TRƯỜNG ĐẠI HỌC BÁCH KHOA

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO

PBL5 – DỰ ÁN KỸ THUẬT MÁY TÍNH

TÊN ĐỀ TÀI: HỆ THỐNG NHẬN DIỆN CẢM XÚC NGƯỜI DÙNG QUA HÌNH ẢNH VÀ ÂM THANH

Giảng viên hướng dẫn: TS. Phạm Công Thắng

NHÓM SINH VIÊN THỰC HIỆN	LỚP HỌC PHẦN
Nguyễn Tiến Quang	22T_DT4
Huỳnh Thị Thanh Nhàn	22T_DT5

Đà Nẵng, 06/2025

MỤC LỤC

DANH SÁCH HÌNH ẢNH 4
DANH SÁCH BẢNG BIỂU 5
Tóm tắt đồ án 6
1. Giới thiệu 1
1.1 Hiện trạng 1
1.2 Các vấn đề cần giải quyết 1
1.3 Đề xuất giải pháp tổng quan 2
2. Giải pháp 3
2.1 Giải pháp phần cứng và truyền thông 3
2.1.1. Sơ đồ tổng quan hệ thống 3
a. Mô tả chung: 6
b. Thông số kỹ thuật: 7
c. Chức năng: 7
a. Mô tả chung: 8
b. Thông số kỹ thuật: 8
c. Chức năng: 9
2.1.2. Sơ đồ kết nối linh kiện phần cứng 9
2.1.3 Truyền thông 10
Tóm tắt luồng truyền thông trong hệ thống 11
2.2 Giải pháp AI/KHDL 12
2.2.1. Giải pháp nhận diện cảm xúc qua hình ảnh (khuôn mặt) 12
2.2.2. Giải pháp nhận diện cảm xúc qua giọng nói (âm thanh) 15
2.3 Giải pháp phần mềm 17
2.3.1 Phát triển bài toán 17
2.3.6 Quy trình thực thi hệ thống 21
3. Kết quả 22
3.1. Nhận diện cảm xúc từ khuôn mặt 22
3.1.1. Tập dữ liệu 22
3.2.2. Huấn luyện mô hình 23
3.2. Nhận diện cảm xúc từ giọng nói 25
3.2.1. Tập dữ liệu 25
3.2.2. Huấn luyện mô hình 25
3.2.3. Kết quả nhận diện 27
3.3. Tích hợp và hiển thị kết quả 27
3.3.1. Giao diện trang chính (/) 27

3.3.2. Ghi dữ liệu cảm xúc vào file CSV 28
3.3.3. Trang thống kê kết quả (/summary) 29
3.3.4 Giao diện thống kê mức độ hài lòng 31
4. KẾT LUẬN 34
4.1. Đánh giá 34
4.2. Hướng phát triển 34
4.2.1. Nhận diện cảm xúc qua hình ảnh 34
4.2.2. Nhận diện cảm xúc qua âm thanh 34
4.2.3. Phần cứng 35
4.3. Chi phí thực hiện 35
DANH MỤC TÀI LIỆU THAM KHẢO 36

DANH SÁCH HÌNH ẢNH

Hình 1: Sơ đồ tổng quát hệ thống 3
Hình 2. Rasberry Pi 4 Model B.....4
Hình 3: Sơ đồ chân của Rasperry Pi 4 5
Hình 4: Raspberry Pi Camera Module V2 8M.....6
Hình 5: Mini Microphone USB MI-305 Raspb 8
Hình 6: Sơ đồ kết nối linh kiện phần cứng.....10
Hình 7. Sơ đồ khối hệ thống 20
Hình 8. Sơ đồ tuần tự luồng thực hiện của hệ thống.....21
Hình 9 . Kết quả train model cảm xúc từ hình ảnh 24
Hình 10. Kết quả train model cảm xúc từ hình ảnh 27
Hình 11: Giao diện chính hiển thị danh sách chuyến đi và cảm xúc mới nhất 28
Hình 12: Ví dụ nội dung file CSV một chuyến đi.....29
Hình 13: Biểu đồ thống kê tỉ lệ cảm xúc theo khuôn mặt và giọng nói.....30
Hình 14: Giao diện thống kê mức độ hài lòng.....33

DANH SÁCH BẢNG BIỂU

Bảng 1: Bảng đề xuất giải pháp tổng quan.....2
Bảng 2: Bảng thông số kỹ thuật của Raspberry Pi Camera 7
Bảng 3: Bảng thông số kỹ thuật của Mini Microphone USB MI-305 9
Bảng 4: chi tiết kết nối phần cứng 10
Bảng 5. Kiến trúc mô hình CNN nhận diện cảm xúc khuôn mặt.....14
Bảng 6. Kiến trúc mô hình CNN nhận diện cảm xúc qua giọng nói.....17
Bảng 7: Bảng công nghệ sử dụng 19
Bảng 8: Hệ thống quy đổi cảm xúc thành hệ số đánh giá 31
Bảng 9: Kế quả phân loại.....31
Bảng 10: Chi phí thực hiện 35

Tóm tắt đồ án

Ngày nay, cảm xúc của người dùng đóng vai trò rất quan trọng trong nhiều lĩnh vực như giáo dục, chăm sóc sức khỏe tinh thần, dịch vụ khách hàng và các hệ thống tương tác người–máy. Tuy nhiên, việc nhận biết và phân tích cảm xúc vẫn còn phụ thuộc nhiều vào đánh giá chủ quan của con người, thiếu sự chính xác và nhất quán. Xuất phát từ thực tế đó, nhóm đã thực hiện đề tài "Hệ thống Nhận Diện Cảm Xúc Người Dùng Qua Hình Ảnh và Âm Thanh" nhằm xây dựng một giải pháp thông minh giúp phát hiện cảm xúc của người dùng một cách tự động và chính xác hơn.

Hệ thống cho phép người dùng tương tác thông qua camera và microphone. Dữ liệu khuôn mặt và giọng nói sẽ được thu nhận, sau đó xử lý bằng các mô hình học sâu (Deep Learning). Đầu tiên, hệ thống sẽ phân tích hình ảnh khuôn mặt để nhận biết cảm xúc qua biểu cảm. Đồng thời, giọng nói của người dùng cũng được trích xuất các đặc trưng như tần số, âm sắc để xác định cảm xúc tương ứng. Kết quả từ cả hai nguồn dữ liệu sẽ được kết hợp lại để đưa ra một nhận định cảm xúc cuối cùng.

Với hệ thống này, người dùng có thể dễ dàng biết được cảm xúc đang thể hiện là gì (như vui vẻ, buồn bã, tức giận, ngạc nhiên,...), đồng thời hệ thống cũng có thể được ứng dụng vào các lĩnh vực như giám sát lớp học, tư vấn tâm lý, đánh giá trải nghiệm sản phẩm, dịch vụ.

Bảng phân công nhiệm vụ

Sinh viên thực hiện	Các nhiệm vụ
Nguyễn Tiến Quang	- Thu thập và xử lý dữ liệu - Xây dựng mô hình nhận diện cảm xúc bằng âm thanh - Triển khai phần cứng - Xây dựng website - Viết báo cáo
Huỳnh Thị Thanh Nhàn	- Thu thập và xử lý dữ liệu - Xây dựng mô hình nhận diện cảm xúc bằng hình ảnh - Xây dựng website - Viết báo cáo

1. Giới thiệu

Trong thời đại số và kỷ nguyên trí tuệ nhân tạo (AI), việc hiểu và phản hồi đúng với cảm xúc của người dùng ngày càng trở thành yếu tố then chốt trong các hệ thống tương tác người–máy. Các hệ thống có khả năng nhận diện cảm xúc người dùng có thể giúp cải thiện trải nghiệm trong giáo dục, chăm sóc sức khỏe tâm thần, hỗ trợ khách hàng và nhiều lĩnh vực khác. Việc kết hợp giữa hình ảnh (khuôn mặt) và âm thanh (giọng nói) là một xu hướng mới nhằm tăng độ chính xác trong việc đánh giá trạng thái cảm xúc thực sự của người dùng.

1.1 Hiện trạng

Trên thế giới, nhiều công ty lớn như Microsoft, Google, Affectiva đã phát triển các hệ thống nhận diện cảm xúc dựa trên hình ảnh khuôn mặt hoặc giọng nói, ứng dụng trong chăm sóc khách hàng, giáo dục, xe hơi thông minh,... Tuy nhiên, các hệ thống này thường có chi phí cao, yêu cầu phần cứng mạnh và phụ thuộc vào nền tảng đám mây.

Tại Việt Nam, một số nghiên cứu tại các trường đại học cũng đã thử nghiệm nhận diện cảm xúc bằng ảnh hoặc âm thanh, nhưng chưa phổ biến và chưa có sản phẩm tích hợp cả hai nguồn dữ liệu. Việc ứng dụng trong thực tế còn hạn chế do khó triển khai trên thiết bị nhỏ gọn, giá rẻ như Raspberry Pi.

1.2 Các vấn đề cần giải quyết

- Làm sao nhận diện được cảm xúc của người dùng chỉ qua hình ảnh hoặc chỉ qua giọng nói, tùy điều kiện thực tế.
- Hệ thống cần hoạt động ổn định trên thiết bị nhỏ gọn, chi phí thấp như Raspberry Pi.
- Giao diện dễ sử dụng, hiển thị cảm xúc theo thời gian thực.
- Mô hình AI phải đủ nhẹ để xử lý cục bộ, không cần dùng đám mây.

