

BÁO CÁO CUỐI KỲ - PROJECT DATA SCIENCE

Nhận diện Mức độ Tập trung và Cảm xúc của Người học Trực tuyến
sử dụng Mô hình Ngôn ngữ-Thị giác Lớn (VLM)

Nhóm: 33

Thành viên:

1. Nguyễn Hoàng Lâm - 20225028
2. Đặng Thanh Tùng - 20225111

Giảng viên hướng dẫn: PGS. TS. Phạm Văn Hải

Ngày 13 tháng 1 năm 2026

Mục lục

TÓM TẮT	4
1 GIỚI THIỆU (INTRODUCTION)	4
1.1 Dặt vấn đề	4
1.2 Mục tiêu nghiên cứu	4
1.3 Cấu trúc báo cáo	4
2 PHƯƠNG PHÁP ĐỀ XUẤT (METHODOLOGY)	5
2.1 Mô hình nền tảng (Backbone): Qwen2.5-VL	5
2.2 Bộ phân loại (Classifier)	5
2.3 Tăng cường dữ liệu (Data Augmentation)	5
2.4 Hàm mất mát (Loss Function)	5
2.5 Siêu tham số (Hyperparameters)	6
3 THỰC NGHIỆM VÀ KẾT QUẢ (EXPERIMENTS & RESULTS)	6
3.1 Liên kết đến mã nguồn và dữ liệu	6
3.2 Thiết lập thực nghiệm	6
3.2.1 Dữ liệu	6
3.2.2 Độ đo đánh giá (Evaluation Metrics)	8
3.3 Kết quả Baseline: Qwen2.5-VL đầy đủ	8
3.4 Kết quả sau khi huấn luyện MLP	9
3.5 Phân tích kết quả	9
3.5.1 So sánh Baseline và MLP	9
3.5.2 Hiệu quả của Augmentation	9
3.5.3 Vấn đề mất cân bằng	9
3.5.4 So sánh Dropout	10
3.6 Kết quả chi tiết theo từng cảm xúc	10
3.6.1 Frustration	10
3.6.2 Confusion	15
3.6.3 Engagement	20
3.6.4 Boredom	25
4 KẾT QUẢ ĐẠT ĐƯỢC (DEPLOYMENT RESULTS)	30
4.1 Triển khai ứng dụng thực tế	30
4.1.1 Kiến trúc triển khai	30
4.1.2 Luồng xử lý	30
4.2 Giao diện ứng dụng	31
4.3 Hiệu năng triển khai	32
4.4 Lợi ích của việc triển khai	32
4.5 Thách thức và giải pháp	32
4.6 Hướng phát triển cho deployment	33
5 KẾT LUẬN (CONCLUSION)	33
5.1 Kết luận	33
5.2 Hướng phát triển	34
5.3 Dóng góp của đề tài	34
5.3.1 Dóng góp về phương pháp	34

5.3.2	Dóng góp về kết quả thực nghiệm	35
5.3.3	Dóng góp về dữ liệu và tài nguyên	35
5.3.4	Dóng góp về insights và hướng nghiên cứu	35

TÓM TẮT (ABSTRACT)

Báo cáo này trình bày kết quả nghiên cứu và thực nghiệm về việc nhận diện trạng thái cảm xúc của người học (Boredom, Confusion, Engagement, Frustration) trong môi trường học tập trực tuyến. Sử dụng tập dữ liệu chuẩn DAiSEE, nhóm nghiên cứu đề xuất phương pháp tiếp cận mới dựa trên Mô hình Ngôn ngữ-Thị giác Lớn (Vision-Language Model - VLM), cụ thể là Qwen2.5-VL-7B-Instruct, để trích xuất đặc trưng ngữ nghĩa mức cao từ video người học. Các đặc trưng này sau đó được sử dụng để huấn luyện một bộ phân lớp MLP (Multi-Layer Perceptron). Bên cạnh đó, nhóm cũng áp dụng kỹ thuật tăng cường dữ liệu sử dụng các bộ dữ liệu khuôn mặt tĩnh (Facial Expression Data) để khắc phục vấn đề mất cân bằng dữ liệu nghiêm trọng của DAiSEE. Kết quả thực nghiệm cho thấy phương pháp đề xuất đạt được độ chính xác khả quan trên một số lớp, tuy nhiên vẫn còn thách thức lớn đối với các lớp thiểu số do bản chất mất cân bằng của dữ liệu. Báo cáo phân tích chi tiết hiệu năng của các cấu hình mô hình khác nhau và đề xuất các hướng cải tiến trong tương lai.

Từ khóa: *Engagement Recognition, Vision-Language Models, Qwen2.5-VL, DAiSEE, Imbalanced Learning.*

1 GIỚI THIỆU (INTRODUCTION)

1.1 Đặt vấn đề

Học tập trực tuyến (E-learning) đã trở thành một phần không thể thiếu của giáo dục hiện đại. Tuy nhiên, một trong những thách thức lớn nhất của hình thức này là thiếu sự tương tác trực tiếp, khiến giáo viên khó nắm bắt được trạng thái cảm xúc và mức độ tập trung của học viên. Việc tự động nhận diện các trạng thái như "Chán nản" (Boredom), "Bối rối" (Confusion) hay "Tập trung" (Engagement) đóng vai trò quan trọng trong việc xây dựng các hệ thống học tập thông minh, có khả năng thích ứng và phản hồi kịp thời để nâng cao hiệu quả giảng dạy.

Gần đây, nghiên cứu của Wang et al. [1] đã chứng minh tiềm năng của các mô hình ngôn ngữ-thị giác (Vision-Language Models - VLM) trong việc phát hiện cảm xúc học tập của sinh viên thông qua biểu cảm khuôn mặt. Được truyền cảm hứng từ công trình này, nhóm quyết định khai thác khả năng của VLM tiên tiến, cụ thể là Qwen2.5-VL, để giải quyết bài toán nhận diện trạng thái cảm xúc trên tập dữ liệu DAiSEE.

1.2 Mục tiêu nghiên cứu

Mục tiêu của đề tài là xây dựng một mô hình học sâu có khả năng phân loại 4 trạng thái cảm xúc của người học dựa trên video ghi hình khuôn mặt, sử dụng tập dữ liệu DAiSEE. Nhóm tập trung khai thác sức mạnh của các mô hình nền tăng đa phương thức (VLM) tiên tiến thay vì các phương pháp CNN truyền thống.

1.3 Cấu trúc báo cáo

Báo cáo được chia thành 5 phần chính: Phần I giới thiệu bài toán và động lực nghiên cứu; Phần II mô tả phương pháp đề xuất; Phần III trình bày thiết lập thực nghiệm và phân tích kết quả chi tiết; Phần IV kết luận và hướng phát triển; và cuối cùng là phần Tài liệu tham khảo.

2 PHƯƠNG PHÁP ĐỀ XUẤT (METHODOLOGY)

Mô hình tổng thể bao gồm hai giai đoạn chính: Trích xuất đặc trưng (Feature Extraction) và Phân loại (Classification).

2.1 Mô hình nền tảng (Backbone): Qwen2.5-VL

Nhóm sử dụng Qwen2.5-VL-7B-Instruct làm backbone. Đây là mô hình đa phương thức mạnh mẽ, được huấn luyện trên lượng dữ liệu khổng lồ, có khả năng chiết xuất các đặc trưng thị giác phong phú và có ý nghĩa ngữ nghĩa cao hơn so với các CNN thông thường.

- **Input:** Video đầu vào được lấy mẫu (sample) với tốc độ **1 FPS**.
- **Feature Extraction:** Mỗi khung hình được đưa qua Qwen2.5-VL để lấy vector embedding (kích thước lớn).
- **Temporal Pooling:** Các vector đặc trưng của các khung hình trong cùng một clip được gộp (Mean Pooling) để tạo thành một vector đại diện duy nhất cho video đó.

2.2 Bộ phân loại (Classifier)

Vector đặc trưng sau khi gộp được đưa vào một mạng MLP đơn giản để phân loại:

- **Kiến trúc:** Linear(Input_Dim → 512) → ReLU → Dropout → Linear(512 → 256) → ReLU → Dropout → Linear(256 → 4).
- **Dropout:** Được sử dụng để tránh overfitting, với các tỷ lệ thử nghiệm là 0, 0.3, và 0.5.

2.3 Tăng cường dữ liệu (Data Augmentation)

Do tập DAiSEE có sự mất cân bằng dữ liệu rất lớn (lớp "Engagement" chiếm đa số, "Confusion"/"Boredom" rất ít), nhóm đã thực hiện tăng cường dữ liệu bằng cách sử dụng thêm các tập dữ liệu khuôn mặt tĩnh (Facial Expression Data) từ các nguồn bên ngoài (như FER-2013 hoặc Mendeley Data). Các ảnh tĩnh này được gán nhãn tương ứng với 4 trạng thái của DAiSEE để bổ sung mẫu cho quá trình huấn luyện.

2.4 Hàm mất mát (Loss Function)

Sử dụng Cross-Entropy Loss tiêu chuẩn:

$$L = -\frac{1}{N} \sum_{i=1}^N \sum_{c=0}^{M-1} y_{i,c} \log(p_{i,c})$$

Trong đó N là số mẫu, $M = 4$ là số lớp.

2.5 Siêu tham số (Hyperparameters)

Các tham số cấu hình cho quá trình huấn luyện:

- **Model Name:** Multi-Layer Perceptron (MLP)
- **Hidden Dimension:** 512
- **Batch Size:** 32
- **Learning Rate:** 1×10^{-3} (0.001)
- **Optimizer:** Adam
- **Scheduler:** ReduceLROnPlateau (Patience=5, Factor=0.5)
- **Epochs:** 50
- **Dropout:** Thay đổi [0, 0.3, 0.5] để tìm cấu hình tối ưu

3 THỰC NGHIỆM VÀ KẾT QUẢ (EXPERIMENTS & RESULTS)

3.1 Liên kết đến mã nguồn và dữ liệu

- Mã nguồn (GitHub): [2]
<https://github.com/nhlam04/data-science-group20>
 - Kết quả thực nghiệm trong thư mục *results/* và *plots/*
- Mô hình triển khai (HuggingFace): [3]
<https://huggingface.co/mapotofu40/qwen-mlp>
- Tập dữ liệu:
 - DAiSEE: [4] <https://people.iith.ac.in/vineethnb/resources/daisee/>
 - Facial Expression Data: [5] <https://data.mendeley.com/datasets/6dbdkb8g3d/>

3.2 Thiết lập thực nghiệm

3.2.1 Dữ liệu

Tập dữ liệu DAiSEE [6] Nhóm sử dụng tập dữ liệu DAiSEE với phân chia chuẩn như sau:

- **Train:** 5,358 samples
- **Validation:** 1,429 samples
- **Test:** 1,784 samples
- **Total:** 8,571 samples

Phân phối lớp trong DAiSEE Dữ liệu DAiSEE có sự mất cân bằng nghiêm trọng giữa các mức độ cảm xúc. Bảng 1 trình bày phân phối chi tiết cho từng trạng thái cảm xúc trong tập huấn luyện.

Bảng 1: Phân phối lớp trong tập huấn luyện DAiSEE (5,358 samples)

Emotion	Level 0	Level 1	Level 2	Level 3
Boredom	2,433 (45.41%)	1,696 (31.65%)	1,073 (20.03%)	156 (2.91%)
Engagement	34 (0.63%)	213 (3.98%)	2,617 (48.84%)	2,494 (46.55%)
Confusion	3,616 (67.49%)	1,245 (23.24%)	431 (8.04%)	66 (1.23%)
Frustration	4,183 (78.07%)	941 (17.56%)	191 (3.56%)	43 (0.80%)

Như có thể thấy từ bảng trên:

- **Boredom:** Tương đối cân bằng hơn, nhưng Level 3 chỉ chiếm 2.91%.
- **Engagement:** Tập trung chủ yếu ở Level 2 và 3, trong khi Level 0 và 1 rất hiếm (< 5%).
- **Confusion:** Mất cân bằng nghiêm trọng với 67.49% là Level 0, Level 3 chỉ 1.23%.
- **Frustration:** Mất cân bằng nghiêm trọng nhất với 78.07% là Level 0, Level 3 chỉ 0.80%.

Dữ liệu tăng cường (Facial Expression Data) Để khắc phục vấn đề mất cân bằng, nhóm đã sử dụng thêm dữ liệu khuôn mặt tĩnh từ Mendeley Facial Expression Dataset:

Bảng 2: Dữ liệu tăng cường từ Facial Expression Dataset

Emotion	Source File	Samples
Boredom	boring.csv	1,931
Engagement	happiness.csv	593
Confusion	confused.csv	1,177
Frustration	surprise.csv	219
Total		3,920

Các ảnh khuôn mặt tĩnh này (kích thước $256 \times 256 \times 3 = 196,608$ pixels) được xử lý thông qua Qwen2.5-VL để trích xuất đặc trưng, sau đó được gán nhãn tương ứng với các mức độ cảm xúc của DAiSEE để bổ sung cho tập huấn luyện.

Môi trường thực nghiệm

- **Platform:** Google Colab / Kaggle T4x2 hoặc Local GPU
- **Framework:** PyTorch
- **Feature Extraction:** Qwen2.5-VL-7B-Instruct với sampling rate 1 FPS

3.2.2 Độ đo đánh giá (Evaluation Metrics)

Để đánh giá hiệu năng mô hình phân loại cảm xúc trên tập dữ liệu DAiSEE, nhóm sử dụng các độ đo sau:

1. **Accuracy (Độ chính xác):** Tỷ lệ mẫu được phân loại đúng.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

2. **Precision (Độ chính xác):** Tỷ lệ mẫu thực sự thuộc lớp dương trong số các mẫu được dự đoán là dương.
3. **Recall (Độ phủ):** Tỷ lệ mẫu dương được phát hiện đúng trên tổng số mẫu dương thực tế.
4. **F1-Score (Macro):** Trung bình điều hòa của Precision và Recall. Nhóm sử dụng **Macro F1** để đánh giá công bằng giữa các lớp, tránh mô hình chỉ học tốt trên lớp đa số mà bỏ qua lớp thiểu số.

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

Do dữ liệu mất cân bằng, **Accuracy** không phản ánh hiệu năng. Nhóm sử dụng thêm **Precision, Recall, và F1-Score (Macro Average)** để đánh giá công bằng hơn.

3.3 Kết quả Baseline: Qwen2.5-VL đầy đủ

Trước khi huấn luyện bộ phân loại MLP, nhóm đã thử nghiệm với mô hình Qwen2.5-VL đầy đủ (full model) trực tiếp trên tập dữ liệu DAiSEE để đánh giá khả năng zero-shot/few-shot của mô hình. Bảng 3 trình bày kết quả baseline này.

Bảng 3: Kết quả Baseline của mô hình Qwen2.5-VL đầy đủ (trước fine-tuning MLP)

Category	Accuracy	Precision	Recall	F1 Score (Macro)	F1 Score (Weighted)
Boredom	32.91%	37.76%	25.88%	14.53%	18.62%
Engagement	51.89%	38.67%	26.46%	19.97%	36.45%
Confusion	69.47%	29.19%	25.43%	21.54%	57.75%
Frustration	78.08%	19.52%	25.00%	21.92%	68.47%

Kết quả baseline cho thấy mô hình Qwen2.5-VL đầy đủ đạt hiệu năng khiêm tốn với F1-Score (Macro) dao động từ 14.53% (Boredom) đến 21.92% (Frustration). Điều này cho thấy cần thiết phải fine-tune thêm một bộ phân loại chuyên biệt (MLP) trên đặc trưng trích xuất từ Qwen2.5-VL để cải thiện hiệu năng cho bài toán cụ thể này.

