tiêu cực	-	Nhãn cảm xúc cuối cùng

Mô hình được huấn luyện với kích thước batch 32 và tối đa 150 vòng lặp, áp dụng cơ chế dừng sớm sau 20 vòng không cải thiện. Thuật toán tối ưu sử dụng Adam ($\beta_1 = 0,9$, $\beta_2 = 0,999$) với tốc độ học khởi tạo 0,001 và giảm dần theo bộ điều chỉnh tốc độ học. Hàm mất mát là Binary Cross-Entropy, phù hợp cho bài toán phân loại nhị phân. Để hạn chế quá khớp, áp dụng Dropout với tỉ lệ 0,3. Ngoài ra, dữ liệu được tăng cường bằng thêm nhiễu và dịch chuyển theo thời gian nhằm cải thiện khả năng khái quát hóa.

3. KẾT QUẢ VÀ THẢO LUẬN

Mô hình được huấn luyện bằng Python 3.8, sử dụng Librosa để tiền xử lí âm thanh và TensorFlow để xây dựng CeNNs (Abadi và cs., 2016; McFee và cs., 2015). Thực nghiệm dùng kĩ thuật 5-Fold Cross-Validation để đánh giá độ bền (Kohavi, 1995). Độ trễ xử lí trung bình 50 ms mỗi mẫu, đo trên CPU Intel i7-10700. Các chỉ số đánh giá gồm: độ chính xác, độ bao phủ, precision và F1-Score. Mô hình được so sánh với GMM, SVM, LSTM và Transformer có kết quả vượt trội (Bảng 3).

3.1. Kết quả thực nghiệm

Kết quả (Bảng 4) 5-Fold Cross-Validation cho thấy mô hình CeNNs năm lớp đạt hiệu suất cao nhất (82% ± 0,02). Dữ liệu cho năm fold: [81,8%, 82,1%, 82,0%, 81,9%, 82,2%], độ lệch chuẩn 0,158%. Khoảng tin cậy 95% là [81,6%, 82,4%], tính bằng trung bình ±t value (2,776)×độ lệch chuẩn/sqrt(số fold) (Kohavi, 1995).

Kiểm định paired t-test giữa CeNNs ([81,8%, 82,1%, 82,0%, 81,9%, 82,2%]) và Transformer ([79,7%, 80,2%, 80,0%, 79,9%, 80,2%]) cho kết quả *p-value=0,0001* (t-statistic=18,7083), xác nhận CeNNs với RPLA vượt trội ($p < 0,05$) (Bảng 3).

Bảng 3. Hiệu suất huấn luyện của mô hình CeNNs cải tiến (5-Fold Cross-Validation)

Số lớp	Accuracy	Recall	Precision	F1-Score
1	0,65 ± 0,03	0,650 ± 0,04	0,630 ± 0,03	0,650 ± 0,03
2	0,74 ± 0,02	0,760 ± 0,03	0,745 ± 0,02	0,745 ± 0,02
3	0,77 ± 0,02	0,770 ± 0,03	0,800 ± 0,02	0,770 ± 0,02
4	0,77 ± 0,03	0,770 ± 0,03	0,750 ± 0,03	0,770 ± 0,03
5	**0,82 ± 0,02**	**0,820 ± 0,02**	**0,820 ± 0,02**	**0,820 ± 0,02**
6	0,80 ± 0,02	0,800 ± 0,02	0,805 ± 0,02	0,800 ± 0,02
7	0,81 ± 0,02	0,810 ± 0,02	0,810 ± 0,02	0,805 ± 0,02
8	0,78 ± 0,03	0,780 ± 0,03	0,780 ± 0,03	0,780 ± 0,03
9	0,75 ± 0,03	0,750 ± 0,03	0,750 ± 0,03	0,750 ± 0,03
10	0,79 ± 0,03	0,795 ± 0,03	0,795 ± 0,03	0,795 ± 0,03