1.3 Đề xuất giải pháp tổng quan

Đề xuất giải pháp tổng quan thể hiện ở bảng 1:

Thành phần	Giải pháp đề xuất
Phần cứng	- Máy tính Windows (đặt server Flask) - Raspberry Pi Camera Module V2 8MP - Microphone rời micro USB) - Raspberry Pi4 4GB
Nhận diện khuôn mặt	-Sử dụng OpenCV Haar Cascade (haarcascade_frontalface_default.xml) để phát hiện khuôn mặt nhanh, chính xác và nhẹ
Nhận diện cảm xúc qua hình ảnh	- Mô hình CNN tự xây dựng huấn luyện trên tập dữ liệu (ảnh grayscale 48x48) - Dự đoán 7 lớp cảm xúc: vui, buồn, tức giận, ngạc nhiên, sợ hãi, chán nản, bình thường
Nhận diện cảm xúc qua âm thanh	- Mô hình CNN dựa trên tập dữ liệu âm thanh - Dự đoán 7 lớp cảm xúc: vui, buồn, tức giận, ngạc nhiên, sợ hãi, chán nản, bình thường
Ứng dụng	– Xây dựng website nhận diện cảm xúc qua hình ảnh và âm thanh để đánh giá mức độ hài lòng của khách hàng sau khi sử dụng dịch vụ. – Hiển thị kết quả cảm xúc dưới dạng văn bản và biểu đồ theo thời gian thực – Có thể phân loại theo nguồn: hình ảnh / âm thanh
Server	-Viết bằng Flask API

Bảng 1: Bảng đề xuất giải pháp tổng quan

2. Giải pháp

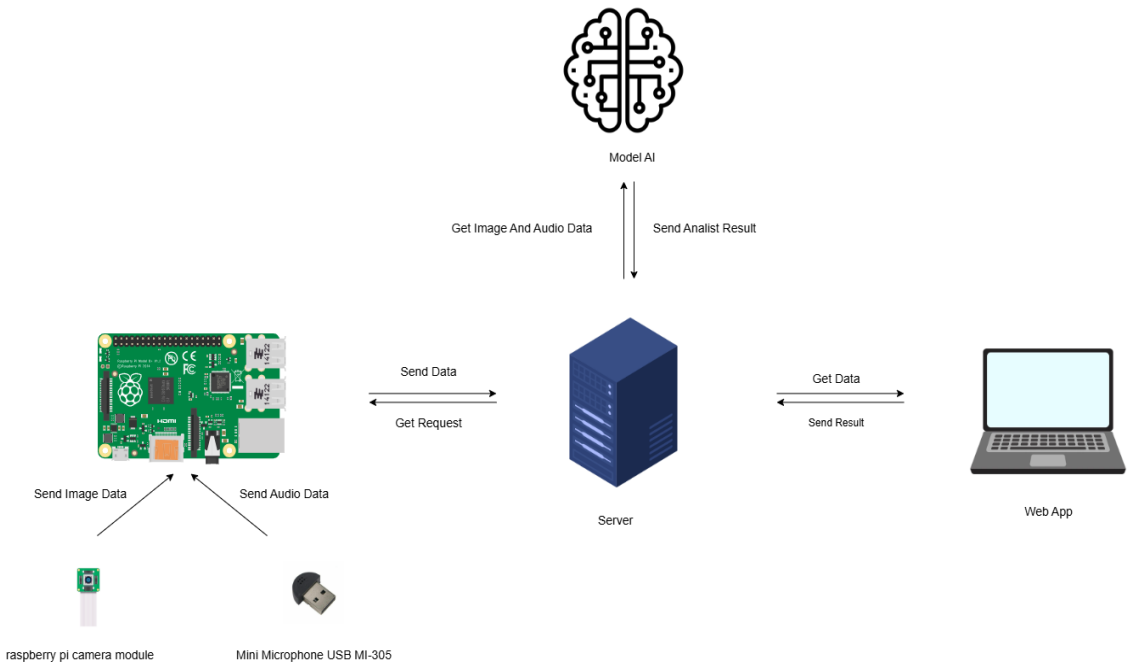
2.1 Giải pháp phần cứng và truyền thông

2.1.1. Sơ đồ tổng quan hệ thống

Hệ thống gồm hai thành phần chính: **client** và **server**.

- **Client (Raspberry Pi 4)**:
  - Gắn camera Pi để thu hình ảnh khuôn mặt.
  - Gắn micro USB để thu âm giọng nói.
  - Xử lý sơ bộ (nếu cần) và gửi dữ liệu hình ảnh, âm thanh lên server.
- **Server (máy tính)**:
  - Nhận dữ liệu từ Raspberry Pi.
  - Thực hiện xử lý nhận diện cảm xúc qua hình ảnh khuôn mặt và giọng nói.
  - Hiển thị kết quả nhận diện hoặc lưu trữ dữ liệu phân tích

Sơ đồ tổng quát hệ thống được thể hiện ở hình 1:



Hình 1: Sơ đồ tổng quát hệ thống

Raspberry pi 4 được thể hiện ở hình 2:

Hình 2. Rasberry Pi 4 Model B

Raspberry Pi 4 Model B

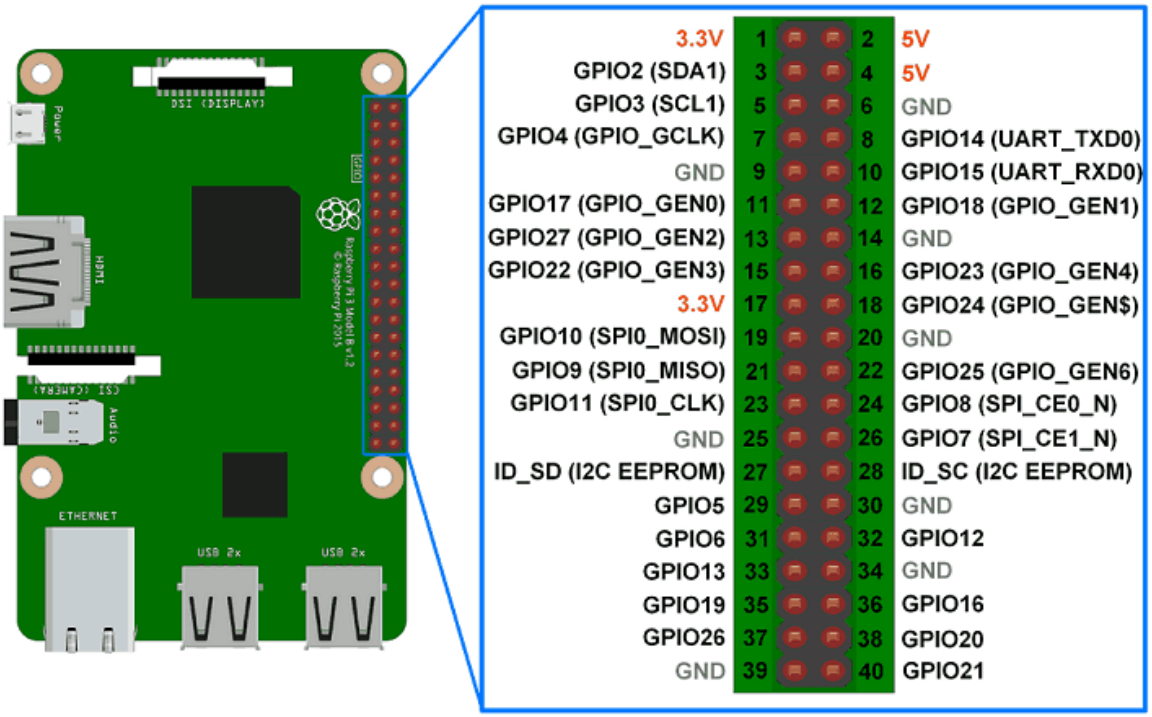
Raspberry Pi 4 Model B là phiên bản mới nhất trong dòng máy tính nhỏ gọn nổi tiếng Raspberry Pi, được thiết kế để mang lại hiệu năng cao hơn và hỗ trợ nhiều tính năng tiên tiến cho các ứng dụng giáo dục, nghiên cứu, và phát triển nhúng. Với khả năng xử lý mạnh mẽ hơn, bộ nhớ RAM đa dạng, và khả năng kết nối phong phú, Raspberry Pi 4 Model B mở rộng đáng kể tiềm năng sáng tạo cho người dùng.

Các thông số kỹ thuật chính của máy tính nhúng:

- Bộ vi xử lý: Broadcom BCM2711, lõi tứ ARM Cortex-A72 (ARM v8) 64-bit, tốc độ 1.5 GHz.
- Bộ nhớ RAM: Tùy chọn 2 GB, 4 GB, hoặc 8 GB LPDDR4-3200 SDRAM.
- Đồ họa: Broadcom VideoCore VI, hỗ trợ OpenGL ES 3.1, 4Kp60 HEVC video decode
- Lưu trữ: Khe cắm thẻ microSD (hỗ trợ UHS-I), USB 3.0, USB 2.0.
- Cổng USB: 2 x USB 3.0, 2 x USB 2.0.
- Kết nối mạng: Ethernet Gigabit, hỗ trợ PoE (yêu cầu bổ sung HAT PoE).

- Không dây: Wi-Fi 802.11 b/g/n/ac (2.4 GHz và 5.0 GHz), Bluetooth 5.0, BLE.
- Cổng màn hình: 2 x micro-HDMI, hỗ trợ độ phân giải lên đến 4Kp60.
- Cổng âm thanh: Jack âm thanh 3.5 mm (âm thanh và video composite), hỗ trợ HDMI audio.
- Nguồn điện: Cổng USB-C, cung cấp nguồn 5V/3A. [1]

Sơ đồ chân của Raspberry Pi 4 được thể hiện ở hình 3:



Hình 3: Sơ đồ chân của Rasperry Pi 4

Giao thức truyền thông và các chân giao tiếp:

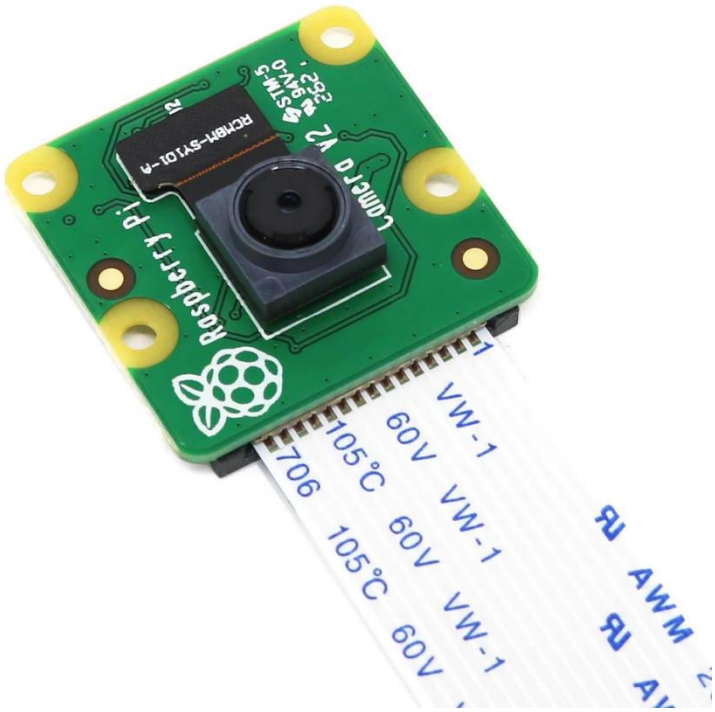
- GPIO: 40 chân GPIO tiêu chuẩn, cung cấp giao diện để kết nối với các thiết bị phần cứng bên ngoài như cảm biến, đèn LED, hoặc module giao tiếp.
- UART: Universal Asynchronous Receiver/Transmitter, dùng để giao tiếp nối tiếp với các thiết bị khác như mô-đun GPS hoặc module không dây.
- SPI: Serial Peripheral Interface, một giao thức truyền thông nối tiếp tốc độ cao dùng để kết nối với các thiết bị như màn hình OLED hoặc cảm biến.
- I2C: Inter-Integrated Circuit, một giao thức truyền thông nối tiếp cho phép kết nối với các thiết bị như bộ giải mã ADC, cảm biến nhiệt độ, hoặc

EEPROM.

- Ethernet: Hỗ trợ truyền thông mạng có dây với tốc độ Gigabit Ethernet, giúp tăng tốc độ truyền dữ liệu và độ tin cậy khi kết nối mạng.
- USB: Các cổng USB 3.0 và USB 2.0 cho phép kết nối với nhiều loại thiết bị ngoại vi như ổ cứng ngoài, bàn phím, chuột, và thiết bị lưu trữ USB.