3.4 Kết quả sau khi huấn luyện MLP

Sau khi huấn luyện bộ phân loại MLP trên đặc trưng từ Qwen2.5-VL, nhóm đã thử nghiệm nhiều cấu hình khác nhau. Bảng 4 tổng hợp kết quả trên tập Test cho các cấu hình tốt nhất của từng trạng thái cảm xúc (dựa trên F1-Score cao nhất).

Bảng 4: Kết quả tổng hợp của các cấu hình tốt nhất cho từng trạng thái

Trạng thái	Cấu hình	Accuracy	Precision	Recall	F1-Score
Boredom	True_512_0.3 (Ep.16)	39.87%	0.3004	0.3010	0.2873
Engagement	True_512_0 (Ep.42)	49.88%	0.3013	0.2684	0.2680
Confusion	True_512_0 (Ep.31)	69.29%	0.4062	0.2671	0.2456
Frustration	True_512_0 (Ep.42)	76.98%	0.2735	0.2528	0.2309

Chú thích cấu hình: [Facial_Augment]_[HiddenDim]_[Dropout]

(Ví dụ: True_512_0.3 nghĩa là có Augmentation, Hidden Dim 512, Dropout 0.3).

3.5 Phân tích kết quả

3.5.1 So sánh Baseline và MLP

So với kết quả baseline (mô hình Qwen2.5-VL đầy đủ), việc huấn luyện thêm bộ phân loại MLP đã mang lại sự cải thiện đáng kể:

- **Boredom:** F1-Score (Macro) tăng từ 14.53% lên **28.73%** (tăng 97.7%), cho thấy MLP đã học được cách phân biệt tốt hơn các mức độ chán nản.
- **Engagement:** F1-Score (Macro) tăng từ 19.97% lên **26.80%** (tăng 34.2%).
- **Confusion:** F1-Score (Macro) tăng từ 21.54% lên **24.56%** (tăng 14.0%).
- **Frustration:** F1-Score (Macro) tăng từ 21.92% lên **23.09%** (tăng 5.3%).

Kết quả này chứng minh rằng việc sử dụng Qwen2.5-VL làm feature extractor kết hợp với MLP classifier là một chiến lược hiệu quả cho bài toán nhận diện cảm xúc học tập.

3.5.2 Hiệu quả của Augmentation

Hầu hết các kết quả tốt nhất đều đến từ cấu hình có sử dụng dữ liệu tăng cường (True), cho thấy việc bổ sung dữ liệu ngoại lai giúp cải thiện khả năng tổng quát hóa của mô hình.

3.5.3 Vấn đề mất cân bằng

- **Frustration:** Có Accuracy rất cao (77-78%) nhưng F1-Score thấp (0.21-0.23). Nhìn vào Confusion Matrix (trong logs), mô hình dự đoán hầu hết là lớp 0 (không bực bội), dẫn đến Recall của lớp dương (lớp 1,2,3) rất thấp.
- **Boredom:** Có Accuracy thấp nhất (~40%) nhưng F1-Score lại cao nhất (0.28). Điều này cho thấy mô hình đã "dám" dự đoán các lớp thiểu số nhiều hơn, chấp nhận Accuracy thấp để đổi lấy khả năng phát hiện đúng lớp hiếm tốt hơn.

3.5.4 So sánh Dropout

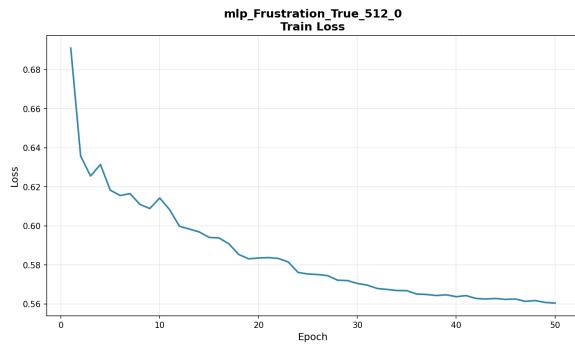
Dropout thấp (0 hoặc 0.3) thường cho kết quả tốt hơn Dropout cao (0.5), có thể do mô hình (MLP head) chưa đủ sâu để cần regularization mạnh, hoặc đặc trưng từ Qwen2.5 đã đủ tốt.

3.6 Kết quả chi tiết theo từng cảm xúc

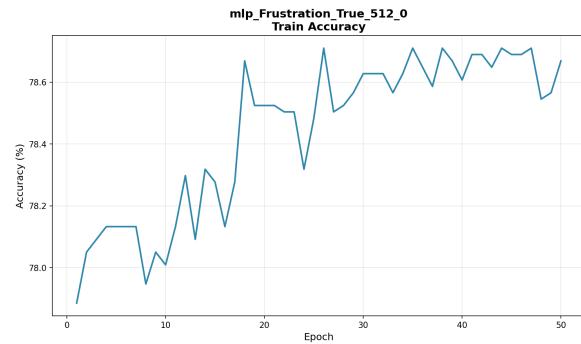
3.6.1 Frustration

Cấu hình: Facial Data = True/False, Hidden Dim = 512, Dropout = [0, 0.3, 0.5]

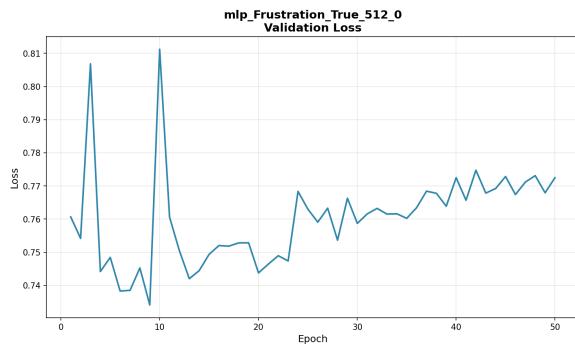
mlp_Frustration_True_512_0 Kết quả: Epoch 42, Val F1: 0.2275, Test Acc: 76.98%, Precision: 0.2735, Recall: 0.2528, F1: 0.2309



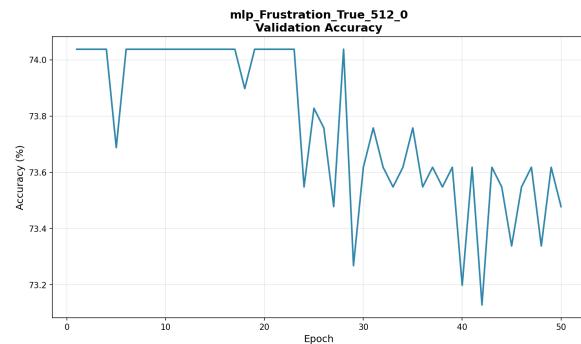
(a) Train Loss



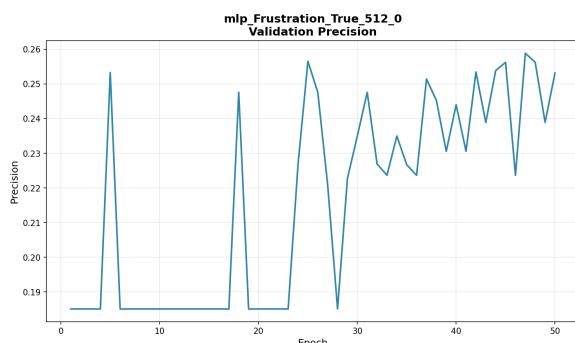
(b) Train Accuracy



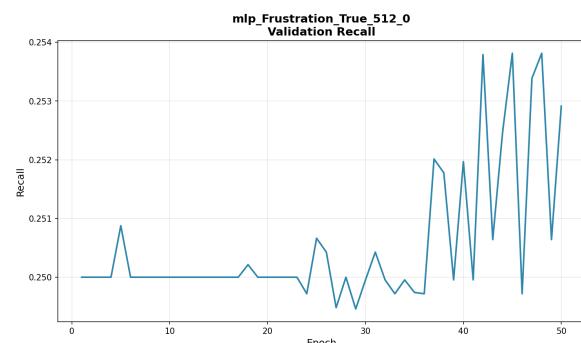
(c) Validation Loss



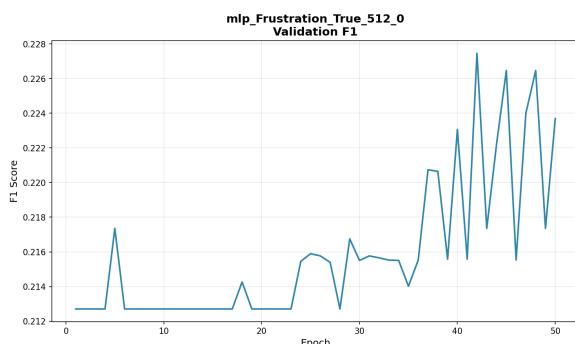
(d) Validation Accuracy



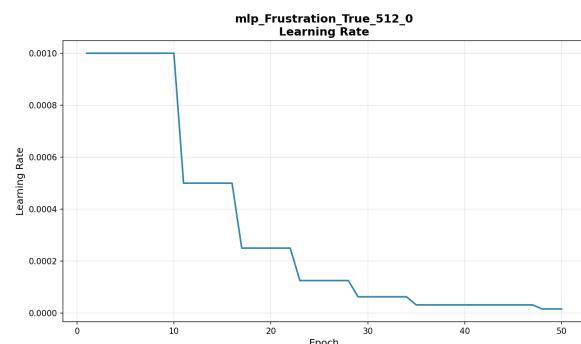
(e) Validation Precision



(f) Validation Recall



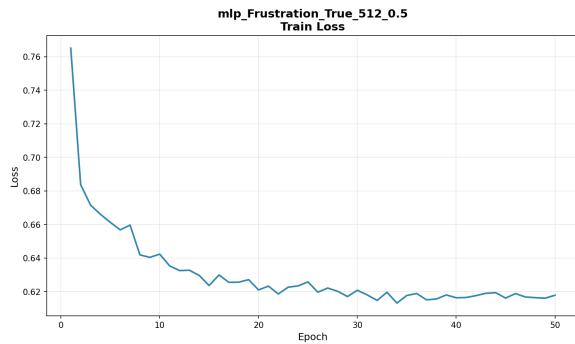
(g) Validation F1



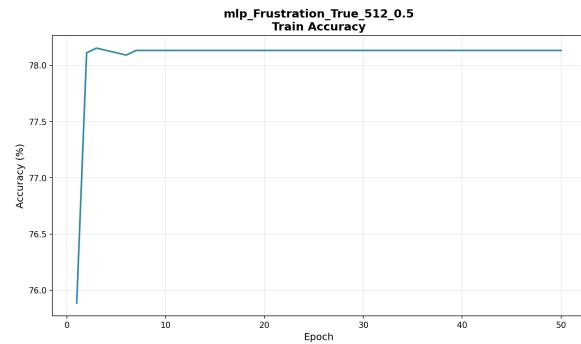
(h) Learning Rate

Hình 1: Đường học: mlp_Frustration_True_512_0

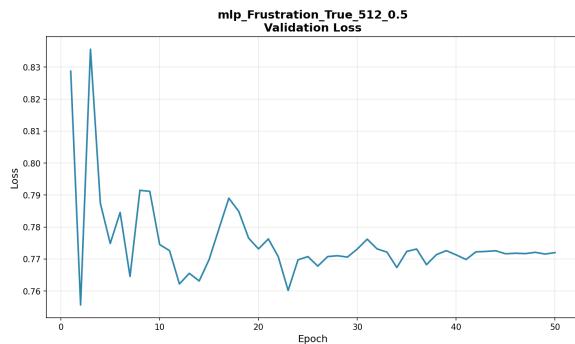
mlp_Frustration_True_512_0.5 Kết quả: Epoch 1, Val F1: 0.2127, Test Acc: 78.08%, Precision: 0.1952, Recall: 0.2500, F1: 0.2192