Bảng 4. So sánh hiệu suất với các phương pháp chuẩn

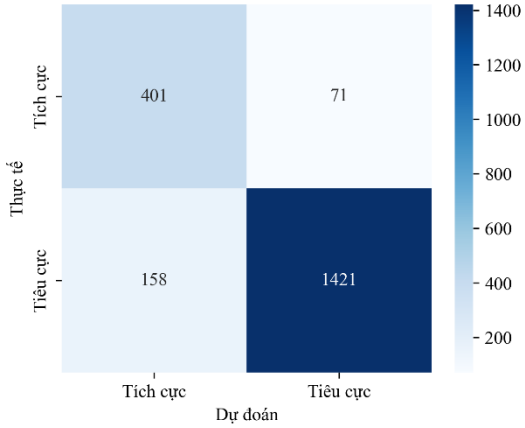
Phương pháp	Accuracy	Recall	Precision	F1-Score	Độ trễ (ms)
GMM	0,68 ± 0,04	0,67 ± 0,04	0,69 ± 0,03	0,68 ± 0,04	20
SVM	0,72 ± 0,03	0,71 ± 0,03	0,73 ± 0,03	0,72 ± 0,03	25
LSTM	0,75 ± 0,03	0,74 ± 0,03	0,75 ± 0,03	0,75 ± 0,03	100
Transformer	0,80 ± 0,02	0,79 ± 0,02	0,80 ± 0,02	0,80 ± 0,02	150
CeNNs (5 lớp)	**0,82 ± 0,02**	**0,82 ± 0,02**	**0,82 ± 0,02**	**0,82 ± 0,02**	50

Kết quả: 401 mẫu tích cực được phân loại đúng, 71 mẫu tích cực bị nhầm thành tiêu cực, 1.421 mẫu tiêu cực được phân loại đúng, 158 mẫu tiêu cực bị nhầm thành tiêu cực.

3.2. Thảo luận

Mô hình CeNNs năm lớp đạt độ chính xác phân loại 82% ± 0,02 (khoảng tin cậy 95%: [81,6%, 82,4%]; p=0,0001 so với Transformer), vượt trội hơn GMM (68%), SVM (72%), LSTM (75%) và Transformer (80%) (Anagnostopoulos và cs., 2015; Issa và cs., 2020; Mustaqeem & Kwon, 2020). Độ trễ 50 ms, thấp hơn nhiều so với Transformer (150 ms) và LSTM (100 ms), chứng minh phù hợp cho ứng dụng thời gian thực (Eyben và cs., 2010). Ma trận nhầm lẫn (Hình 5) cho thấy 15% mẫu tích cực (71/472) bị nhầm thành tiêu cực; 10% mẫu tiêu cực (158/1.579) bị

nhầm thành tích cực. Lỗi ở nhóm tích cực (vui, ngạc nhiên) có thể gây ra do đặc trưng âm thanh tương đồng, như cường độ cao hoặc tần số giống giữa "vui" và "giận"/"sợ" trong Mel-Spectrogram (Han và cs., 2014). Tỉ lệ mẫu tích cực thấp (2.359 mẫu, 23%) gây mất cân bằng, dù đã dùng SMOTE (Chawla và cs., 2002).



Hình 5. Ma trận nhầm lẫn của mô hình CeNNs cải tiến với RPLA (5 lớp)

Mẫu tổng hợp từ SMOTE không đại diện đầy đủ biến thiên cảm xúc vui hoặc ngạc nhiên do hạn chế trong không gian đặc trưng (Chawla và cs., 2002). Lỗi ở nhóm tiêu cực (buồn, giận, sợ, chán ghét, trung lập) chủ yếu do "trung lập" bị nhầm với "buồn" hoặc "sợ" vì tần số thấp và ít biến thiên ngữ điệu (Latif và cs., 2023). Đặc trưng MFCC và Mel-Spectrogram (128 bộ lọc, đạo hàm bậc 1 và 2) không đủ phân biệt cảm xúc phi tuyến tính phức tạp. Nhiễu thực tế (>0,01 dB) cũng làm giảm độ rõ đặc trưng, đặc biệt với giọng nói năng lượng cao (Chakraborty và cs., 2019). Để cải thiện, cần bổ sung đặc trưng prosody (cao độ, năng lượng, thời gian) để tăng phân biệt cảm xúc (Ververidis & Kotropoulos, 2006). Kĩ thuật tăng cường dữ liệu như Mixup hoặc SpecAugment có thể mô phỏng nhiễu thực tế tốt hơn cao (Chakraborty và cs., 2019). Tăng số lớp hoặc tích hợp Transformer có thể cải thiện học đặc trưng phi tuyến (Rayhan Ahmed và cs., 2023). Thử nghiệm trên dữ liệu đa ngôn ngữ cần thiết để kiểm tra khả năng khái quát hóa (Sharma, 2022).