Raspberry Pi Camera Module

Raspberry Pi Camera Module được thể hiện ở hình 4:



Hình 4: Raspberry Pi Camera Module V2 8M

a. Mô tả chung:

Camera Module V2 là một mô-đun camera chính hãng của Raspberry Pi Foundation, sử dụng cảm biến hình ảnh Sony IMX219 có độ phân giải 8 megapixel. Thiết bị được thiết kế để kết nối trực tiếp với bo mạch Raspberry Pi thông qua cổng CSI (Camera Serial Interface), hỗ trợ chụp ảnh và quay video chất lượng cao.

b. Thông số kỹ thuật:

Thông số kỹ thuật của camera được thể hiện ở bảng 2:

Thông số	Giá trị
Cảm biến	Sony IMX219
Độ phân giải ảnh	8 Megapixel (3280 × 2464)
Giao tiếp	CSI (Camera Serial Interface)
Ống kính	Fixed focus (tiêu cự cố định)
Khả năng quay video	1080p30, 720p60, 640x480p90
Kích thước	23.86 x 25 mm
Trọng lượng	Khoảng 3g

Bảng 2: Bảng thông số kỹ thuật của Raspberry Pi Camera

c. Chức năng:

- Ghi lại hình ảnh và video để phục vụ xử lý hình ảnh trong các ứng dụng như: nhận diện khuôn mặt, nhận diện vật thể, theo dõi đối tượng.
- Có thể sử dụng trong các hệ thống giám sát, robot tự hành, nhận diện cảm xúc, v.v.

Mini Microphone USB MI-305 Raspberry Pi

Mini Microphone USB MI-305 được thể hiện ở hình 5:

Hình 5: Mini Microphone USB MI-305 Raspb

a. Mô tả chung:

Microphone USB MI-305 là một loại micro nhỏ gọn, cắm trực tiếp vào cổng USB và không yêu cầu driver. Linh kiện này tương thích tốt với Raspberry Pi, đặc biệt trong các ứng dụng xử lý âm thanh như nhận diện giọng nói, ghi âm, hoặc điều khiển bằng giọng nói.

b. Thông số kỹ thuật:

Thông số kỹ thuật của Mini Microphone USB MI-305 được thể hiện ở bảng 3:

Thông số	**Giá trị**
Loại micro	Condenser Microphone
Giao tiếp	USB 2.0
Hướng thu âm	Đơn hướng (Unidirectional)
Độ nhạy	-30 ± 2dB

Tần số đáp ứng	100Hz – 16kHz
Tỷ lệ tín hiệu trên nhiễu	≥ 60 dB
Điện áp hoạt động	5V DC (qua USB)
Cắm & chạy	Không cần cài đặt driver
Kích thước	Rất nhỏ gọn (mini size ~ 2.5cm)

Bảng 3: Bảng thông số kỹ thuật của Mini Microphone USB MI-305

c. Chức năng:

- Thu âm giọng nói với chất lượng tương đối tốt.
- Dùng làm thiết bị đầu vào âm thanh cho các hệ thống điều khiển bằng giọng nói hoặc nhận diện cảm xúc qua giọng nói.
- Kết hợp tốt với Raspberry Pi và thư viện âm thanh như sounddevice, pyaudio, speech_recognition.

2.1.2. Sơ đồ kết nối linh kiện phần cứng

- **Raspberry Pi 4**:

  - Kết nối camera Pi qua cổng CSI (Camera Serial Interface).

  - Micro USB được cắm vào một trong các cổng USB của Raspberry Pi.

  - Kết nối mạng (WiFi hoặc Ethernet) để truyền dữ liệu đến server.

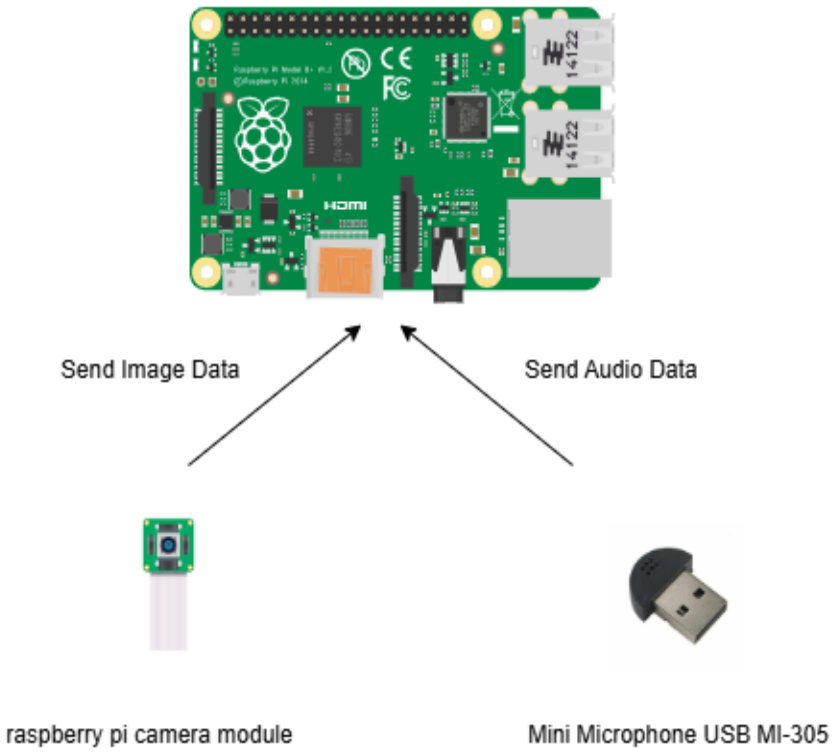
Chi tiết kết nối phần cứng được thể hiện qua bảng 4:

Thiết bị	**Cổng kết nối trên Raspberry Pi 4**	**Ghi chú**
Camera Pi	Cổng CSI (ribbon cable)	Cần kích hoạt trong cấu hình hệ thống
Micro USB	Cổng USB 2.0 hoặc USB 3.0	Raspberry Pi tự động nhận dạng mic USB

Server PC	Kết nối mạng (LAN/WiFi)	Giao tiếp qua TCP/IP hoặc HTTP/WebSocket

Bảng 4: chi tiết kết nối phần cứng

Sơ đồ kết nối linh kiện phần cứng được thể hiện ở hình 6:



Hình 6: Sơ đồ kết nối linh kiện phần cứng

2.1.3 Truyền thông

Giao thức truyền thông:

Socket Programming:

Socket là một phương thức truyền thông mạng dựa trên mô hình client-server, cho phép thiết lập kết nối giữa hai thiết bị thông qua địa chỉ IP và cổng giao tiếp (port). Socket hoạt động theo cơ chế truyền dữ liệu song công (bi-directional), thích hợp cho các ứng dụng thời gian thực. Trong đồ án này, socket TCP được sử dụng để truyền dữ liệu liên tục từ Raspberry Pi (client) đến máy chủ xử lý (server).

Socket hỗ trợ phân biệt loại dữ liệu thông qua header đặc trưng (b'IMG' cho ảnh và b'AUD' cho âm thanh), giúp hệ thống xử lý chính xác và đồng thời cả hai luồng dữ liệu. Đây là phương pháp truyền thông đơn giản nhưng hiệu quả, đặc biệt trong các hệ thống nhúng không yêu cầu giao thức tầng cao như HTTP hoặc MQTT.

HTTP (Hypertext Transfer Protocol):

HTTP là giao thức truyền thông ở tầng ứng dụng trong mô hình OSI, được sử dụng rộng rãi để truyền tải dữ liệu trên nền tảng web. Trong đồ án, HTTP được ứng dụng để tạo giao tiếp giữa server và các thành phần phụ trợ, thông qua các RESTful API như: lấy kết quả phân tích cảm xúc, điều khiển trạng thái hệ thống, hoặc nhận tín hiệu khởi động từ phía giao diện người dùng. HTTP sử dụng các phương thức chuẩn như GET, POST, giúp việc tích hợp giữa thiết bị và giao diện điều khiển trở nên dễ dàng và có tính mở rộng cao.

API (Application Programming Interface):

API là giao diện lập trình ứng dụng cho phép các thành phần trong hệ thống – từ phần cứng đến phần mềm – giao tiếp với nhau. Trong đồ án, API đóng vai trò trung gian giữa thiết bị Raspberry Pi và server điều khiển. Các loại API sử dụng bao gồm:

- **RESTful API**: Giao tiếp qua HTTP để truyền nhận dữ liệu xử lý ảnh, âm thanh, trạng thái cảm xúc.
- **WebSocket API**: Truyền dữ liệu thời gian thực như tín hiệu khởi động, trạng thái hoạt động hoặc thông báo cảm xúc.
- **Local API**: Giao tiếp nội bộ giữa các tiến trình server như bộ xử lý hình ảnh và bộ xử lý âm thanh.

FastAPI:

FastAPI là một framework hiện đại và hiệu suất cao dùng để xây dựng các RESTful API với Python. Dựa trên chuẩn ASGI, FastAPI hỗ trợ xử lý bất đồng bộ (async/await) và tự động sinh tài liệu API với Swagger UI. Trong đồ án, FastAPI được sử dụng để triển khai server điều khiển, cung cấp các endpoint như:

- Nhận tín hiệu bắt đầu hoặc dừng truyền dữ liệu từ Raspberry Pi.
- Gửi kết quả nhận diện cảm xúc khuôn mặt và giọng nói.
- Kết nối với hệ thống giám sát trung tâm để hiển thị tình trạng người dùng theo thời gian thực.

Tóm tắt luồng truyền thông trong hệ thống

Hệ thống sử dụng mô hình client–server, trong đó Raspberry Pi đóng vai trò client thu thập dữ liệu và gửi đến server để xử lý. Luồng truyền thông cụ thể như sau:

1. **Tại client (Raspberry Pi)**

- Camera ghi lại hình ảnh khuôn mặt → chuyển thành dữ liệu byte → gắn tiền tố b'IMG' → gửi qua TCP socket đến server.
  - Micro USB ghi lại âm thanh giọng nói → chia thành các khung nhỏ → gắn tiền tố b'AUD' → gửi qua TCP socket đến server.
- Client liên tục gửi song song cả hai luồng ảnh và âm thanh trong thời gian thực.

2. **Tại server (máy tính xử lý):**

- Lắng nghe socket và phân biệt luồng dữ liệu qua tiền tố b'IMG' hoặc b'AUD'.
- Dữ liệu ảnh được xử lý bằng mô hình nhận diện cảm xúc khuôn mặt (ví dụ: DeepFace).
- Dữ liệu âm thanh được xử lý bằng mô hình nhận diện cảm xúc từ giọng nói (ví dụ: SpeechEmotionRecognition).
- Kết quả cảm xúc được gửi về giao diện người dùng thông qua API RESTful hoặc WebSocket.