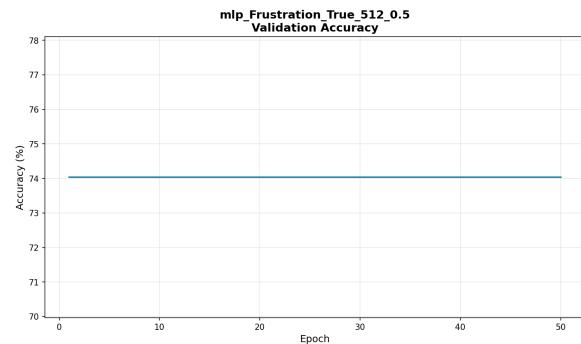
(a) Train Loss



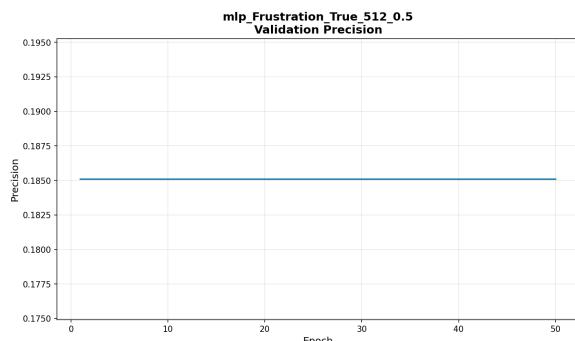
(b) Train Accuracy



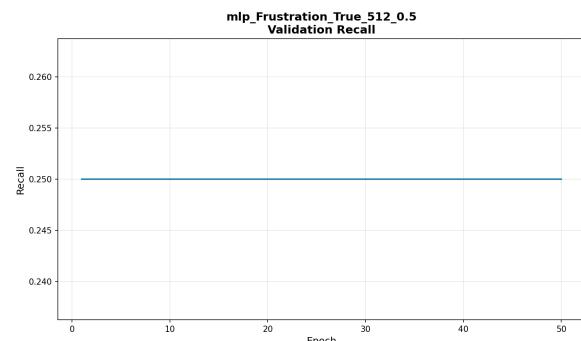
(c) Validation Loss



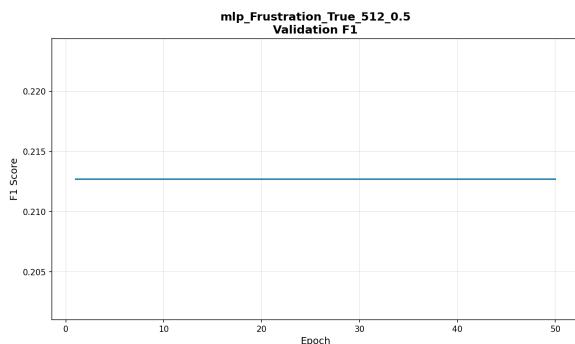
(d) Validation Accuracy



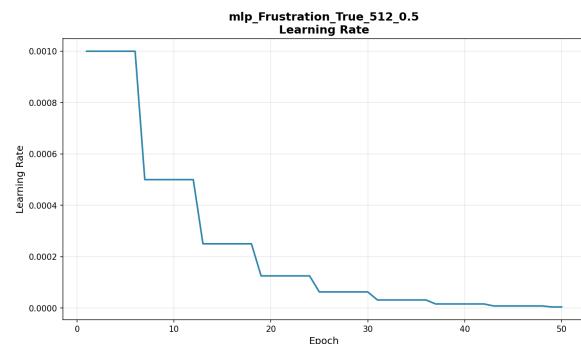
(e) Validation Precision



(f) Validation Recall



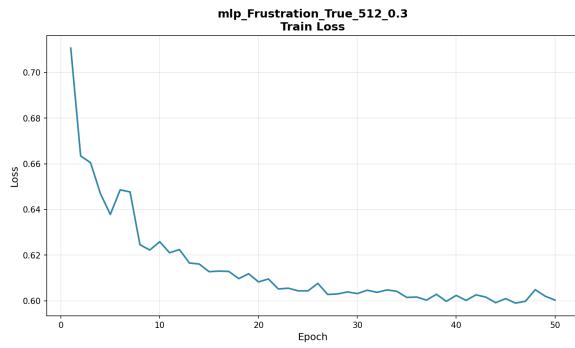
(g) Validation F1



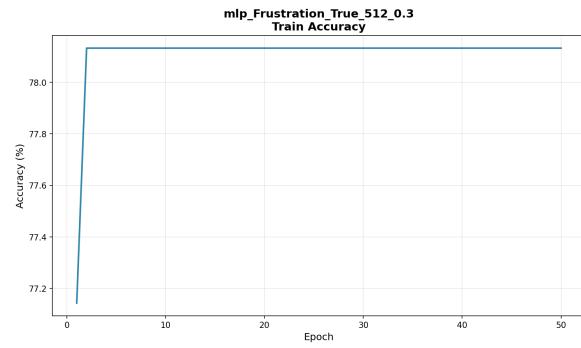
(h) Learning Rate

Hình 2: Đường học: mlp_Frustration_True_512_0.5

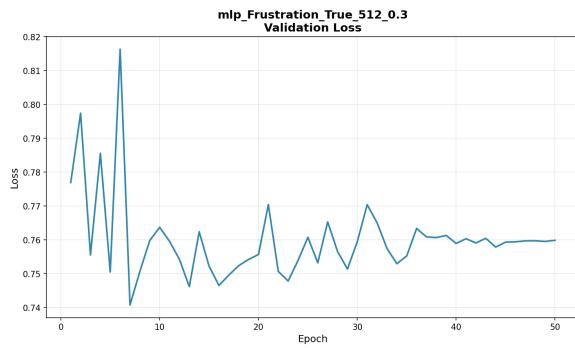
mlp_Frustration_True_512_0.3 Kết quả: Epoch 1, Val F1: 0.2127, Test Acc: 78.08%, Precision: 0.1952, Recall: 0.2500, F1: 0.2192



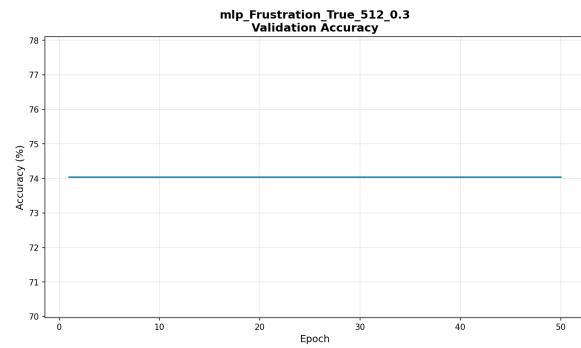
(a) Train Loss



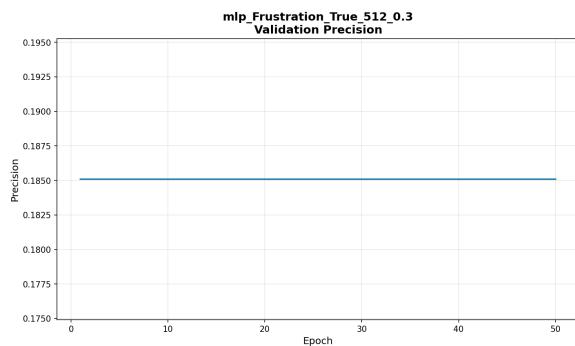
(b) Train Accuracy



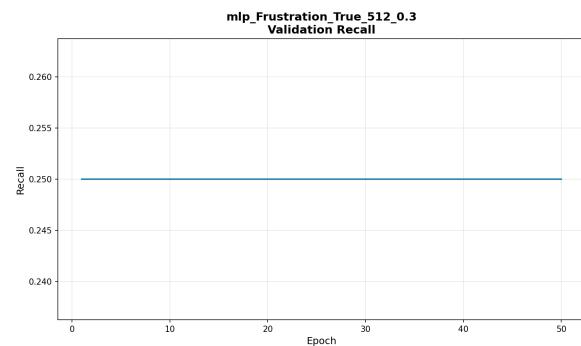
(c) Validation Loss



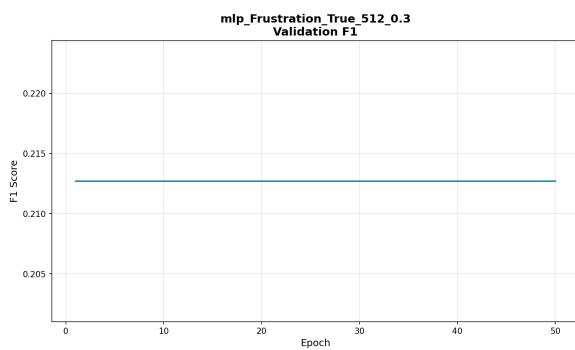
(d) Validation Accuracy



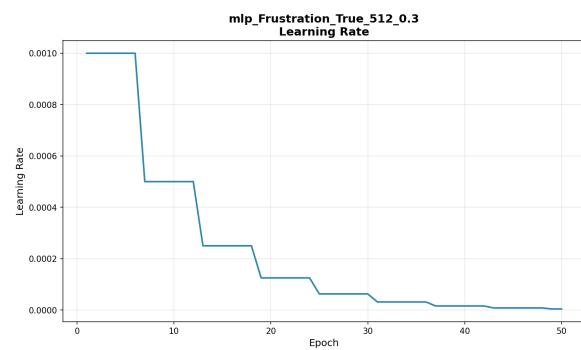
(e) Validation Precision



(f) Validation Recall



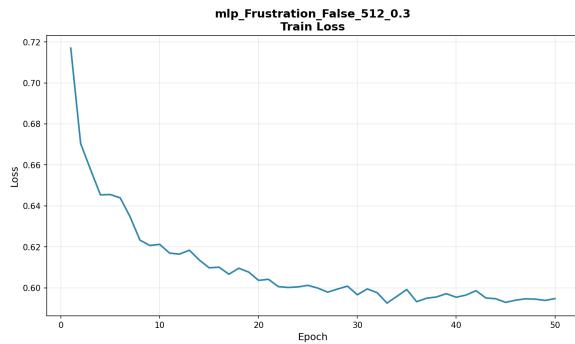
(g) Validation F1



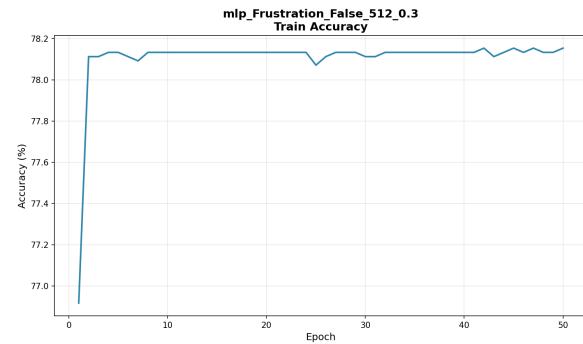
(h) Learning Rate

Hình 3: Đường học: mlp_Frustration_True_512_0.3

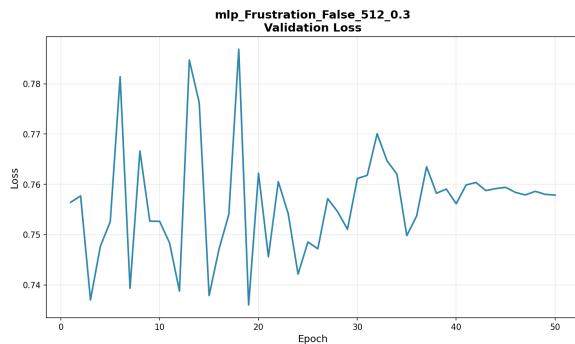
mlp_Frustration_False_512_0.3 Kết quả: Epoch 1, Val F1: 0.2127, Test Acc: 78.08%, Precision: 0.1952, Recall: 0.2500, F1: 0.2192



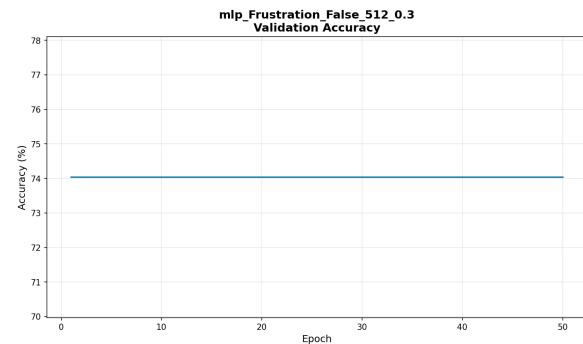
(a) Train Loss



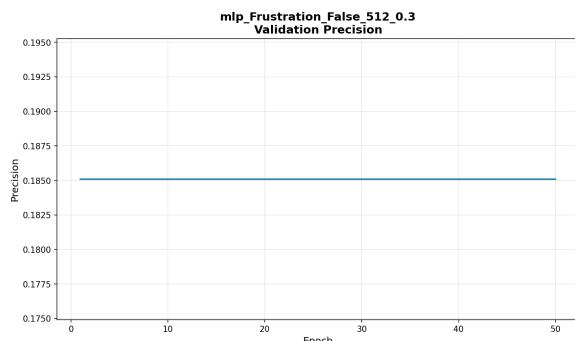
(b) Train Accuracy



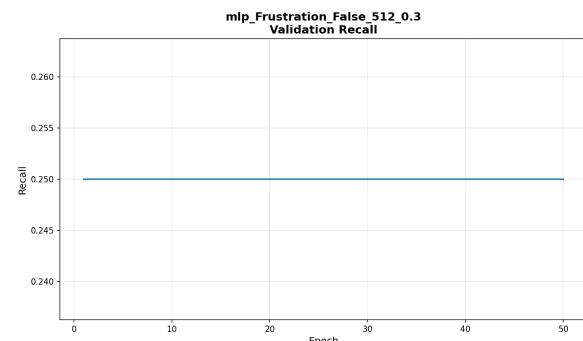
(c) Validation Loss



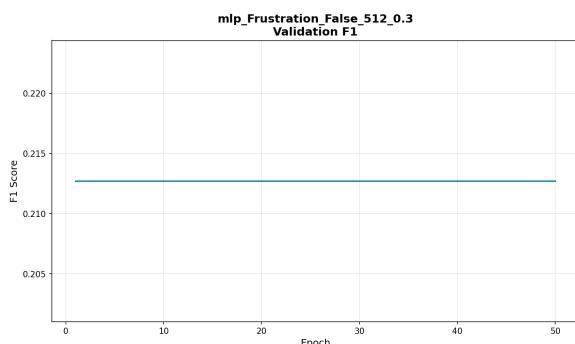
(d) Validation Accuracy



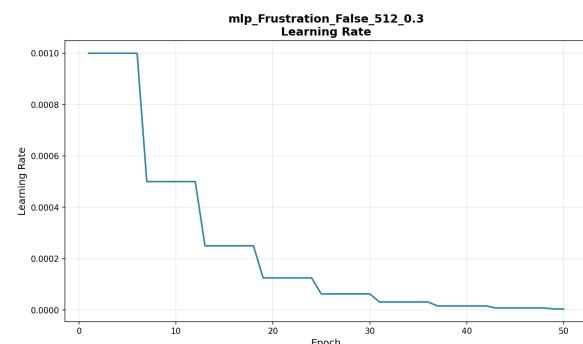
(e) Validation Precision



(f) Validation Recall



(g) Validation F1



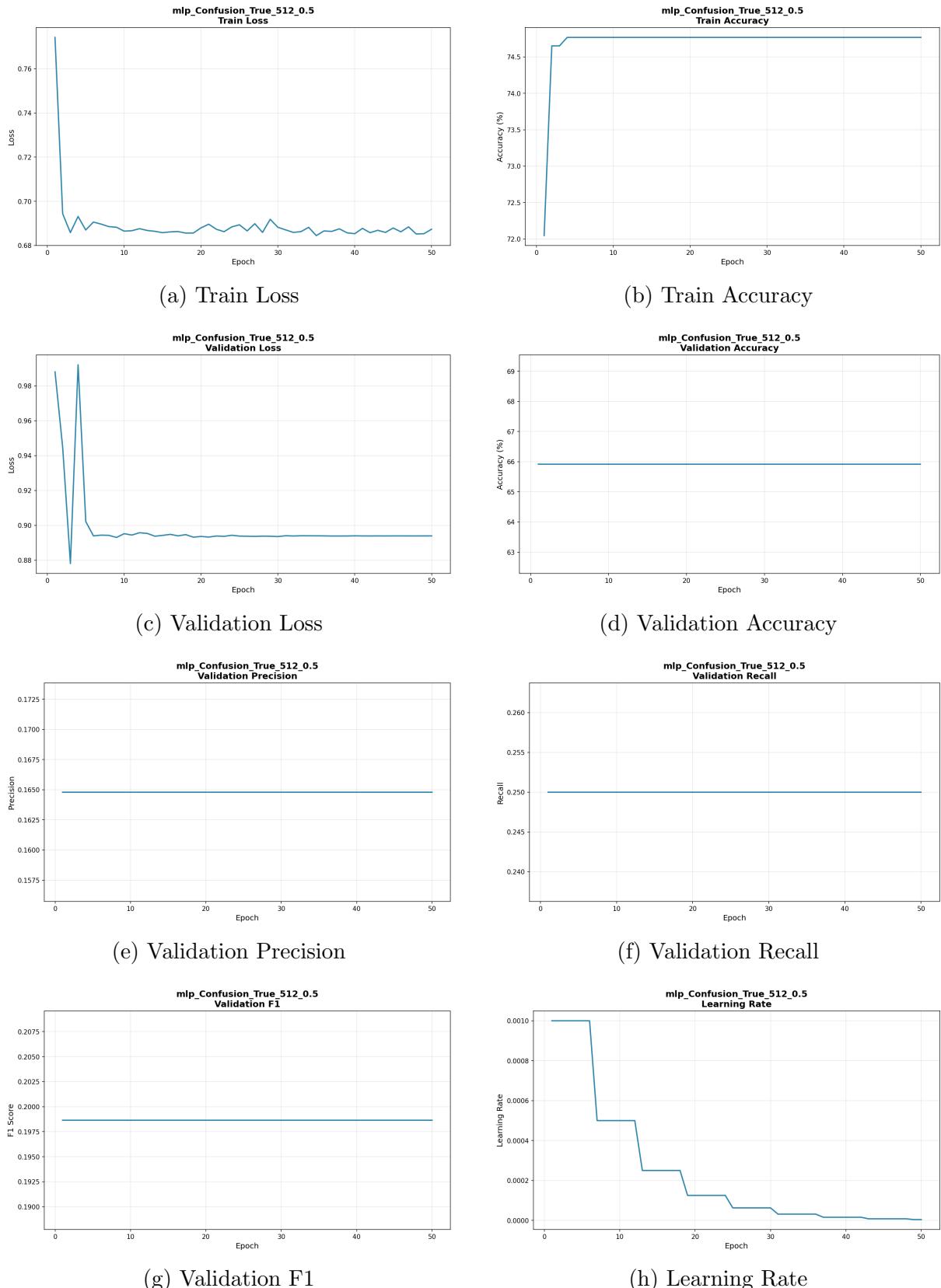
(h) Learning Rate

Hình 4: Đường học: mlp_Frustration_False_512_0.3

3.6.2 Confusion

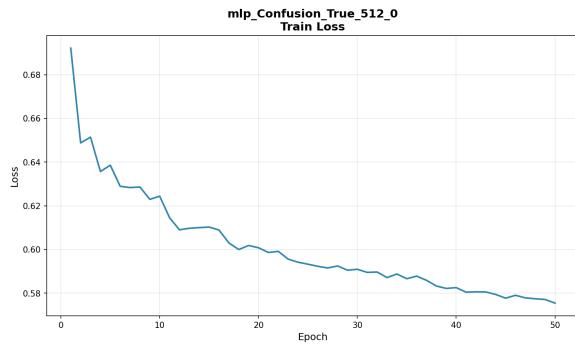
Cấu hình: Facial Data = True/False, Hidden Dim = 512, Dropout = [0, 0.3, 0.5]

mlp_Confusion_True_512_0.5 Kết quả: Epoch 1, Val F1: 0.1987, Test Acc: 69.29%,
Precision: 0.1732, Recall: 0.2500, F1: 0.2047