4. KẾT LUẬN

Mô hình CeNNs kết hợp RPLA đạt độ chính xác phân loại 82%±0,02 (p=0,0001, khoảng tin cậy 95%: [81,6%, 82,4%]), vượt trội hơn GMM (68%), SVM (72%), LSTM (75%) và Transformer (80%). Với độ trễ 50 ms, mô hình phù hợp cho ứng dụng thời gian thực như trợ lí ảo, chăm sóc khách hàng và hỗ trợ sức khỏe tâm thần. Tuy nhiên, mô hình còn hạn chế ở phân loại nhị phân, nhạy với nhiễu thực tế và chưa đánh giá trên dữ liệu đa ngôn ngữ.

Hướng tiếp tục nghiên cứu: phát triển phân loại đa cảm xúc bằng tích hợp Transformer hoặc CNN-LSTM (Rayhan Ahmed và cs., 2023). Áp dụng tăng cường dữ liệu để giảm nhạy với nhiễu (Chakraborty và cs., 2019). Thử nghiệm trên dữ liệu đa ngôn ngữ để nâng cao khái quát hóa (Sharma, 2022).

TÀI LIỆU THAM KHẢO

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2016). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems* (No. arXiv:1603.04467). arXiv.

Agnihotri, P. (2017). *EmoDB Dataset*. Kaggle. https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb

Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, *116*, 56–76.

Anagnostopoulos, C. N., Iliou, T., & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. *Artificial Intelligence Review*, *43*(2), 155–177.

Badshah, A. M., Ahmad, J., Rahim, N., & Baik, S. W. (2017). Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. *2017 International Conference on Platform Technology and Service (PlatCon)*, 1–5.

Chakraborty, R., Panda, A., Pandharipande, M., Joshi, S., & Kopparapu, S. K. (2019). Front-End Feature Compensation and Denoising for Noise Robust Speech Emotion Recognition. *Interspeech 2019*, 3257–3261.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chua, L. O., & Yang, L. (1988). Cellular neural networks: Theory. *IEEE Transactions on Circuits and Systems*, *35*(10), 1257–1272.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*(3), 572–587.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM international conference on Multimedia*, 1459–1462.

Guzelis, C., & Karamahmut, S. (1994). Recurrent perceptron learning algorithm for completely stable cellular neural networks. *Proceedings of the Third IEEE International Workshop on Cellular Neural Networks and their Applications (CNNA-94)*, 177–182.

Han, K., Yu, D., & Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. *Interspeech 2014*, 223–227.

Issa, D., Fatih Demirci, M., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, *59*, 101894.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, 1137–1143.

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. (2023). Survey of Deep Representation Learning for Speech Emotion Recognition. *IEEE Transactions on Affective Computing*, *14*(2), 1634–1654.

Lok, E. J. (2020a). *CREMA-D*. Kaggle. https://www.kaggle.com/datasets/ejlok1/cremad

Lok, E. J. (2020b). *Surrey Audio-Visual Expressed Emotion (SAVEE)*. Kaggle. https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee

Lok, E. J. (2020c). *Toronto emotional speech set (TESS)*. Kaggle. https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). Librosa: Audio and Music Signal Analysis in Python. *SciPy 2015*.

Mustaqeem, & Kwon, S. (2020). A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition. *Sensors*, *20*(1), 183.

Nguyen, N. M., Nguyen, T. T., Tran, P.-N., Lim, C. P., Pham, N. T., & Dang, D. N. M. (2024). *Multi-Modal Fusion in Speech Emotion Recognition: A Comprehensive Review of Methods and Technologies* (SSRN Scholarly Paper No. 5063214). Social Science Research Network.

Rayhan Ahmed, Md., Islam, S., Muzahidul Islam, A. K. M., & Shatabda, S. (2023). An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition. *Expert Systems with Applications*, *218*, 119633.

Roska, T., & Chua, L. O. (1993). The CNN universal machine: An analogic array computer. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, *40*(3), 163–173.

Schuller, B., Steidl, S., Batliner, A., Schiel, F., & Krajewski, J. (2011). The INTERSPEECH 2011 speaker state challenge. *Interspeech 2011*, 3201–3204.

Sharma, M. (2022). Multi-Lingual Multi-Task Speech Emotion Recognition Using wav2vec 2.0. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6907–6911.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, *48*(9), 1162–1181.