3. **Giao tiếp điều khiển (hai chiều):**

- Server sử dụng FastAPI để cung cấp REST API, nhận tín hiệu điều khiển từ người dùng (bắt đầu/dừng thu thập).
- Server có thể phản hồi lại client qua socket hoặc trigger một hành động xử lý.

2.2 Giải pháp AI/KHDL

2.2.1. Giải pháp nhận diện cảm xúc qua hình ảnh (khuôn mặt)

Trong đồ án này, nhóm phát triển một mô hình học sâu (Deep Learning) sử dụng mạng nơ-ron tích chập (CNN) để nhận diện cảm xúc khuôn mặt người từ ảnh tĩnh. Mục tiêu là phân loại khuôn mặt thành 7 cảm xúc cơ bản: giận dữ (angry), khinh thường (disgust), sợ hãi (fear), hạnh phúc (happy), buồn bã (sad), ngạc nhiên (surprise), bình thường (neutral).

Mô hình được huấn luyện từ đầu (train from scratch) trên tập dữ liệu gồm một tập ảnh grayscale kích thước 48×48 pixel, thường dùng cho các bài toán nhận diện cảm xúc trong học thuật. Việc huấn luyện được thực hiện trên GPU và tối ưu bằng kỹ thuật như data augmentation, batch normalization, và early stopping để tăng độ chính xác và giảm overfitting.

Quy trình nhận diện cảm xúc khuôn mặt được thực hiện như sau:

Bước 1: Phát hiện khuôn mặt từ ảnh đầu vào

Nhóm sử dụng OpenCV hoặc MediaPipe để phát hiện khuôn mặt trong khung hình. Khuôn mặt được cắt, chuyển sang ảnh xám (grayscale) và resize về 48×48 pixels.

Bước 2: Dự đoán cảm xúc

Ảnh được đưa vào mô hình CNN đã huấn luyện để xuất ra xác suất của 7 lớp cảm xúc. Cảm xúc có xác suất cao nhất sẽ là đầu ra của hệ thống.

Bước 3: Hiển thị kết quả

Cảm xúc được hiển thị trực tiếp trên ảnh đầu vào (camera), cùng với khung bao quanh khuôn mặt.

2.2.1.1 . Mục đích của mô hình nhận diện cảm xúc

Hệ thống giúp tự động xác định trạng thái cảm xúc của người dùng trong các tình huống như học tập, thi cử, hoặc giám sát hành vi, đánh giá mức độ hài lòng khi người dùng sử dụng dịch vụ…. Dữ liệu cảm xúc có thể được sử dụng để đưa ra cảnh báo, thống kê, hoặc điều chỉnh môi trường học tập/phòng thi phù hợp hơn.

2.2.1.2 Giới thiệu mô hình CNN nhận diện cảm xúc

a. Cấu trúc tổng quát

Mô hình học các đặc trưng phi tuyến từ khuôn mặt qua nhiều lớp tích chập (Conv2D) và pooling, kết hợp dropout để giảm overfitting. Sau đó, đặc trưng được gom lại bằng GlobalAveragePooling và đưa vào các lớp Dense để phân loại cảm xúc.

Trong mô hình, nhóm sử dụng một **residual block** – kỹ thuật vay mượn từ ResNet – nhằm giúp mô hình học tốt hơn ở tầng sâu bằng cách cộng đầu vào ban đầu với đầu ra của các lớp tích chập bên trong block. Điều này giúp giảm hiện tượng mất thông tin (information degradation) hoặc biến mất gradient (vanishing gradient).

Kiến trúc mô hình CNN nhận diện cảm xúc khuôn mặt được thể hiện qua bảng 5:

Tên lớp	Chi tiết	Kích thước đầu ra
Input	Ảnh grayscale 48×48×1	48×48×1
Conv2D + BN + MaxPool + Dropout	64 filters, kernel 3×3, ReLU + chuẩn hóa và dropout 0.2	24×24×64
Conv2D×2 + BN + MaxPool + Dropout	128 filters × 2, dropout 0.3	12×12×128
Residual Block	Conv2D×2 với 256 filters + skip connection (1×1 conv)	6×6×256
MaxPooling2D + Dropout	Pooling + dropout 0.4	3×3×256
GlobalAveragePooling2D	Gom toàn bộ đặc trưng không gian thành vector	1×1×256 → 256
Dense + Dropout	Lớp ẩn 256 node, ReLU + dropout 0.5	256
Output Dense	Softmax với 7 node tương ứng 7 cảm xúc	7

Bảng 5. Kiến trúc mô hình CNN nhận diện cảm xúc khuôn mặt

b. Các kỹ thuật tối ưu

- Data Augmentation: Lật ảnh ngang, zoom, dịch ảnh để tăng tính đa dạng của dữ liệu huấn luyện.

- Batch Normalization: Giúp ổn định và tăng tốc quá trình huấn luyện.

- Dropout: Tránh overfitting khi huấn luyện trên tập dữ liệu nhỏ như FER-2013.

- Early Stopping + Model Checkpoint: Dừng sớm nếu mô hình không cải thiện và lưu mô hình tốt nhất.

- Optimizer: Adam với learning rate được điều chỉnh linh hoạt bằng ReduceLROnPlateau

c. Đặc điểm và ưu điểm của giải pháp

- Mô hình có thể triển khai hiệu quả trên các thiết bị cấu hình trung bình nhờ vào kiến trúc gọn nhẹ.
- Việc sử dụng residual block giúp tăng khả năng biểu diễn mà không làm tăng độ phức tập huấn luyện quá nhiều.
- Nhờ augmentation, mô hình có khả năng tổng quát tốt và nhận diện cảm xúc trong điều kiện ảnh đa dạng (độ sáng, tư thế,…).
- Hệ thống đạt độ chính xác >85% trên tập kiểm thử, đáp ứng mục tiêu đề ra của đồ án.

2.2.2. Giải pháp nhận diện cảm xúc qua giọng nói (âm thanh)

Bên cạnh khuôn mặt, giọng nói cũng là một chỉ số quan trọng thể hiện cảm xúc con người. Trong đồ án này, nhóm xây dựng một hệ thống nhận diện cảm xúc từ giọng nói sử dụng kỹ thuật học sâu (Deep Learning) với đầu vào là các đoạn âm thanh ngắn đã được xử lý thành biểu diễn đặc trưng MFCC.

Mục tiêu của mô hình là phân loại cảm xúc người nói thành 7 loại cơ bản: giận dữ (angry), khinh thường (disgust), sợ hãi (fear), hạnh phúc (happy), buồn bã (sad), ngạc nhiên (surprise), và bình thường (neutral).

Quy trình nhận diện cảm xúc qua âm thanh gồm các bước:

Bước 1: Trích xuất đặc trưng MFCC từ âm thanh

Đoạn âm thanh đầu vào được xử lý bằng thư viện librosa để chuẩn hóa tốc độ lấy mẫu về 16.000 Hz và kéo dài (hoặc cắt ngắn) sao cho độ dài cố định là 3 giây (48000 mẫu). Từ tín hiệu này, đặc trưng MFCC với 40 hệ số được trích xuất và chuẩn hóa về phân phối chuẩn, sau đó reshape để phù hợp với mô hình đầu vào.

Bước 2: Huấn luyện mô hình CNN nhận diện cảm xúc

Mô hình sử dụng mạng CNN gồm nhiều lớp tích chập 2D (Conv2D) và pooling để học các đặc trưng cục bộ từ ảnh MFCC. Lớp cuối sử dụng Dense + Softmax để phân loại ra 7 cảm xúc.

Bước 3: Dự đoán cảm xúc từ âm thanh mới

Đoạn âm thanh được xử lý như ở bước 1, sau đó đưa qua mô hình đã huấn luyện. Kết

quả trả về là xác suất thuộc về mỗi lớp cảm xúc, và lớp có xác suất cao nhất được chọn làm đầu ra.

2.2.2.1. Mục đích của mô hình nhận diện cảm xúc từ giọng nói

Mô hình giúp xác định cảm xúc của người nói trong các đoạn hội thoại, tình huống giao tiếp, hoặc trong môi trường học tập và làm việc từ xa. Hệ thống có thể ứng dụng vào việc theo dõi mức độ căng thẳng, sự hài lòng của học viên, hoặc phân tích tổng đài chăm sóc khách hàng, chatbot,…

2.2.2.2. Giới thiệu mô hình CNN xử lý tín hiệu âm thanh

a. Cấu trúc tổng quát

Dữ liệu âm thanh đầu vào được chuyển thành ảnh MFCC và đưa vào mạng CNN gồm 4 tầng Conv2D tăng dần số filters (32→64→128→256), xen kẽ với các kỹ thuật tối ưu như:

- **LeakyReLU:** giúp mô hình học tốt hơn trên dữ liệu âm thanh không tuyến tính.
- **BatchNormalization:** ổn định quá trình huấn luyện.
- **MaxPooling2D:** giảm chiều kích thước, tăng trích xuất đặc trưng.
- **Dropout:** giảm overfitting.

Sau khi trích xuất đặc trưng, toàn bộ tensor được flatten và đưa qua một lớp Dense lớn (512 node) rồi đến lớp softmax để phân loại.

Kiến trúc mô hình CNN nhận diện cảm xúc qua giọng nói được thể hiện qua bảng 6:

Tên lớp	Chi tiết	Kích thước đầu ra
Input	MFCC (40, ~94, 1)	40×94×1
Conv2D + LeakyReLU + BN + Pool + Dropout	32 filters, 3×3 kernel, dropout 0.2	~20×47×32
Conv2D + LeakyReLU + BN + Pool + Dropout	64 filters, dropout 0.3	~10×23×64

Conv2D + LeakyReLU + BN + Pool + Dropout	128 filters, dropout 0.4	~5×11×128
Conv2D + LeakyReLU + BN + Pool + Dropout	256 filters, dropout 0.4	~2×5×256
Flatten	Vector hóa đặc trưng	~2560
Dense + LeakyReLU + Dropout	512 node, dropout 0.5	512
Output Dense	Softmax với 7 node tương ứng 7 cảm xúc	7

Bảng 6. Kiến trúc mô hình CNN nhận diện cảm xúc qua giọng nói

b. Các kỹ thuật tối ưu

- Label Smoothing: làm mềm nhãn giúp giảm độ tự tin sai của mô hình.
- Early Stopping: dừng sớm khi mô hình không cải thiện trên tập validation.
- Model Checkpoint: lưu lại mô hình có độ chính xác cao nhất.
- ReduceLROnPlateau: giảm learning rate khi validation loss không giảm.

c. Đặc điểm và ưu điểm của giải pháp

- Mô hình đơn giản, dễ triển khai nhưng vẫn đạt độ chính xác cao trên tập kiểm thử (≥ 75%).
- Không cần dữ liệu video hay hình ảnh – chỉ cần audio đầu vào (có thể lấy từ micro).
- Có thể kết hợp với mô hình nhận diện khuôn mặt để xây dựng hệ thống đa modal, giúp tăng độ chính xác tổng thể trong nhận diện cảm xúc người dùng.