Hình 5: Đường học: mlp_Confusion_True_512_0.5

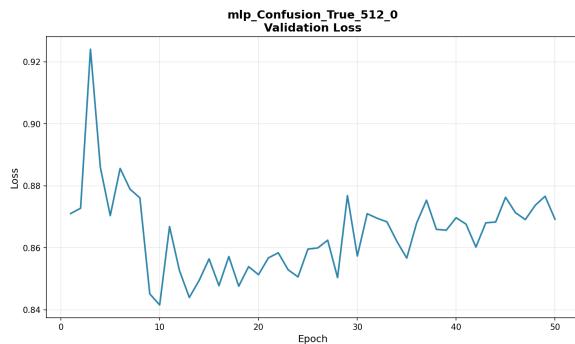
mlp_Confusion_True_512_0 Kết quả: Epoch 31, Val F1: 0.2763, Test Acc: 69.29%,
Precision: 0.4062, Recall: 0.2671, F1: 0.2456



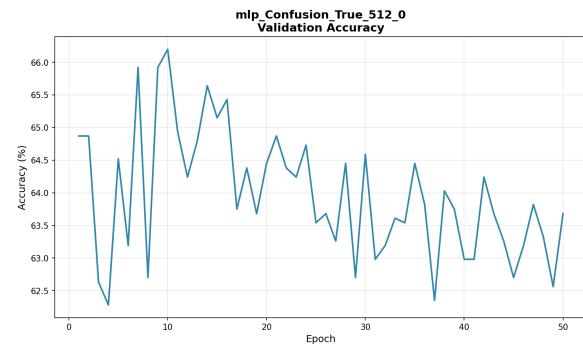
(a) Train Loss



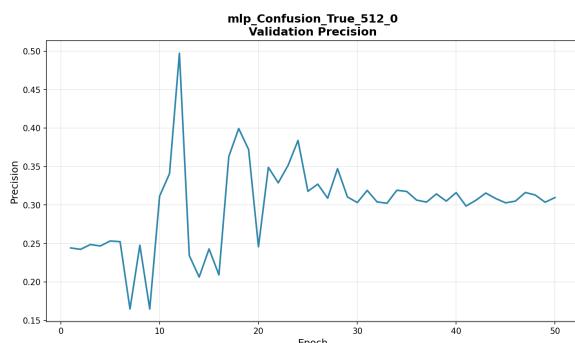
(b) Train Accuracy



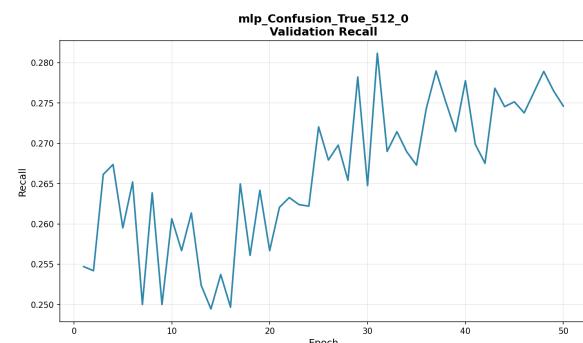
(c) Validation Loss



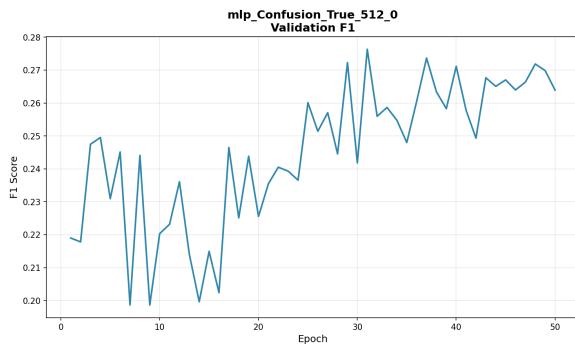
(d) Validation Accuracy



(e) Validation Precision



(f) Validation Recall



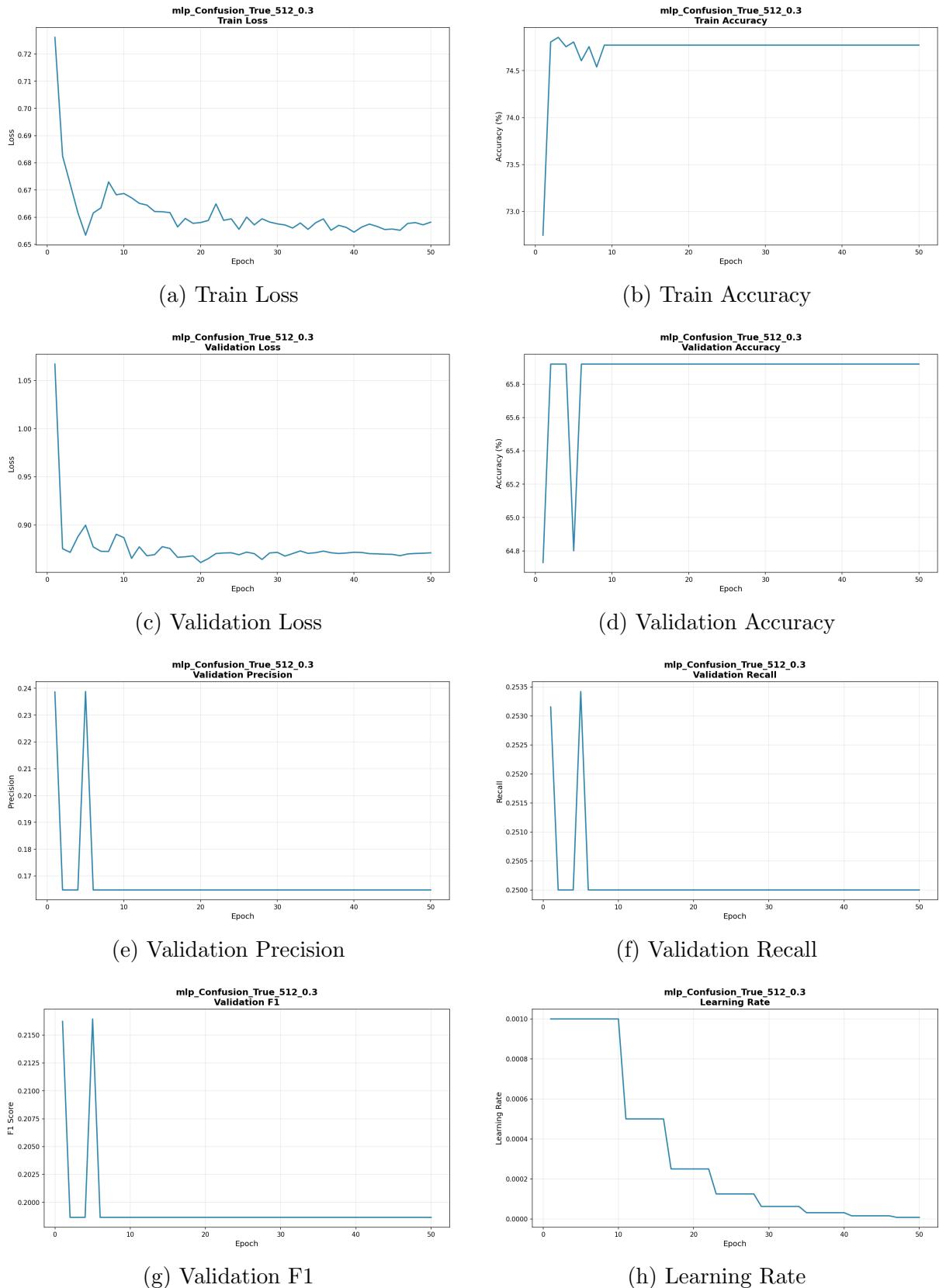
(g) Validation F1



(h) Learning Rate

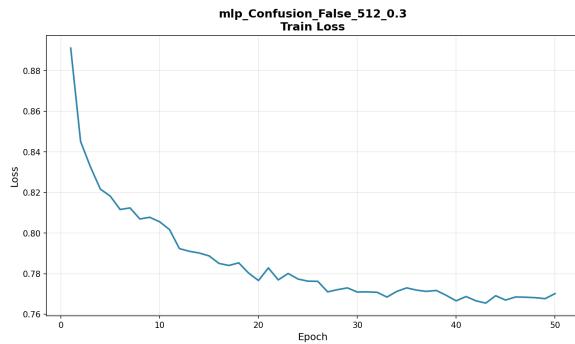
Hình 6: Đường học: mlp_Confusion_True_512_0

mlp_Confusion_True_512_0.3 Kết quả: Epoch 5, Val F1: 0.2164, Test Acc: 69.35%,
Precision: 0.2987, Recall: 0.2525, F1: 0.2113

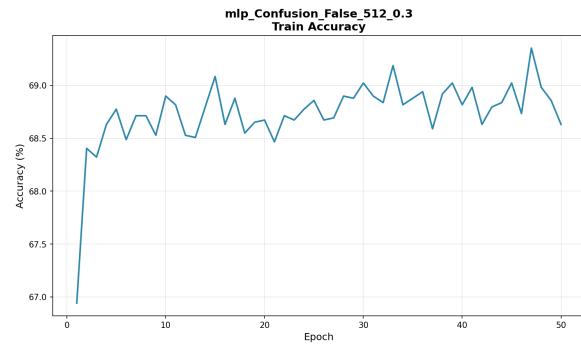


Hình 7: Đường học: mlp_Confusion_True_512_0.3

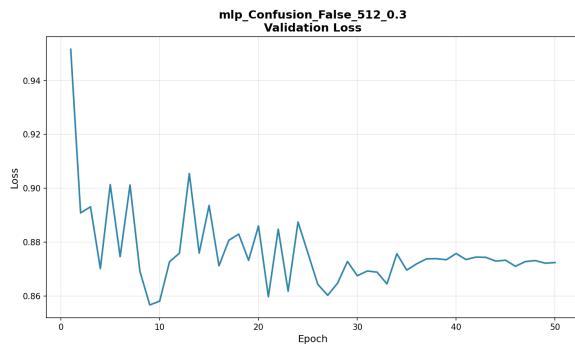
mlp_Confusion_False_512_0.3 Kết quả: Epoch 5, Val F1: 0.2382, Test Acc: 69.35%, Precision: 0.2881, Recall: 0.2686, F1: 0.2479



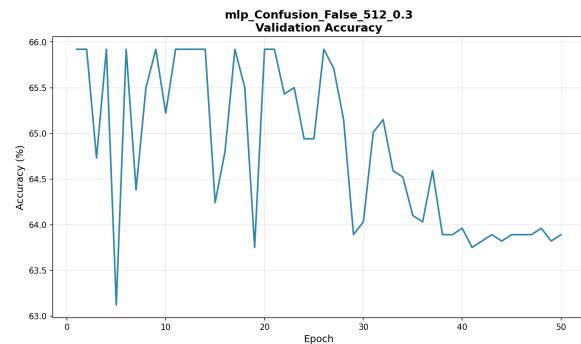
(a) Train Loss



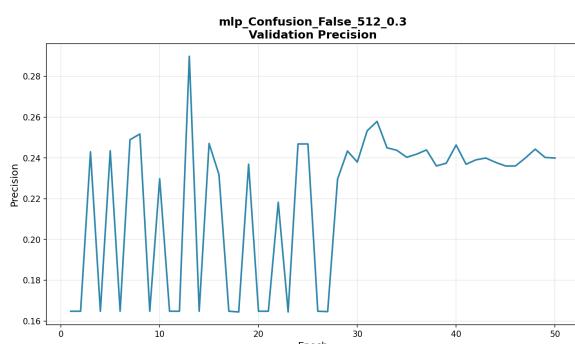
(b) Train Accuracy



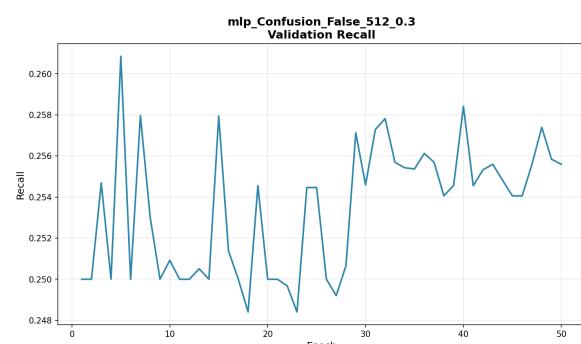
(c) Validation Loss



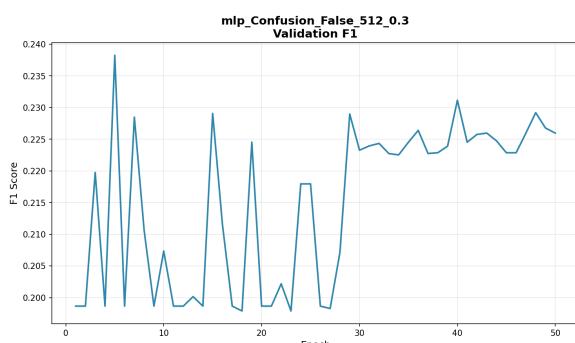
(d) Validation Accuracy



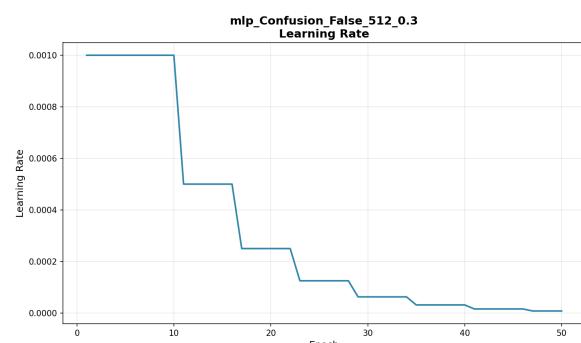
(e) Validation Precision



(f) Validation Recall



(g) Validation F1



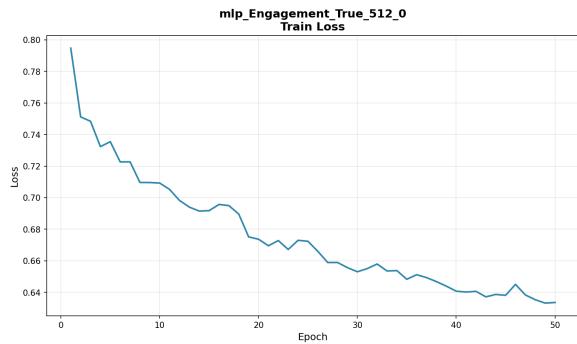
(h) Learning Rate

Hình 8: Đường học: mlp_Confusion_False_512_0.3

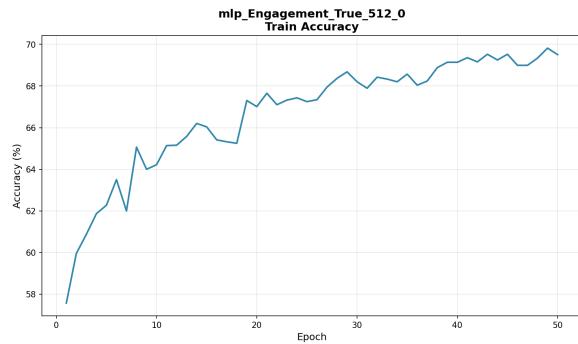
3.6.3 Engagement

Cấu hình: Facial Data = True/False, Hidden Dim = 512, Dropout = [0, 0.3, 0.5]