2.3 Giải pháp phần mềm

2.3.1 Phát triển bài toán

a. **Bài toán đặt ra**

Trong lĩnh vực vận tải hành khách, việc đánh giá mức độ hài lòng của khách hàng

thường phụ thuộc vào khảo sát thủ công, dễ gây thiếu khách quan và mất thời gian. Do đó, hệ thống được xây dựng với mục tiêu:

- Tự động thu thập và phân tích cảm xúc của hành khách trong suốt chuyến đi thông qua hình ảnh khuôn mặt và âm thanh giọng nói.
  - Đánh giá mức độ hài lòng tổng thể của hành khách dựa trên tần suất và tỷ lệ xuất hiện của các cảm xúc tích cực (happy, neutral) hoặc tiêu cực (angry, sad, disgust, etc).
- Cung cấp giao diện web cho tài xế hoặc quản lý để xem thống kê chi tiết theo từng chuyến xe, phục vụ cho cải tiến dịch vụ.

b. Mục tiêu phát triển

- Phát triển hệ thống phần mềm có khả năng nhận diện cảm xúc theo thời gian thực trên thiết bị đặt trong xe.
- Tự động lưu trữ dữ liệu cảm xúc theo từng mã chuyến đi (trip ID) và thời gian, phục vụ phân tích và đánh giá sau.
- Tính toán mức độ hài lòng cuối chuyến dựa trên phân tích cảm xúc và hiển thị kết quả trên web.
- Không can thiệp chủ động từ hành khách, đảm bảo sự tự nhiên và khách quan trong quá trình thu thập dữ liệu.

2.3.2 Công nghệ sử dụng

Công nghệ sử dụng trong dự án được thể hiện qua bảng 7:

Thành phần	**Công nghệ sử dụng**
Ngôn ngữ chính	Python
Backend Framework	Flask
Frontend	HTML, CSS, JavaScript
Mô hình nhận diện hình ảnh	Keras (emotion_cnn_best16.keras)
Mô hình nhận diện âm thanh	Keras (audio_emotion_audio5.keras)

Xử lý ảnh	OpenCV
Xử lý âm thanh	Librosa
Tính năng học máy	TensorFlow / Keras
Lưu trữ dữ liệu	Pandas (CSV)
Biểu đồ thống kê	matplotlib / seaborn (nếu dùng)

Bảng 7: Bảng công nghệ sử dụng

2. Thiết bị Raspberry

Là thiết bị phần cứng tại hiện trường, có chức năng thu thập dữ liệu cảm xúc.

- Gửi dữ liệu hình ảnh và gửi dữ liệu âm thanh: Truyền dữ liệu đến hệ thống xử lý để phân tích cảm xúc.

3. Hệ thống (nội bộ)

Đảm nhiệm việc xử lý dữ liệu và lưu trữ thông tin.

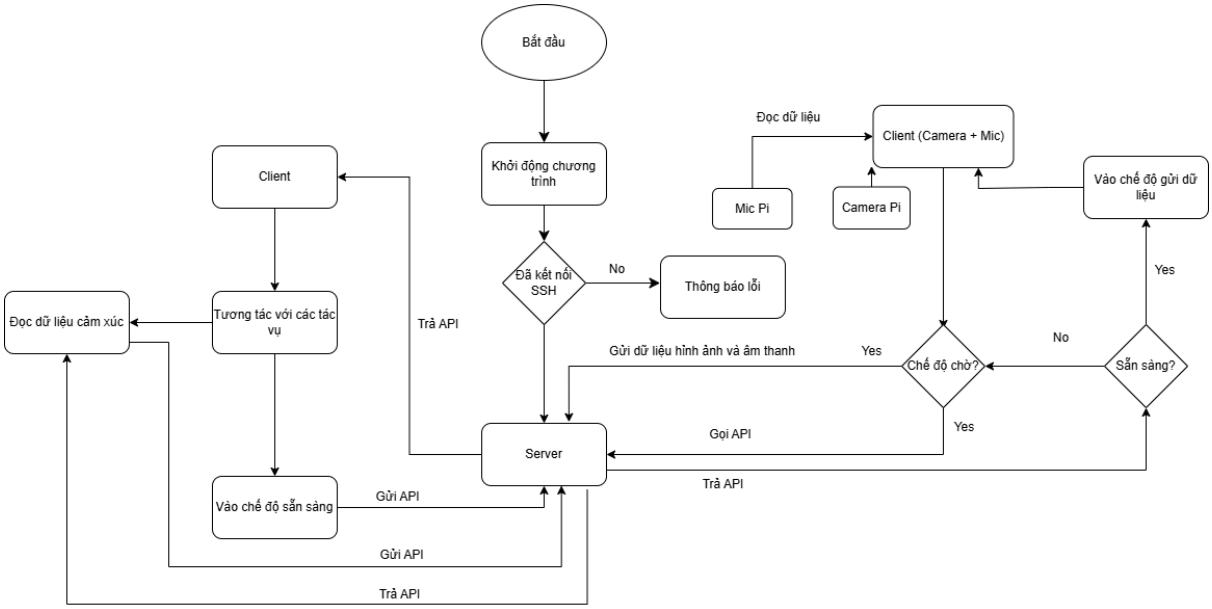
Hệ thống xử lý: Phân tích dữ liệu nhận được để xác định cảm xúc.

- Bao gồm: Nhận dữ liệu âm thanh và hình ảnh (<<include>>).
- Có thể kích hoạt lưu trữ dữ liệu (<<extension>> đến Hệ thống lưu trữ).

Hệ thống lưu trữ: Lưu kết quả phân tích để phục vụ truy xuất, thống kê và hiển thị cho người dùng.

2.3.5 Sơ đồ khối hệ thống

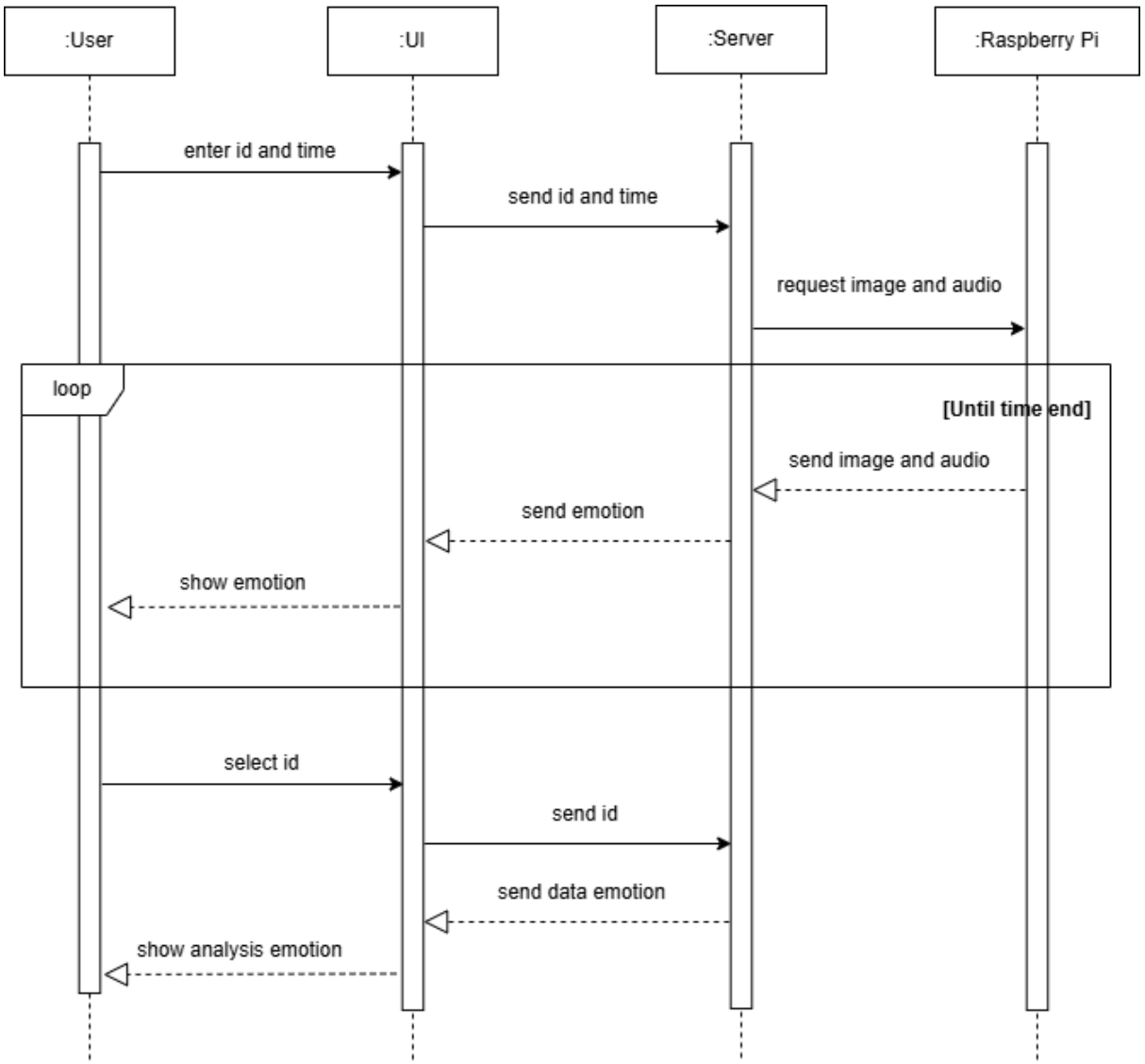
Sơ đồ khối hệ thống được thể hiện ở hình 7:



Hình 7. Sơ đồ khối hệ thống

2.3.6 Quy trình thực thi hệ thống

Luồng thực hiện của hệ thống được thể hiện ở hình 8:



Hình 8. Sơ đồ tuần tự luồng thực hiện của hệ thống.

Sơ đồ tuần tự trên mô tả luồng tương tác giữa người dùng, giao diện người dùng (UI), máy chủ (Server), và thiết bị Raspberry Pi trong hệ thống nhận diện cảm xúc thời gian thực. Ban đầu, người dùng nhập mã định danh (ID) và khoảng thời gian muốn theo dõi. Giao diện sẽ gửi thông tin này đến máy chủ, sau đó server yêu cầu Raspberry Pi bắt đầu thu thập và gửi dữ liệu hình ảnh và âm thanh. Trong suốt khoảng thời gian được chỉ định, Raspberry Pi liên tục truyền dữ liệu đến server, nơi các mô hình học sâu sẽ xử lý và phân tích cảm xúc từ khuôn mặt và giọng nói. Kết quả được gửi về UI để hiển thị cho người dùng theo thời gian thực. Sau khi quá trình theo dõi kết thúc, người dùng có thể chọn lại ID để xem lại dữ liệu

cảm xúc đã lưu, thông qua phân tích tổng hợp từ server. Quy trình này đảm bảo sự liên tục trong việc thu thập, xử lý và hiển thị cảm xúc một cách trực quan và hiệu quả.

4. Kết quả

3.1. Nhận diện cảm xúc từ khuôn mặt

3.1.1. Tập dữ liệu

Tên tập dữ liệu: Data

- Định dạng ảnh: 48×48 pixel, ảnh xám (grayscale)
- Số lớp cảm xúc: 7 (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral)
- Số lượng ảnh: khoảng 6414 ảnh khuôn mặt
- Cấu trúc tập dữ liệu:

- Tỉ lệ chia dữ liệu:

    - 70% ảnh để huấn luyện (training)
    - 15% ảnh để xác thực (validation)
    - 15% ảnh để kiểm tra (testing)

Các ảnh được phân bố vào 7 thư mục con tương ứng với từng nhãn cảm xúc

- Nguồn dữ liệu:

- Được thu thập từ các hình ảnh gắn nhãn trên Internet và nạp sẵn trong cuộc thi Kaggle “Challenges in Representation Learning”
- Ảnh có biểu cảm rõ ràng nhưng độ phân giải thấp (48x48), phù hợp với huấn luyện mô hình nhẹ

Ảnh được căn chỉnh và cắt sao cho khuôn mặt nằm chính giữa khung hình. Dữ liệu có sự đa dạng nhất định về tuổi tác, giới tính, và biểu cảm, tuy nhiên cũng có một số ảnh bị nhiễu, không rõ nét.

Để cải thiện hiệu quả mô hình, tập huấn luyện được áp dụng data augmentation gồm:

- Xoay nhẹ góc ảnh (±10 độ)

- Dịch chuyển trái/phải/trên/dưới (tối đa 20%)

- Zoom in/out (tối đa 10%)

- Lật ngang ảnh

- Chuẩn hóa pixel về khoảng [0, 1]

3.2.2. Huấn luyện mô hình

Kiến trúc mô hình

Mô hình CNN được thiết kế chuyên biệt cho bài toán nhận diện cảm xúc với ảnh grayscale kích thước 48x48. Cấu trúc bao gồm:

- Các lớp Conv2D kết hợp với BatchNormalization, MaxPooling2D và Dropout

- Một khối residual giúp tăng khả năng học đặc trưng sâu hơn

- Lớp GlobalAveragePooling2D và các lớp Dense để phân loại 7 cảm xúc

Cấu hình huấn luyện

- **Input shape**: (48, 48, 1)

- **Batch size**: 64

- **Số lớp đầu ra**: 7 (ứng với 7 cảm xúc)

- **Optimizer**: Adam

- **Loss function**: CategoricalCrossentropy có smoothing

- **Số epoch**: 150

- **Callbacks**:

    + Lưu mô hình tốt nhất (ModelCheckpoint)

    + Dừng sớm khi overfitting (EarlyStopping)

    + Giảm learning rate khi validation loss không giảm (ReduceLROnPlateau)

Theo dõi và đánh giá

- Mô hình được huấn luyện và đánh giá dựa trên accuracy và validation accuracy trên cả tập huấn luyện và kiểm tra.

- Kết quả được biểu diễn bằng đồ thị train accuracy và validation qua các epoch để quan sát quá trình học.

Kết quả huấn luyện:

Sau khi huấn luyện mô hình trên tập dữ liệu cảm xúc gồm ảnh grayscale kích thước nhỏ, mô hình đạt kết quả khả quan với độ chính xác cao trên tập kiểm tra. Cụ thể:

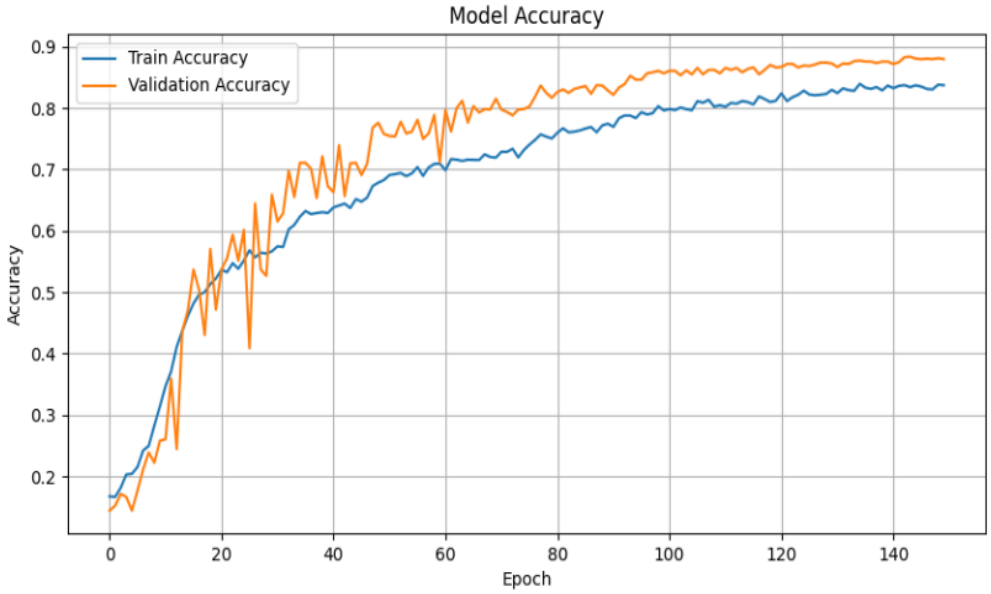
Validation Accuracy: ≈ *86-87%*

Train Accuracy: ≈ *84-85%*

- **Mô hình tốt nhất được lưu tại**:

E:/Python/PBL5/models/emotion_cnn_best15.keras

Kết quả train model cảm xúc hình ảnh được thể hiện ở hình 9:



Hình 9 . Kết quả train model cảm xúc từ hình ảnh

3.2.3. Kết quả nhận diện

- Mô hình nhận diện tốt các cảm xúc phổ biến như *vui*, *buồn*, *giận dữ*.

- Với các cảm xúc trừu tượng hơn như *ngạc nhiên* hay *sợ hãi*, mô hình vẫn phân loại tốt nhưng dễ nhầm lẫn trong một số trường hợp ánh sáng yếu hoặc khuôn mặt bị nghiêng.

- Mô hình hoạt động tốt trên ảnh grayscale, giúp giảm tải xử lý và phù hợp với các thiết bị hạn chế phần cứng.

Việc thiết kế mô hình CNN có khối residual và áp dụng tăng cường dữ liệu đã cải thiện đáng kể khả năng nhận diện cảm xúc trên ảnh nhỏ. Mô hình nhẹ, độ chính xác cao và sẵn sàng triển khai trong các hệ thống như giám sát cảm xúc học sinh trong lớp học, chăm sóc sức khỏe tinh thần, hoặc chatbot thông minh.

3.2. Nhận diện cảm xúc từ giọng nói

3.2.1. Tập dữ liệu

- Tên tập dữ liệu: data_audio
- Định dạng: file .wav âm thanh mono
- Tần số mẫu (sampling rate): 16.000 Hz
- Thời lượng mỗi đoạn âm thanh: 3 giây
- Đặc trưng sử dụng: MFCC (Mel Frequency Cepstral Coefficients) với 40 hệ số
- Số lớp cảm xúc: 7 (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral)
- Tỉ lệ chia dữ liệu:

    70% dùng để huấn luyện (train)

    15% xác thực (validation)

    15% kiểm tra (test)

- Tiền xử lý dữ liệu:
    - Âm thanh được cắt hoặc đệm để đủ 3 giây (48000 mẫu)
    - Trích xuất 40 hệ số MFCC
    - Chuẩn hóa đặc trưng (zero-mean, unit-variance)
    - Chuyển đổi nhãn về dạng one-hot

3.2.2. Huấn luyện mô hình

Kiến trúc mô hình

Mô hình CNN 2D được thiết kế để xử lý biểu diễn MFCC dưới dạng ảnh (40 x time_steps x 1)

Bao gồm:

- Các lớp Conv2D với LeakyReLU, BatchNormalization, MaxPooling2D, Dropout
- Tổng cộng 4 tầng chồng nhau, tăng dần số filters từ 32 → 256
- Flatten đầu ra và kết nối với lớp Dense 512 neurons

- Lớp đầu ra Dense với softmax phân loại 7 cảm xúc

Cấu hình huấn luyện:

- Input shape: (40, time_steps, 1) (MFCC đầu vào)

- Batch size: 32

- Số lớp đầu ra: 7

- Optimizer: Adam với learning_rate=1e-3

- Loss function: CategoricalCrossentropy (có label_smoothing=0.1)

- Epochs: 150

- Callbacks:

+ ModelCheckpoint: lưu mô hình tốt nhất tại E:/Python/PBL5/models/audio_emotion_audio5.keras

+ EarlyStopping: dừng sớm nếu validation không cải thiện sau 15 epoch

+ ReduceLROnPlateau: giảm learning rate khi validation loss không giảm

Kết quả huấn luyện

Sau khi huấn luyện, mô hình đạt:

Train Accuracy: ~81%

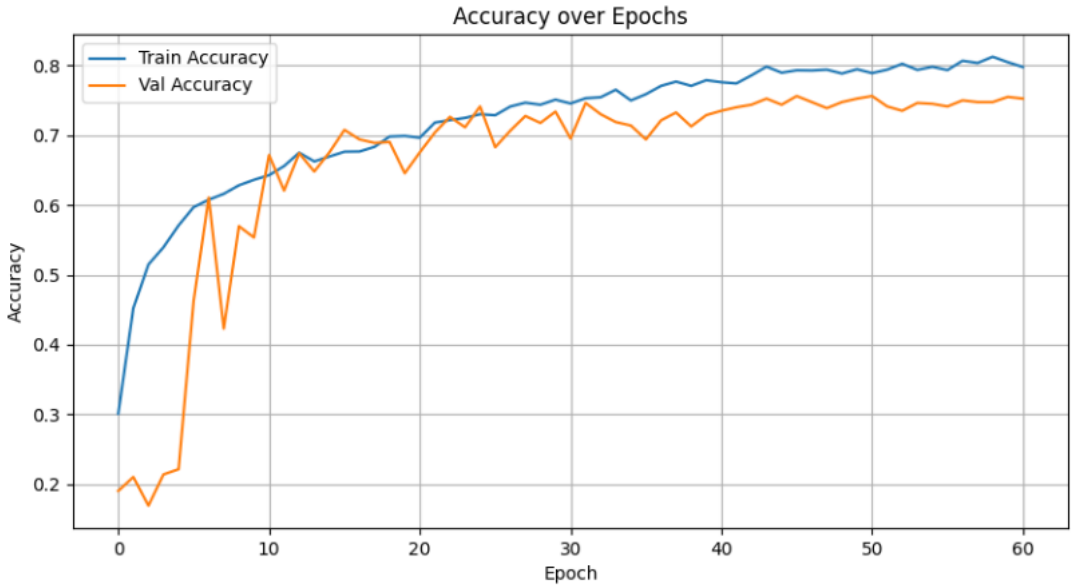
Validation Accuracy: ~74–75%

Test Accuracy: ~75.04%

Kết quả này cho thấy mô hình học tốt đặc trưng từ MFCC và có khả năng tổng quát hóa khá tốt.