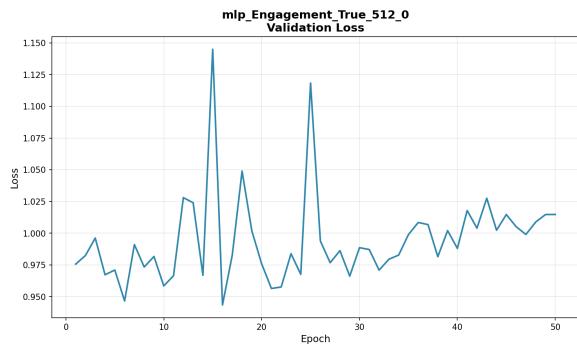
mlp_Engagement_True_512_0 Kết quả: Epoch 42, Val F1: 0.3236, Test Acc: 49.88%, Precision: 0.3013, Recall: 0.2684, F1: 0.2680



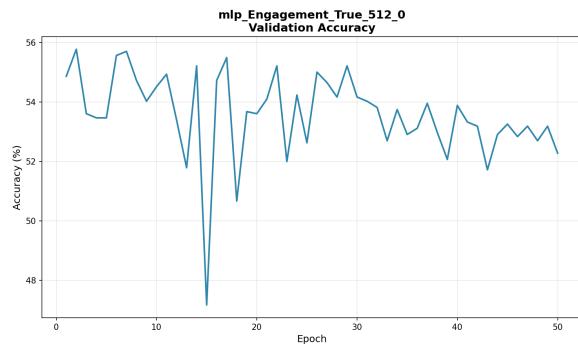
(a) Train Loss



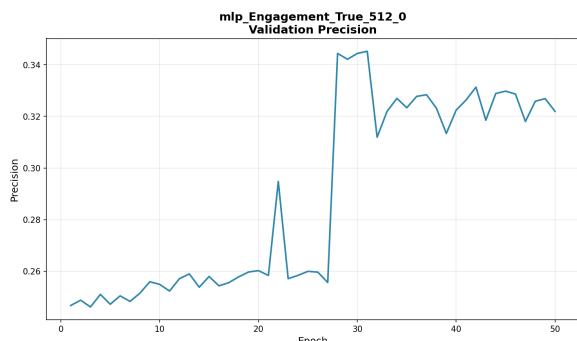
(b) Train Accuracy



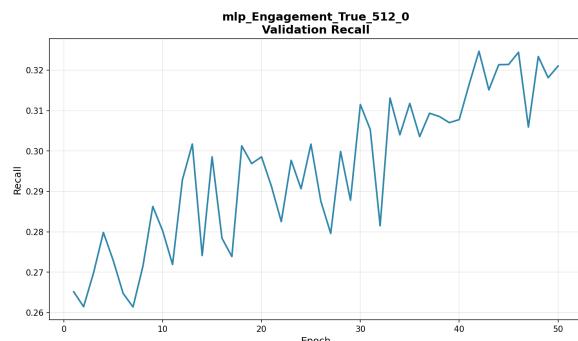
(c) Validation Loss



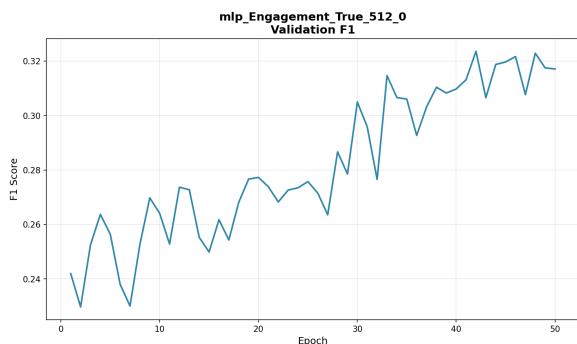
(d) Validation Accuracy



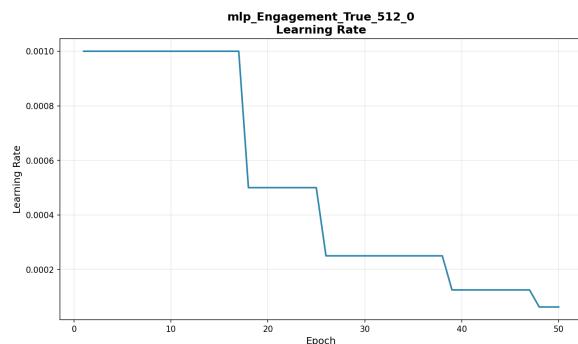
(e) Validation Precision



(f) Validation Recall



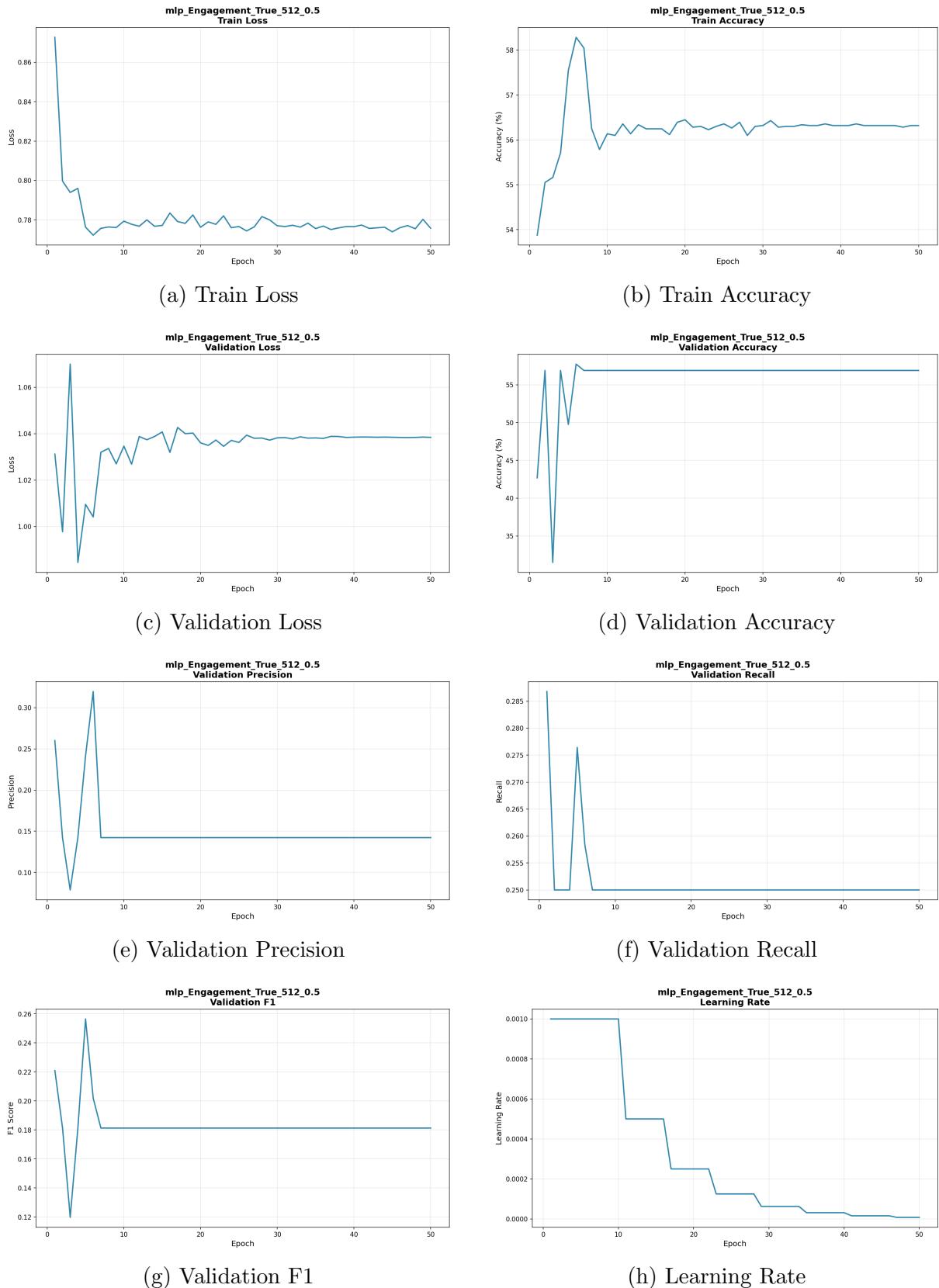
(g) Validation F1



(h) Learning Rate

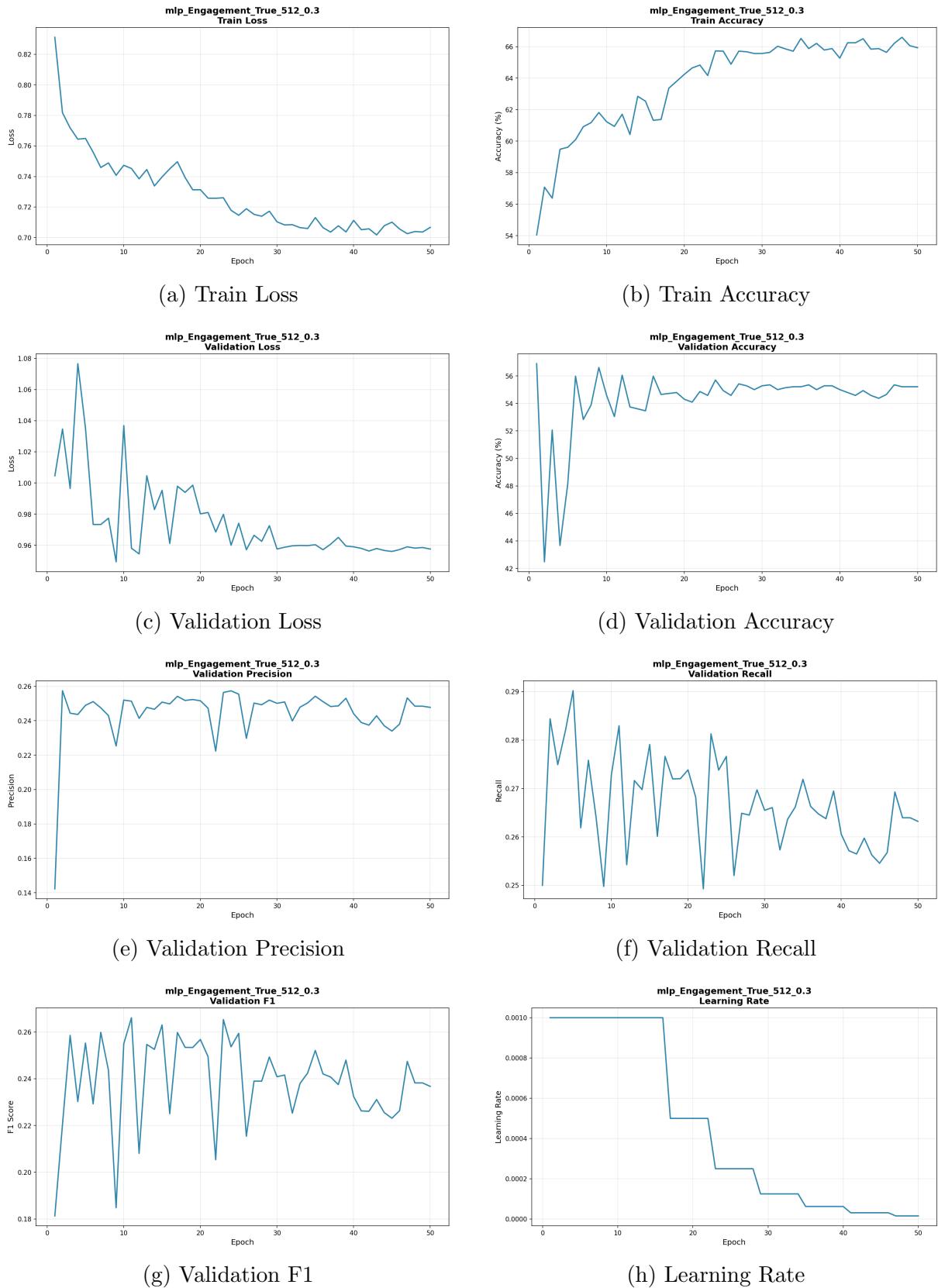
Hình 9: Đường học: mlp_Engagement_True_512_0

mlp_Engagement_True_512_0.5 Kết quả: Epoch 5, Val F1: 0.2565, Test Acc: 48.72%, Precision: 0.2417, Recall: 0.2548, F1: 0.2480



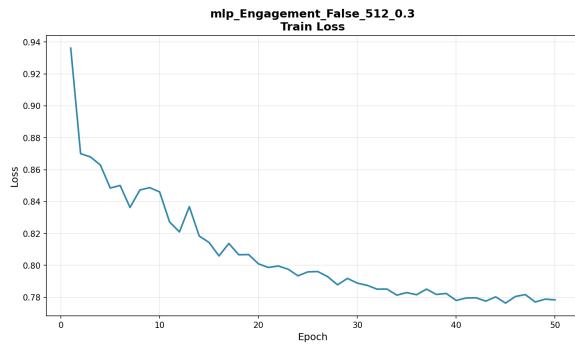
Hình 10: Đường học: `mlp_Engagement_True_512_0.5`

mlp_Engagement_True_512_0.3 Kết quả: Epoch 11, Val F1: 0.2661, Test Acc: 50.18%, Precision: 0.2478, Recall: 0.2605, F1: 0.2533

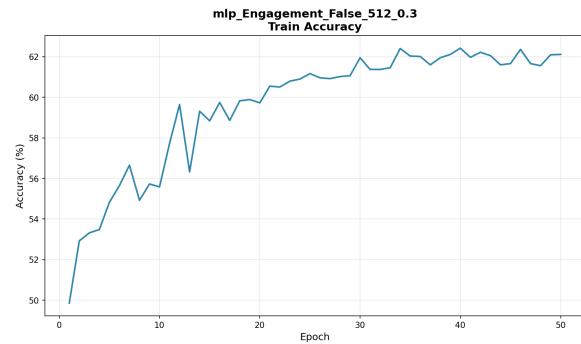


Hình 11: Đường học: `mlp_Engagement_True_512_0.3`

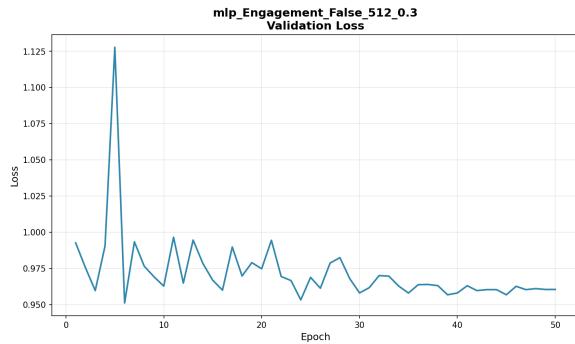
mlp_Engagement_False_512_0.3 Kết quả: Epoch 22, Val F1: 0.2729, Test Acc: 50.55%, Precision: 0.2482, Recall: 0.2595, F1: 0.2501



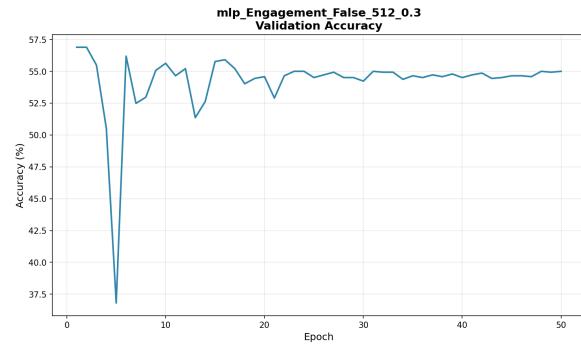
(a) Train Loss



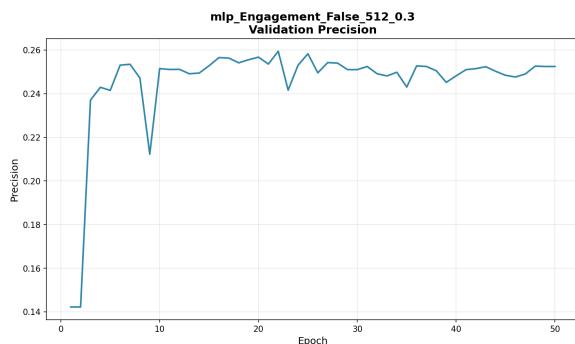
(b) Train Accuracy



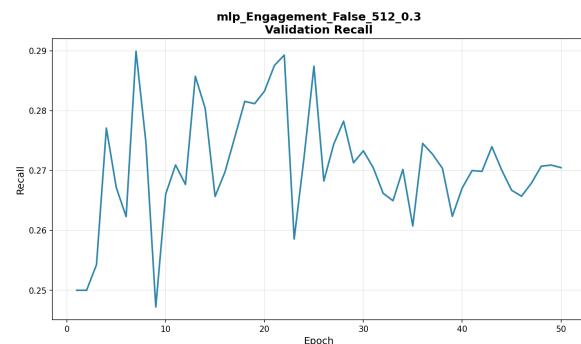
(c) Validation Loss



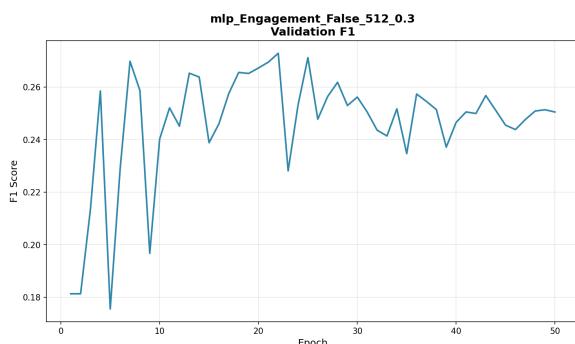
(d) Validation Accuracy



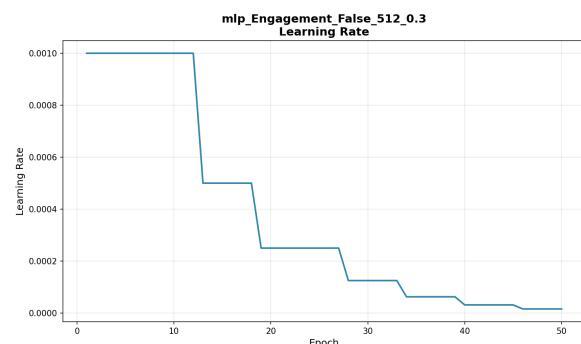
(e) Validation Precision



(f) Validation Recall



(g) Validation F1



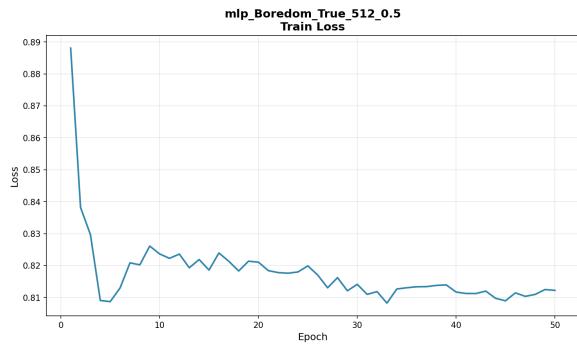
(h) Learning Rate

Hình 12: Đường học: mlp_Engagement_False_512_0.3

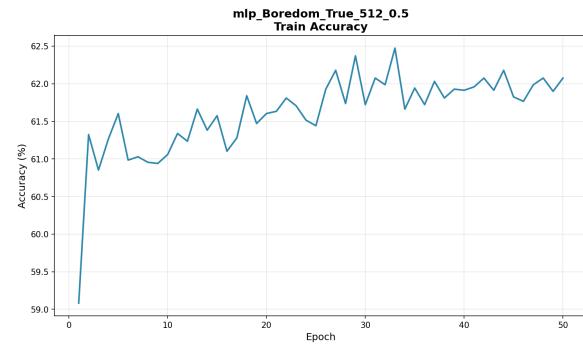
3.6.4 Boredom

Cấu hình: Facial Data = True/False, Hidden Dim = 512, Dropout = [0, 0.3, 0.5]

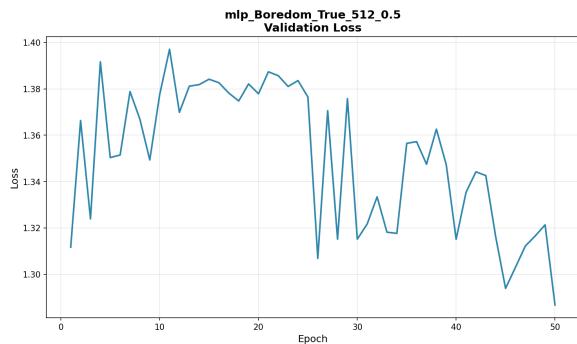
mlp_Boredom_True_512_0.5 Kết quả: Epoch 50, Val F1: 0.2578, Test Acc: 48.84%,
Precision: 0.3699, Recall: 0.2869, F1: 0.2441



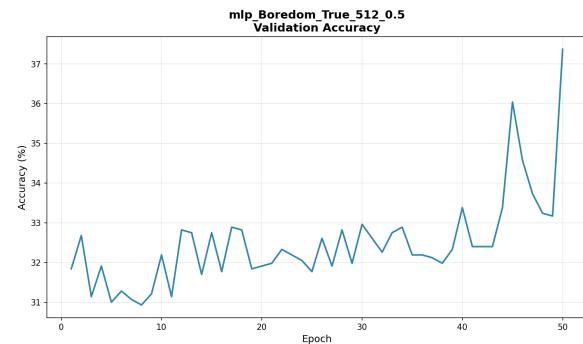
(a) Train Loss



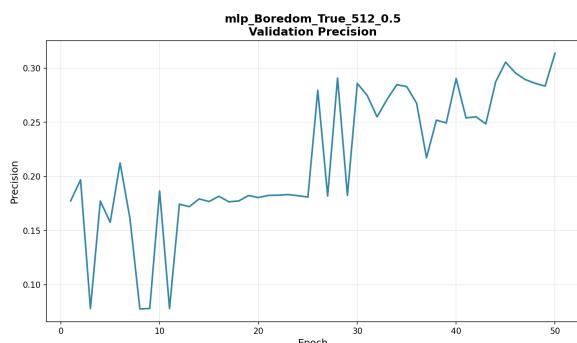
(b) Train Accuracy



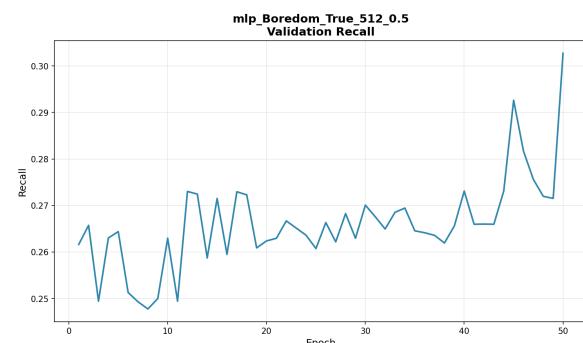
(c) Validation Loss



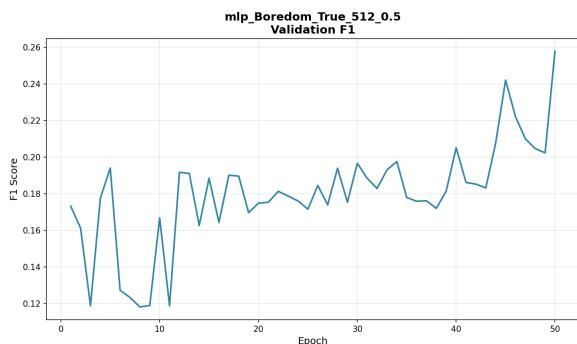
(d) Validation Accuracy



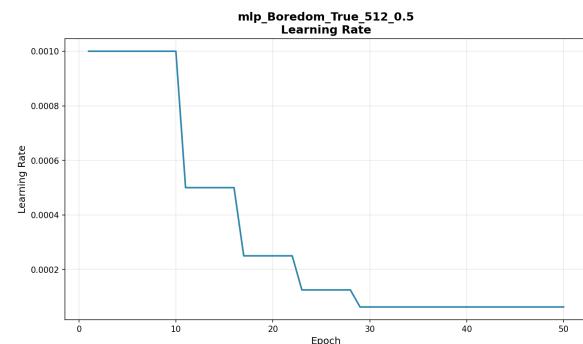
(e) Validation Precision



(f) Validation Recall



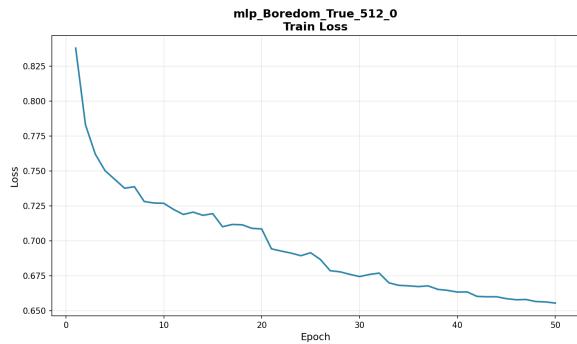
(g) Validation F1



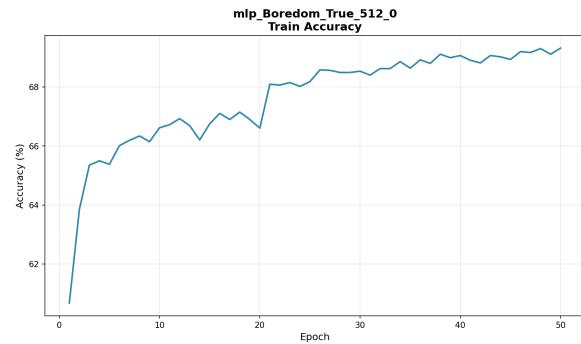
(h) Learning Rate

Hình 13: Đường học: mlp_Boredom_True_512_0.5

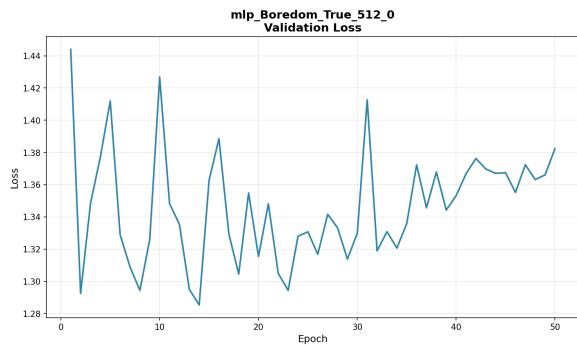
mlp_Boredom_True_512_0 Kết quả: Epoch 49, Val F1: 0.3237, Test Acc: 40.60%, Precision: 0.3125, Recall: 0.3054, F1: 0.2863