Biểu đồ Accuracy qua các Epoch:

Kết quả train model cảm xúc hình ảnh được thể hiện qua hình 10:



Hình 10. Kết quả train model cảm xúc từ hình ảnh

3.2.3. Kết quả nhận diện

- Mô hình nhận diện khá tốt các cảm xúc phổ biến qua giọng nói như **vui**, **buồn**, **giận dữ**.

- Với các cảm xúc khó phân biệt như **disgust** hoặc **fear**, vẫn còn nhầm lẫn nhẹ, đặc biệt trong môi trường ồn hoặc khi giọng không rõ ràng.

- Việc sử dụng MFCC giúp giảm thiểu ảnh hưởng bởi nhiễu nền và tốc độ nói.

- Mô hình nhẹ, phù hợp triển khai trên các thiết bị nhúng như Raspberry Pi.

3.3. Tích hợp và hiển thị kết quả

Sau khi nhận diện cảm xúc từ khuôn mặt và giọng nói, hệ thống tiến hành lưu trữ, xử lý và hiển thị kết quả thông qua giao diện web. Phần này trình bày quy trình ghi dữ liệu, thống kê cảm xúc và đánh giá mức độ hài lòng của người dùng.

3.3.1. Giao diện trang chính (/)

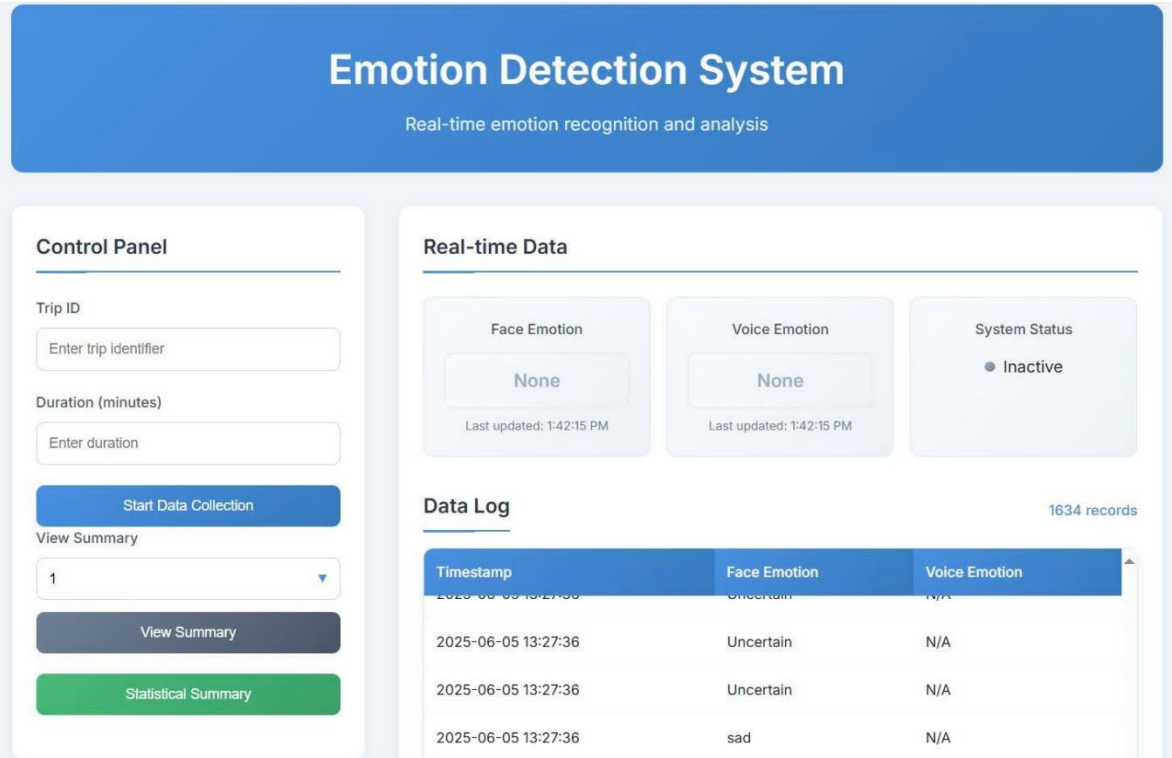
Giao diện chính hiển thị thông tin tổng quan của hệ thống bao gồm:

+ Danh sách các chuyến đi (trip_id) đã ghi nhận.

+ Kết quả cảm xúc mới nhất từ khuôn mặt và giọng nói (hiển thị realtime).

+ Nút điều khiển để bắt đầu hoặc dừng thu thập dữ liệu.

Giao diện chính hiển thị danh sách chuyến đi và cảm xúc mới nhất được thể hiện qua hình 11:



Hình 11: Giao diện chính hiển thị danh sách chuyến đi và cảm xúc mới nhất

3.3.2. Ghi dữ liệu cảm xúc vào file CSV

Mỗi chuyến đi được tạo một tệp .csv riêng biệt và dữ liệu thu thập được ghi theo thời gian thực

Trường "N/A" được dùng nếu chưa có dữ liệu ở một trong hai nguồn tại thời điểm ghi.

Ví dụ nội dung file CSV một chuyến đi được thể hiện qua hình 12:

Hình 12: Ví dụ nội dung file CSV một chuyến đi

3.3.3. Trang thống kê kết quả (/summary)

Giao diện thống kê giúp người dùng dễ dàng đánh giá toàn bộ hành trình theo từng chuyến đi cụ thể.

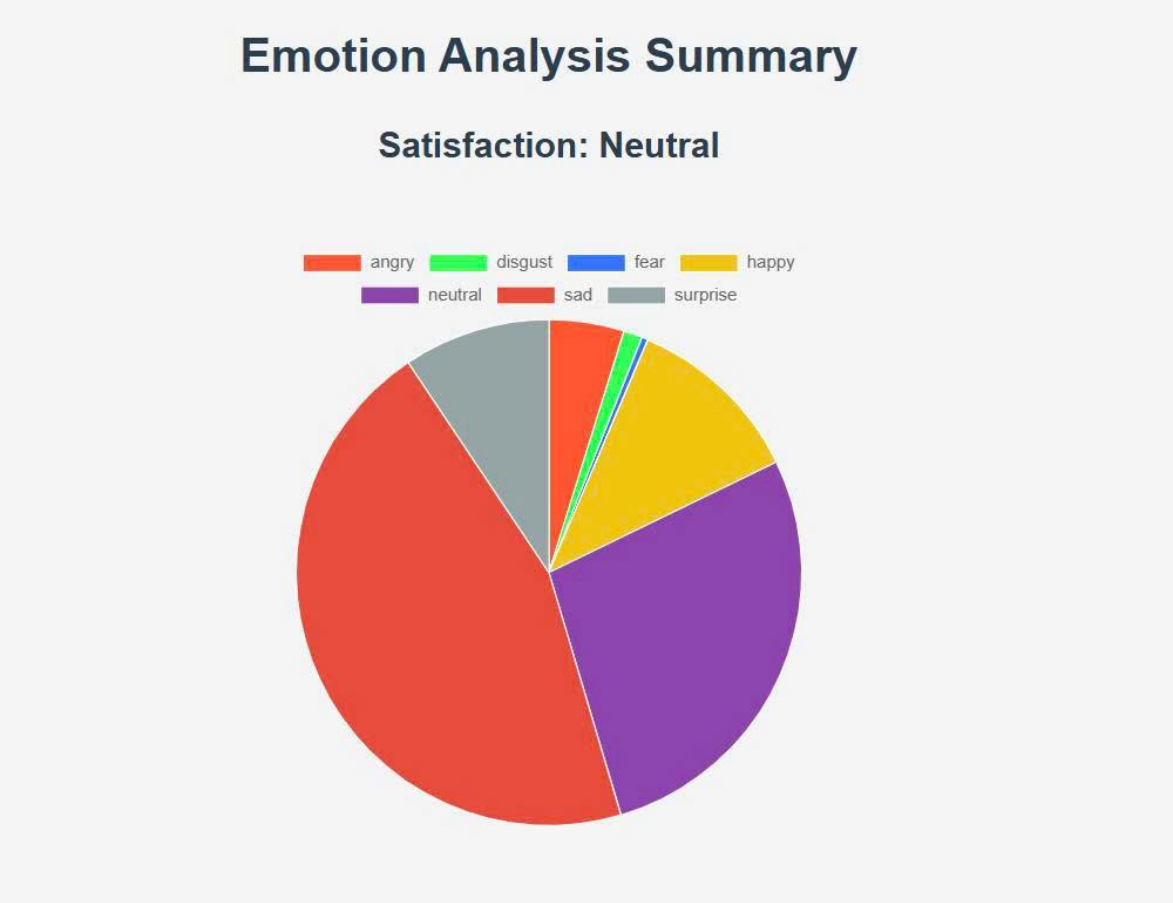
a. Thống kê tỉ lệ cảm xúc

Hệ thống loại bỏ các giá trị "N/A" hoặc "Uncertain" trước khi tính toán.

Tỉ lệ phần trăm mỗi cảm xúc được tính theo công thức:

$$\text{Tỉ lệ (\%)} = \frac{\text{Số dòng cảm xúc X}}{\text{Tổng số dòng hợp lệ}} \times 100$$

Dữ liệu được hiển thị bằng biểu đồ cột hoặc biểu đồ tròn được thể hiện qua hình 13:.



Hình 13: Biểu đồ thống kê tỉ lệ cảm xúc theo khuôn mặt và giọng nói

b. Tính điểm hài lòng

Hệ thống quy đổi cảm xúc thành hệ số đánh giá được thể hiện qua bảng 8:

Emotion	**Hệ số đánh giá**
happy	+1.0
neutral	+0.5
surprise	0.2
sad	–0.5

fear	–0.7
disgust	–0.8
angry	–1.0

Bảng 8: Hệ thống quy đổi cảm xúc thành hệ số đánh giá

Công thức tính điểm tổng:

$$\text{Điểm hài lòng} = \sum(\text{Tỉ lệ cảm xúc} \times \text{Hệ số})$$

Kết quả phân loại được thể hiện qua bảng 9:

Giá trị normalized_score	**Mức độ hài lòng**
≥ 0.4	Very Satisfied (Rất hài lòng)
>=0.1 và <0.4	Satisfied (Hài lòng)
>=-0.1 và <0.1	Neutral (Bình thường)
>=-0.4 và <-0.1	Dissatisfied (Không hài lòng)
<-0.4	Very Dissatisfied (Rất không hài lòng)

Bảng 9: Kế quả phân loại

3.3.4 Giao diện thống kê mức độ hài lòng

Mục tiêu

Trang web "Satisfaction Statistics" được thiết kế nhằm hiển thị trực quan và chi tiết các số liệu liên quan đến mức độ hài lòng của người dùng sau mỗi chuyến đi. Đây là một phần quan trọng của hệ thống đánh giá cảm xúc và phản hồi của người dùng, giúp đưa ra cái nhìn tổng quan về chất lượng dịch vụ thông qua thống kê và biểu đồ.