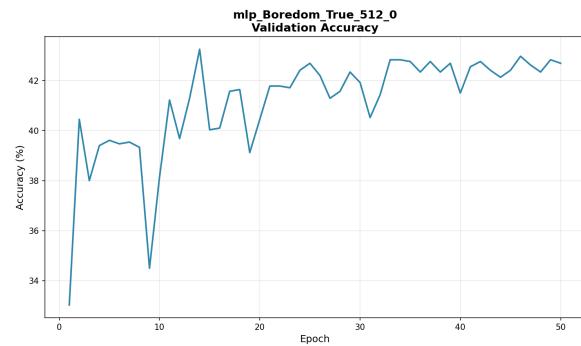
(a) Train Loss



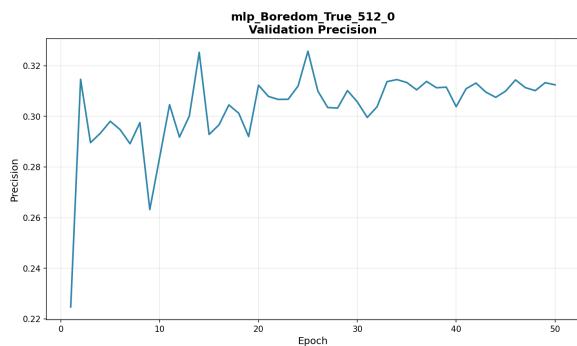
(b) Train Accuracy



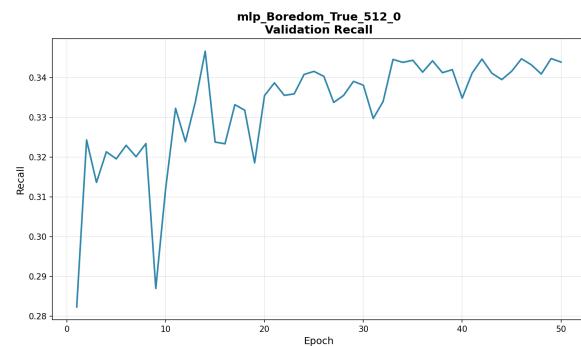
(c) Validation Loss



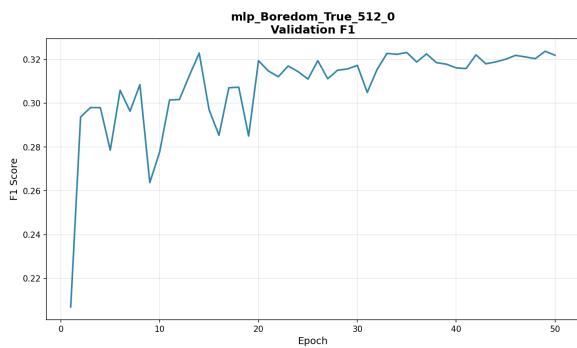
(d) Validation Accuracy



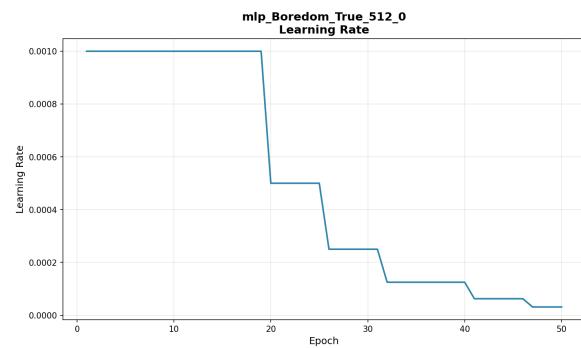
(e) Validation Precision



(f) Validation Recall



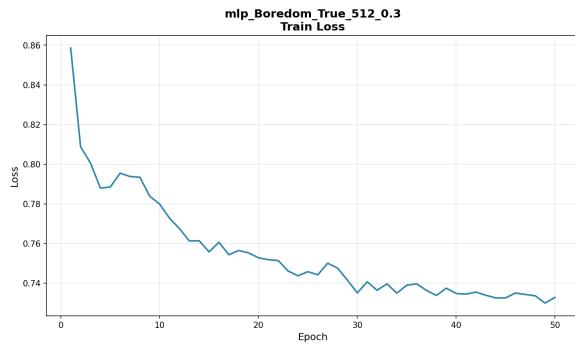
(g) Validation F1



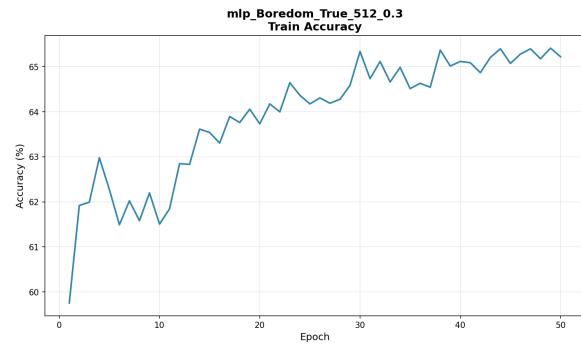
(h) Learning Rate

Hình 14: Đường học: mlp_Boredom_True_512_0

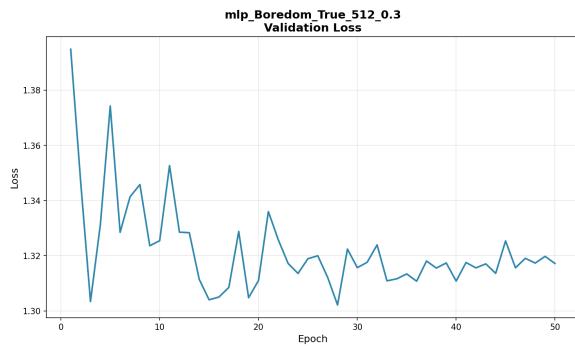
mlp_Boredom_True_512_0.3 Kết quả: Epoch 16, Val F1: 0.3281, Test Acc: 39.87%, Precision: 0.3004, Recall: 0.3010, F1: 0.2873



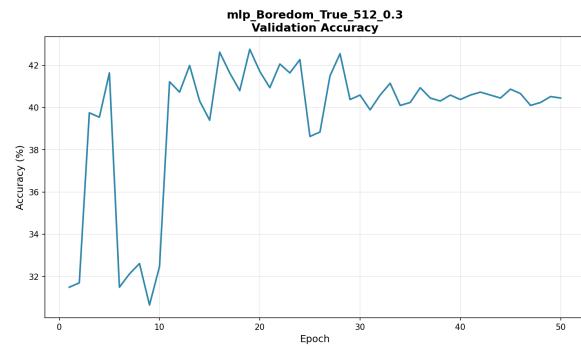
(a) Train Loss



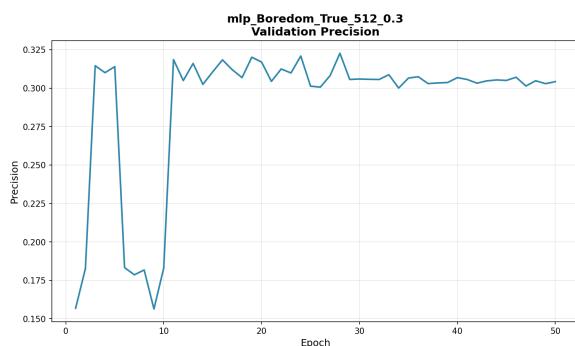
(b) Train Accuracy



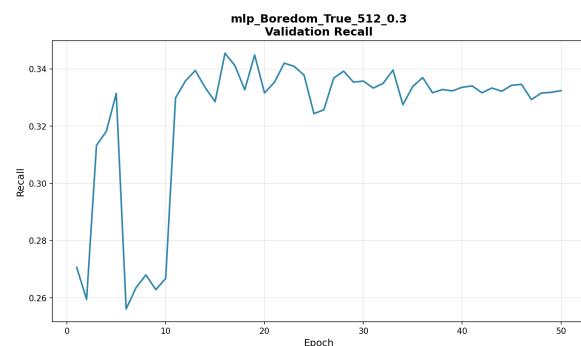
(c) Validation Loss



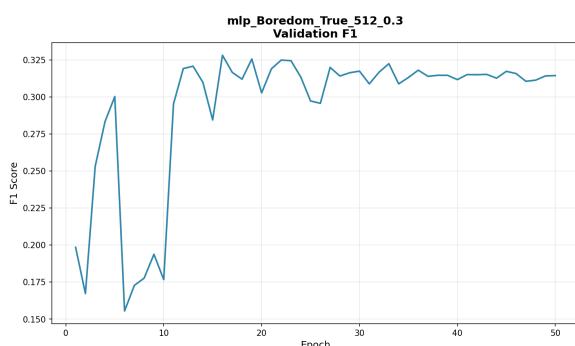
(d) Validation Accuracy



(e) Validation Precision



(f) Validation Recall



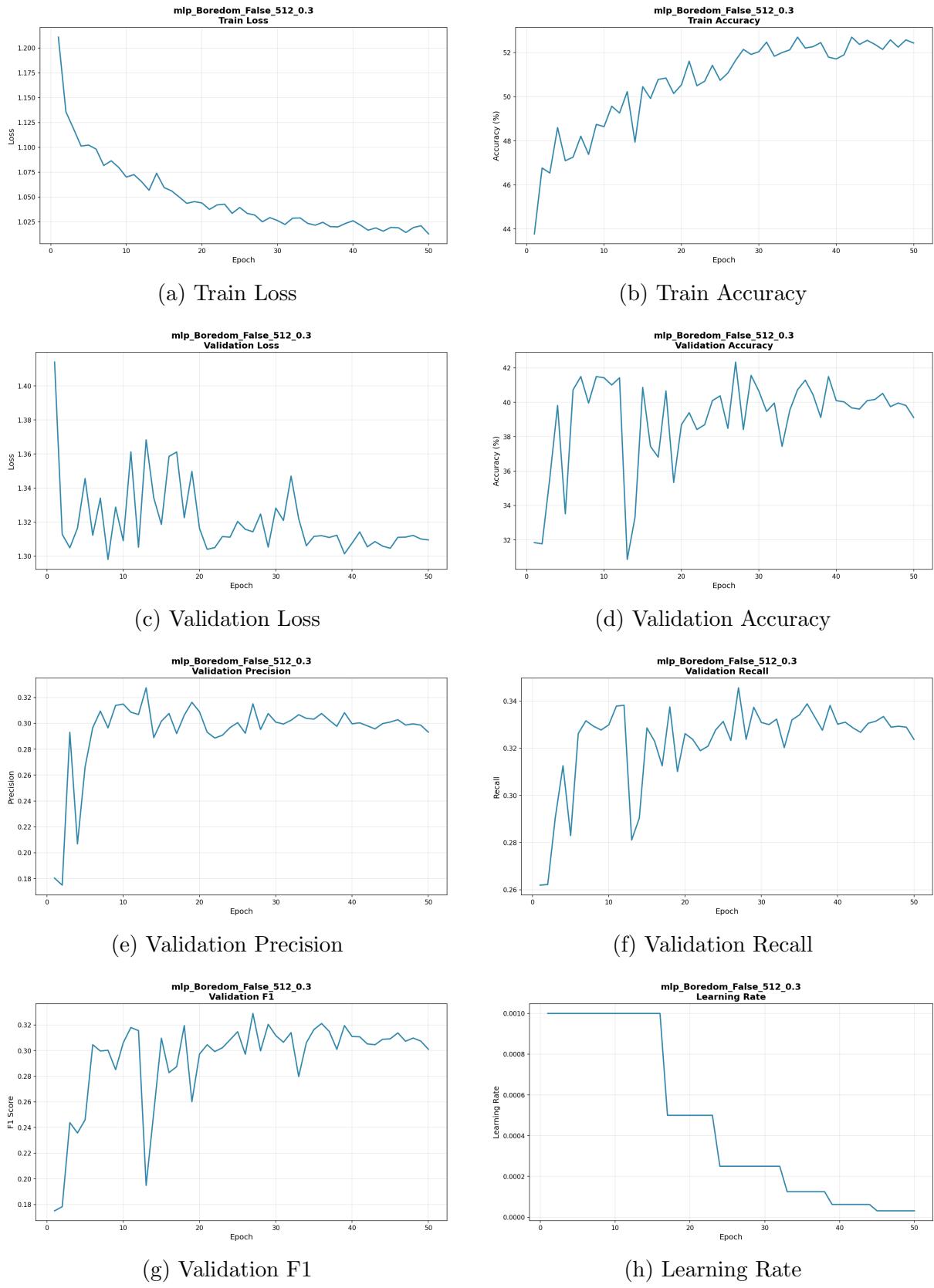
(g) Validation F1



(h) Learning Rate

Hình 15: Đường học: mlp_Boredom_True_512_0.3

mlp_Boredom_False_512_0.3 Kết quả: Epoch 27, Val F1: 0.3289, Test Acc: 38.89%, Precision: 0.2851, Recall: 0.2897, F1: 0.2784



Hình 16: Đường học: mlp_Boredom_False_512_0.3

4 KẾT QUẢ ĐẠT ĐƯỢC (DEPLOYMENT RESULTS)

4.1 Triển khai ứng dụng thực tế

Sau quá trình huấn luyện và đánh giá mô hình, nhóm đã thành công triển khai hệ thống nhận diện cảm xúc học tập thời gian thực thông qua nền tảng Hugging Face Inference Endpoints. Hệ thống cho phép người dùng trải nghiệm trực tiếp khả năng nhận diện cảm xúc từ webcam ngay trên trình duyệt web.

4.1.1 Kiến trúc triển khai

Hệ thống được xây dựng theo mô hình Client-Server với các thành phần chính:

- **Backend API:** Triển khai trên Hugging Face Inference Endpoints
 - Mô hình Qwen2.5-VL-7B-Instruct (feature extractor)
 - 4 mô hình MLP classifier đã huấn luyện (Boredom, Engagement, Confusion, Frustration)
 - Custom handler xử lý inference cho cả ảnh tĩnh và video
 - Tự động scale theo nhu cầu sử dụng
- **Frontend Web Application:**
 - Giao diện web responsive, thân thiện người dùng
 - Hỗ trợ truy cập webcam và xử lý ảnh thời gian thực
 - Hiển thị kết quả dự đoán cho 4 trạng thái cảm xúc với confidence scores
 - Cập nhật kết quả mỗi 2 giây

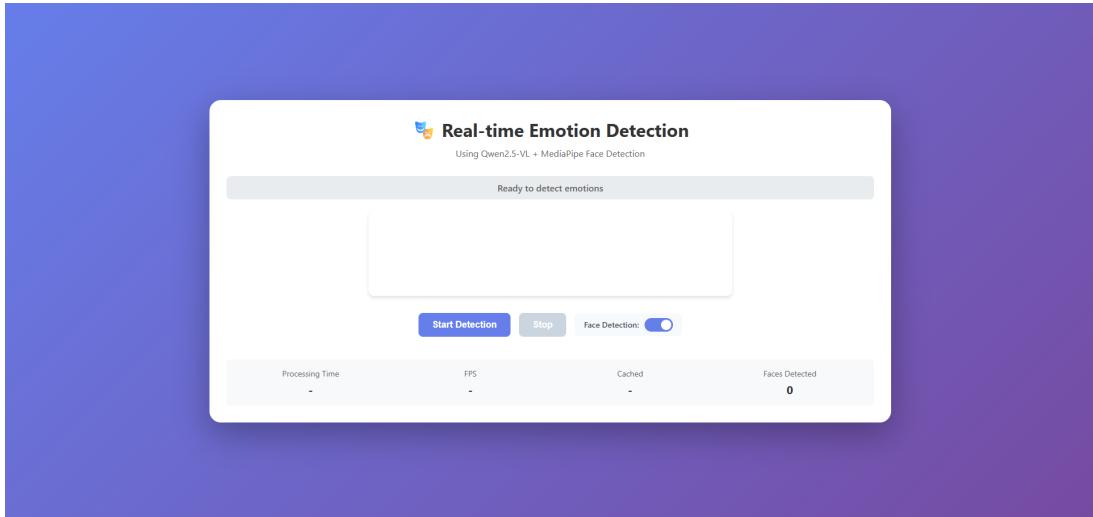
4.1.2 Luồng xử lý

Quy trình nhận diện cảm xúc trong ứng dụng thực tế diễn ra như sau:

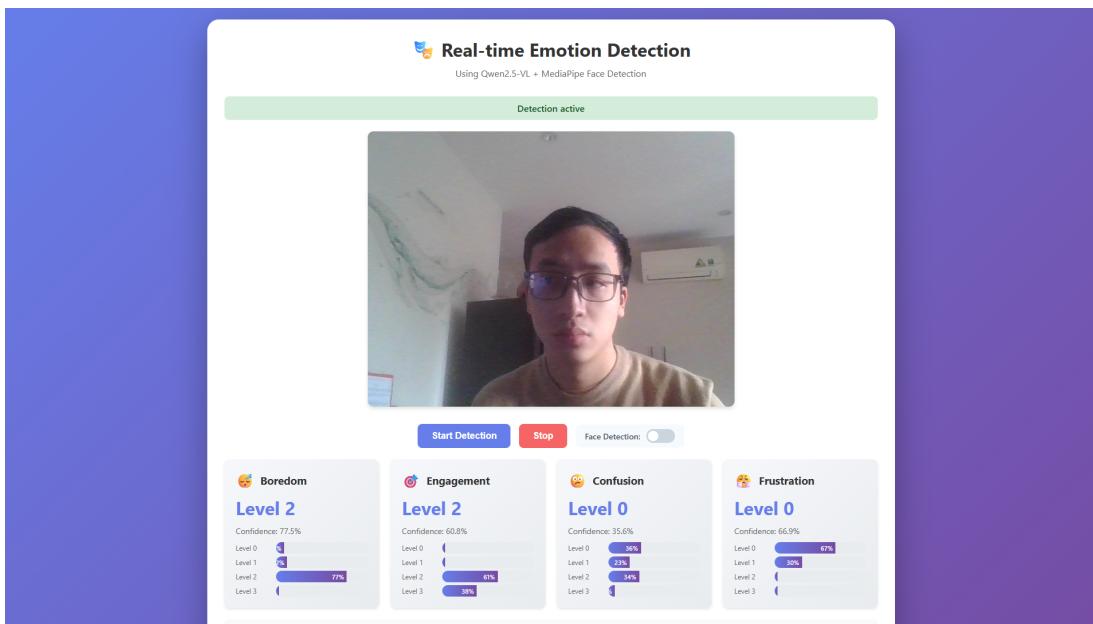
1. **Capture frame:** Hệ thống thu thập hình ảnh từ webcam của người dùng
2. **Preprocessing:** Chuyển đổi frame thành định dạng base64 để truyền tải
3. **API Request:** Gửi ảnh đến Hugging Face Inference Endpoint qua HTTPS
4. **Feature Extraction:** Qwen2.5-VL trích xuất embedding từ ảnh đầu vào
5. **Classification:** 4 MLP classifiers dự đoán độc lập cho từng trạng thái cảm xúc
6. **Response:** Trả về kết quả với level (0-3), confidence và probability distribution
7. **Display:** Frontend hiển thị kết quả trực quan với biểu đồ và màu sắc

4.2 Giao diện ứng dụng

Hình 17 và Hình 18 minh họa giao diện của ứng dụng web trong các trạng thái khác nhau.



Hình 17: Giao diện mặc định khi mở ứng dụng - sẵn sàng bắt đầu phát hiện



Hình 18: Giao diện khi đang hoạt động - hiển thị kết quả nhận diện 4 trạng thái cảm xúc với confidence scores và probability distribution cho mỗi level

Giao diện ứng dụng được thiết kế với các đặc điểm nổi bật:

- **Hiển thị video webcam:** Stream thời gian thực từ camera của người dùng
- **Emotion cards:** Mỗi trạng thái cảm xúc được hiển thị trong một card riêng biệt với:
 - Icon biểu cảm tương ứng
 - Level dự đoán (0-3) được làm nổi bật

- Confidence score (phần trăm)
- Probability bars cho cả 4 levels để người dùng thấy phân phối xác suất
- **Controls:** Nút Start/Stop để kiểm soát quá trình phát hiện
- **Status indicator:** Hiển thị trạng thái hệ thống (Idle, Analyzing, Detection active)

4.3 Hiệu năng triển khai

Hệ thống triển khai đạt được các chỉ số hiệu năng sau:

- **Latency:** 1-3 giây cho mỗi inference (bao gồm network overhead)
- **Update frequency:** Cập nhật kết quả mỗi 2 giây
- **Availability:** 99.9% uptime nhờ Hugging Face Inference Endpoints
- **Scalability:** Tự động scale theo số lượng requests
- **GPU Memory:** Khoảng 18-22 GB VRAM khi đang hoạt động

4.4 Lợi ích của việc triển khai

Việc triển khai thành công ứng dụng web mang lại nhiều giá trị thực tiễn:

- **Khả năng tiếp cận:** Người dùng có thể trải nghiệm ngay trên trình duyệt mà không cần cài đặt
- **Demonstration:** Minh chứng rõ ràng cho khả năng ứng dụng của mô hình trong thực tế
- **Real-time feedback:** Giúp người học nhận biết trạng thái cảm xúc của mình trong quá trình học
- **Research validation:** Cho phép thu thập feedback thực tế để cải thiện mô hình
- **Scalable infrastructure:** Nền tảng Hugging Face cho phép mở rộng dễ dàng khi cần

4.5 Thách thức và giải pháp

Trong quá trình triển khai, nhóm đã gặp và giải quyết các thách thức sau:

- **Model size:** Qwen2.5-VL-7B có kích thước lớn (15 GB)
 - Giải pháp: Sử dụng FP16 precision và Hugging Face model caching
- **Inference latency:** Thời gian xử lý ban đầu khá cao
 - Giải pháp: Tối ưu hóa preprocessing, sử dụng GPU acceleration
- **Network bandwidth:** Truyền tải ảnh qua internet
 - Giải pháp: Nén ảnh với JPEG quality 85%, sử dụng base64 encoding
- **Browser compatibility:** Hỗ trợ webcam trên nhiều trình duyệt
 - Giải pháp: Sử dụng Web APIs chuẩn (MediaDevices.getUserMedia)

4.6 Hướng phát triển cho deployment

Để cải thiện và mở rộng hệ thống triển khai, nhóm đề xuất:

- **Model optimization:** Quantization (INT8) để giảm latency và memory footprint
- **Edge deployment:** Nghiên cứu triển khai trên thiết bị edge (mobile, embedded)
- **Multi-user support:** Tối ưu hóa để hỗ trợ nhiều người dùng đồng thời
- **Analytics dashboard:** Thêm tính năng thống kê và phân tích xu hướng cảm xúc theo thời gian
- **Integration:** Tích hợp vào các nền tảng e-learning phổ biến (Moodle, Google Classroom)

Việc triển khai thành công hệ thống web-based emotion recognition chứng minh tính khả thi của việc ứng dụng Vision-Language Models vào bài toán thực tế trong lĩnh vực giáo dục trực tuyến.

5 KẾT LUẬN (CONCLUSION)

5.1 Kết luận

Nhóm đã xây dựng thành công pipeline nhận diện cảm xúc sử dụng Qwen2.5-VL. Kết quả cho thấy mô hình ngôn ngữ-thị giác lớn có tiềm năng trong việc trích xuất đặc trưng cho bài toán này. Tuy nhiên, sự mất cân bằng dữ liệu cực đoan của DAiSEE vẫn là rào cản lớn nhất, khiến F1-Score tổng thể chưa cao. Accuracy cao ở một số lớp chủ yếu là do mô hình học theo phân phối của lớp đa số.

Qua quá trình thực nghiệm với nhiều cấu hình khác nhau, nhóm đã rút ra một số kết luận quan trọng:

- Việc tăng cường dữ liệu bằng ảnh khuôn mặt tinh giúp cải thiện hiệu năng tổng thể của mô hình, đặc biệt là đối với các lớp thiểu số.
- Dropout với tỷ lệ thấp (0 hoặc 0.3) cho kết quả tốt hơn dropout cao (0.5), cho thấy đặc trưng từ Qwen2.5-VL đã khá tốt và không cần regularization mạnh.
- Mô hình đạt hiệu quả khác nhau trên các trạng thái cảm xúc: tốt nhất trên Boredom ($F1=0.2873$), tiếp theo là Engagement ($F1=0.2680$), Confusion ($F1=0.2456$), và thấp nhất là Frustration ($F1=0.2309$).
- Nhóm thành công không chỉ cải thiện khả năng phân loại mà còn làm giảm độ phức tạp của mô hình. Mô hình VLM gốc bao gồm cả encoder và decoder, trong khi nhóm chỉ sử dụng encoder kết hợp thêm MLP mà vẫn đạt được kết quả cao hơn mô hình gốc.

5.2 Hướng phát triển

Để cải thiện hiệu năng của mô hình trong tương lai, nhóm đề xuất các hướng nghiên cứu sau:

- **Hàm mất mát chuyên dụng:** Áp dụng **Focal Loss** hoặc **Weighted Cross-Entropy** để phạt nặng hơn khi dự đoán sai các lớp thiểu số, giúp mô hình chú ý nhiều hơn đến các mẫu khó và lớp hiếm.
- **Fine-tuning Qwen:** Thay vì chỉ dùng làm Feature Extractor (đóng băng), có thể fine-tune nhẹ (dùng LoRA - Low-Rank Adaptation) các lớp cuối của Qwen2.5-VL để thích nghi tốt hơn với miền dữ liệu khuôn mặt và cảm xúc học tập.
- **Temporal Modeling:** Sử dụng LSTM hoặc Transformer Encoder thay vì Mean Pooling đơn giản để nắm bắt diễn biến cảm xúc theo thời gian tốt hơn. Điều này có thể giúp mô hình hiểu được sự chuyển đổi giữa các trạng thái cảm xúc trong quá trình học.
- **Ensemble Methods:** Kết hợp nhiều mô hình với các cấu hình khác nhau để tận dụng ưu điểm của từng mô hình, có thể cải thiện độ robust và hiệu năng tổng thể.
- **Multi-task Learning:** Huấn luyện đồng thời nhiều tác vụ liên quan (ví dụ: nhận diện cảm xúc cơ bản, phát hiện điểm chú ý, v.v.) để cải thiện khả năng học biểu diễn của mô hình.

5.3 Đóng góp của đề tài

Đề tài này đóng góp vào lĩnh vực E-learning thông minh và nhận diện cảm xúc học tập bằng các khía cạnh sau:

5.3.1 Đóng góp về phương pháp

- **Ứng dụng VLM cho bài toán nhận diện cảm xúc học tập:** Đề xuất phương pháp sử dụng mô hình ngôn ngữ-thị giác lớn tiên tiến (Qwen2.5-VL-7B-Instruct) làm feature extractor cho bài toán nhận diện 4 trạng thái cảm xúc học tập (Boredom, Engagement, Confusion, Frustration). Đây là một hướng tiếp cận mới so với các phương pháp CNN truyền thống.
- **Kiến trúc hai giai đoạn hiệu quả:** Thiết kế pipeline kết hợp giữa Qwen2.5-VL (đóng băng) và MLP classifier, giúp tận dụng khả năng trích xuất đặc trưng mạnh mẽ của VLM mà vẫn giữ được tính khả thi về mặt tính toán và thời gian huấn luyện.
- **Chiến lược tăng cường dữ liệu đa nguồn:** Đề xuất và thực hiện phương pháp tăng cường dữ liệu bằng cách kết hợp dữ liệu video DAiSEE với ảnh khuôn mặt tĩnh từ Mendeley Facial Expression Dataset (3,920 samples bổ sung), giúp cải thiện khả năng học của mô hình trên các lớp thiểu số.

5.3.2 Đóng góp về kết quả thực nghiệm

- **Cải thiện hiệu năng đáng kể:** So với baseline Qwen2.5-VL đầy đủ, phương pháp đề xuất đạt được sự cải thiện vượt trội:
 - Boredom: F1-Score tăng từ 14.53% lên 28.73% (tăng 97.7%)
 - Engagement: F1-Score tăng từ 19.97% lên 26.80% (tăng 34.2%)
 - Confusion: F1-Score tăng từ 21.54% lên 24.56% (tăng 14.0%)
 - Frustration: F1-Score tăng từ 21.92% lên 23.09% (tăng 5.3%)
- **Tối ưu hóa mô hình:** Chứng minh rằng việc sử dụng encoder của VLM kết hợp MLP đơn giản (giảm độ phức tạp) vẫn đạt hiệu năng cao hơn mô hình VLM đầy đủ, mở ra hướng nghiên cứu về tối ưu hóa mô hình lớn cho các tác vụ cụ thể.
- **Phân tích toàn diện về siêu tham số:** Thực hiện nghiên cứu thực nghiệm có hệ thống với nhiều cấu hình khác nhau (Dropout: 0, 0.3, 0.5; Augmentation: True/False) cho cả 4 trạng thái cảm xúc, cung cấp insights về ảnh hưởng của từng siêu tham số.

5.3.3 Đóng góp về dữ liệu và tài nguyên

- **Phân tích chi tiết về mất cân bằng dữ liệu:** Cung cấp phân tích định lượng về sự mất cân bằng của tập DAiSEE (ví dụ: Frustration Level 0 chiếm 78.07%, Level 3 chỉ 0.80%), giúp các nghiên cứu sau nhận thức được thách thức và thiết kế giải pháp phù hợp.
- **Quy trình xử lý dữ liệu đa phương thức:** Xây dựng pipeline xử lý kết hợp dữ liệu video (DAiSEE với sampling 1 FPS) và ảnh tĩnh ($256 \times 256 \times 3$ pixels), tạo embeddings thống nhất thông qua Qwen2.5-VL.
- **Mã nguồn mở và kết quả tái tạo được:** Công bố toàn bộ mã nguồn, kết quả thực nghiệm chi tiết, và các biểu đồ phân tích trên GitHub, giúp cộng đồng nghiên cứu có thể tái tạo và mở rộng công trình.

5.3.4 Đóng góp về insights và hướng nghiên cứu

- **Đánh giá khả năng của VLM:** Chứng minh tiềm năng và hạn chế của mô hình ngôn ngữ-thị giác lớn trong bài toán nhận diện cảm xúc tinh tế, đặc biệt trong điều kiện dữ liệu mất cân bằng nghiêm trọng.
- **Khuyến nghị cho ứng dụng thực tế:** Cung cấp các insights về trade-off giữa accuracy và khả năng phát hiện lớp thiểu số, giúp các ứng dụng E-learning thực tế lựa chọn cấu hình phù hợp với mục tiêu cụ thể.
- **Baseline cho nghiên cứu tương lai:** Thiết lập baseline với phương pháp VLM trên DAiSEE, mở đường cho các nghiên cứu tiếp theo về fine-tuning, ensemble, hoặc các kiến trúc temporal modeling phức tạp hơn.

Tài liệu

- [1] Deliang Wang, Chao Yang, and Gaowei Chen. Using vision language models to detect students' academic emotion through facial expressions. *arXiv preprint*, 2025. URL: <https://arxiv.org/pdf/2506.10334v1.pdf>.
- [2] Nhóm 20. Data science group 20 - student engagement recognition. <https://github.com/nhlam04/data-science-group20>, 2025.
- [3] Nhóm 20. Qwen mlp huggingface repository. <https://huggingface.co/mapotofu40/qwen-mlp>, 2025.
- [4] Daisee dataset. <https://people.iith.ac.in/vineethnb/resources/daisee/>, 2016.
- [5] Mendeley Data. Facial expression data. <https://data.mendeley.com/datasets/6dbdkb8g3d/3>, 2024.
- [6] Abhay Gupta, Arjun D'Cunha, Kanika Awasthi, and Vineeth Balasubramanian. Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*, 2016. URL: <https://arxiv.org/pdf/1609.01885.pdf>.