Mô tả giao diện

Tiêu đề: "Satisfaction Statistics" được căn giữa và định dạng nổi bật giúp người dùng dễ nhận diện mục đích trang.

Lưới thống kê (Stats Grid): Hiển thị 4 thông tin chính:

- Tổng số chuyến đi hài lòng (Very Satisfied + Satisfied)

- Số chuyến đi trung lập (Neutral)

- Số chuyến đi không hài lòng (Dissatisfied + Very Dissatisfied)

- Tổng số chuyến đi (all_trips_data|length)

Biểu đồ tròn (Pie Chart): Trình bày phân bố tỉ lệ từng mức độ hài lòng, giúp dễ dàng nhận biết tỷ lệ phần trăm.

Màu sắc biểu thị rõ ràng: Mỗi mức độ hài lòng được tô màu riêng, tương ứng với cảm xúc:

- Rất hài lòng: Xanh lá

- Hài lòng: Xanh dương

- Trung lập: Vàng

- Không hài lòng: Cam

- Rất không hài lòng: Đỏ

Tính năng tương tác:

- Biểu đồ hiển thị tooltip kèm số lượng và phần trăm khi rê chuột.
- Các khối thống kê có hiệu ứng hover nhẹ giúp giao diện sinh động.

Xử lý dữ liệu

- Dữ liệu thống kê được truyền từ backend (Flask) thông qua biến satisfaction_counts.

- JavaScript xử lý dữ liệu bằng cách chuyển JSON từ server thành cấu trúc phù hợp cho biểu đồ.

- Sử dụng Chart.js để hiển thị biểu đồ tương tác theo thời gian thực.

Vai trò trong hệ thống

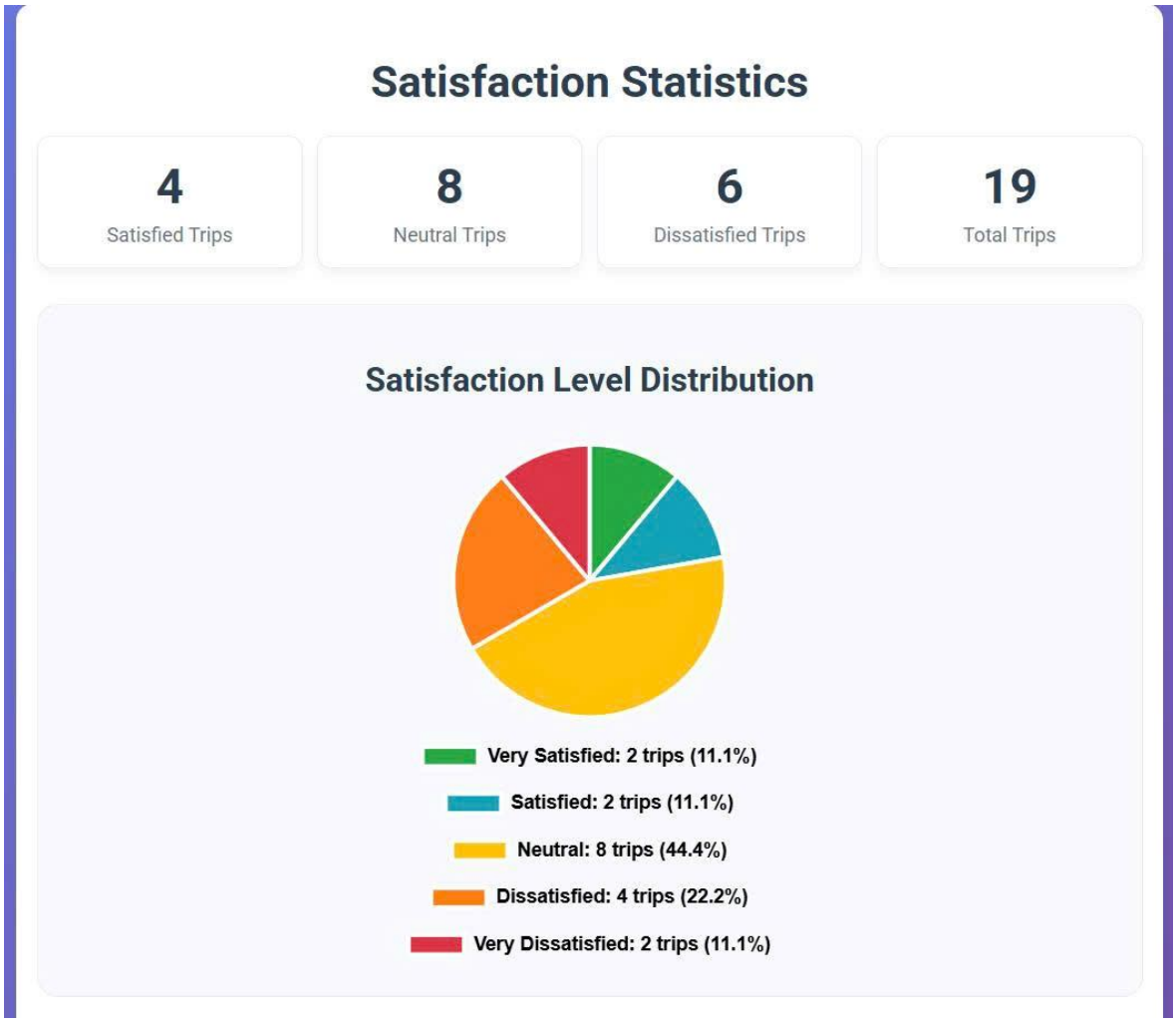
Giao diện này đóng vai trò phản ánh tổng quan hiệu suất hệ thống, giúp:

- Người quản lý đánh giá chất lượng trải nghiệm người dùng.

- Người dùng có thể theo dõi thống kê chuyến đi một cách minh bạch.

- Là một phần bổ sung cho hệ thống nhận diện cảm xúc và phản hồi sau chuyến đi (được trích xuất từ hình ảnh hoặc âm thanh).

Giao diện thống kê mức độ hài lòng được thể hiện qua hình 14:



Hình 14: Giao diện thống kê mức độ hài lòng

4. KẾT LUẬN

4.1. Đánh giá

Đề tài "**Hệ thống nhận diện cảm xúc người dùng qua hình ảnh và âm thanh**" đã hoàn thành các mục tiêu cốt lõi đề ra ban đầu. Hệ thống có khả năng:

- Thu thập và xử lý dữ liệu cảm xúc từ hình ảnh và âm thanh theo thời gian thực.
- Ghi lại và phân tích dữ liệu cảm xúc theo từng phiên (chuyến đi).
- Tổng hợp và đánh giá mức độ hài lòng của người dùng thông qua cảm xúc ghi nhận được.
- Cung cấp giao diện web đơn giản, trực quan, dễ sử dụng cho người dùng và dễ mở rộng.

Thông qua việc kết hợp nhiều công nghệ như OpenCV, TensorFlow, librosa, Flask, và biểu đồ thống kê, hệ thống đã đáp ứng yêu cầu của một ứng dụng giám sát cảm xúc cơ bản. Tuy vẫn còn một số hạn chế nhất định (ví dụ: độ chính xác của mô hình âm thanh khi có tiếng ồn), nhưng kết quả đạt được là ổn đinh, phù hợp với thời gian và phạm vi nghiên cứu của đồ án lần này.

4.2. Hướng phát triển

Trong tương lai, hệ thống có thể được phát triển theo nhiều hướng chuyên sâu hơn để cải thiện độ chính xác, độ tin cậy và phạm vi ứng dụng.

4.2.1. Nhận diện cảm xúc qua hình ảnh

- Sử dụng mô hình học sâu hiện đại hơn như ResNet50, EfficientNet, hoặc Vision Transformer (ViT) để tăng độ chính xác.
- Áp dụng kỹ thuật transfer learning từ các tập dữ liệu lớn (FER+, AffectNet...) để tối ưu hóa mô hình trên dữ liệu thực tế.
- Cải thiện khả năng nhận diện khuôn mặt trong điều kiện ánh sáng yếu hoặc khi khuôn mặt bị che khuất.

4.2.2. Nhận diện cảm xúc qua âm thanh

- Mở rộng mô hình âm thanh bằng cách huấn luyện với ngữ liệu tiếng Việt đa dạng hơn, đảm bảo phản ánh đúng đặc điểm ngôn ngữ và văn hóa.
- Áp dụng transformer-based models như Wav2Vec2 hoặc Whisper để nâng cao độ chính xác trong môi trường có tiếng ồn.
- Kết hợp thêm phân tích ngữ điệu và tốc độ nói để đánh giá cảm xúc sâu hơn.

4.2.3. Phần cứng

- Tích hợp cảm biến hình ảnh và micro chất lượng cao hơn trên các thiết bị như Raspberry Pi 5, Jetson Nano, hoặc các thiết bị IoT chuyên dụng.
- Bổ sung các cảm biến khác như nhiệt độ da, nhịp tim (MAX30102) để hỗ trợ đánh giá trạng thái tâm lý sinh lý người dùng.
- Nâng cấp màn hình hiển thị OLED hoặc cảm ứng để trực quan hơn trong ứng dụng thực tế (xe buýt, lớp học, tư vấn tâm lý…).

4.3. Chi phí thực hiện

Chi phí ước lượng trong quá trình thực hiện đề tài được thể hiện ở bảng 10 như sau:

Thành phần	**Mô tả**	**Chi phí (VNĐ)**
Raspberry Pi 4 (4GB RAM)	Bộ xử lý trung tâm	1.450.000
Camera USB/CSI	Dùng để ghi hình ảnh khuôn mặt	100.000
Microphone USB	Thu âm giọng nói	30.000
Dây nối, breadboard, nguồn, vỏ hộp	Phụ kiện phần cứng	200.000
Tổng chi phí		**1.780.000**

Bảng 10: Chi phí thực hiện

DANH MỤC TÀI LIỆU THAM KHẢO

1. Ejlok1. (2021). Audio Emotion - Part 1 | Explore Data. Truy cập ngày 10/05/2025, từ: https://www.kaggle.com/code/ejlok1/audio-emotion-part-1-explore-data/notebook

2. msambare. (n.d.). FER2013. Truy cập ngày 12/04/2025, từ: https://www.kaggle.com/datasets/msambare/fer2013

3. FPT Aptech. (n.d.). CNN là gì? Tìm hiểu về mạng nơ ron tích chập trong học sâu. Truy cập ngày 20/04/2025, từ: https://aptech.fpt.edu.vn/cnn-la-gi.html

4. Nguyễn Văn Đạt. (19/6/2019). Building CNN model, Layer Patterns and Rules! Viblo. Truy cập ngày 10/04/2025, từ:https://viblo.asia/p/building-cnn-model-layer-patterns-and-rules-Do754qXQKM6