MIT Applied AI & Data Science

Facial Emotion Recognition

Using Deep Learning

Capstone Project · MIT Applied AI & Data Science Program

Kinn Coelho Julião | March 2026

7.5 min presentation
+ 10 min Q&A

Agenda

1. Business Context
2. Problem Statement
3. Dataset Overview
4. Exploratory Data Analysis
5. Model Architectures
6. Model Performance Comparison
7. Final Model & Detailed Results
8. Business Applications
9. Key Takeaways & Future Work

Total presentation time: 7.5 minutes | Q&A: 10 minutes

Why Emotion Recognition Matters

The Foundation

- 55% of sentiment communicated through facial expressions
- Source: MIT Capstone Brief (Mehrabian, 1967)
- Affective Computing trains machines to read this signal
- Uses: customer service · healthcare · education · HCI

The Priority: US Public Transportation

- 17.6% of fatal US crashes involve drowsy driving (AAA Foundation, 2024)
- Fatigue costs US society $109 billion/year (NHTSA, 2024)
- 451,238 school buses · 21.4M children daily (NHTSA)
- Fatigue builds gradually — detectable via the 'Sad' facial signature

Why Now — It Is Already Law

- EU GSR: DMS mandatory on ALL new vehicles from July 2024
- US NHTSA developing parallel requirements
- Governments: both regulator AND largest fleet operator
- This is a legal requirement — not a speculative market

Sources: AAA Foundation aaafoundation.org/drowsy-driving-in-fatal-crashes-united-states-2017-2021 · NHTSA nhtsa.gov/risky-driving/drowsy-driving · NHTSA school bus nhtsa.gov/crashworthiness/school-bus-crashworthiness-research · EU GSR smarteye.se/blog/the-general-safety-regulations-gsr-and-driver-monitoring-systems-dms

Problem Statement

Build a deep learning model to classify facial expressions into 4 categories from 48×48 grayscale images — each mapping directly to a driver safety state

Happy → Alert & engaged

Sad → Fatigue onset (primary target)

Surprise → Sudden event / hazard

Neutral → Baseline driving state

Task Type	Multi-class image classification (4 classes → 4 driver states)
Input	48 × 48 pixel grayscale facial images
Output	Softmax probability over 4 emotion classes
Success Metrics	Accuracy + Macro F1-score on held-out test set
Domain	Affective Computing / Computer Vision / Public Safety

Dataset Overview

Training Set

15,109

images

Validation Set

4,977

images

Test Set

128

images (32/class)

Training Set — Class Distribution

Class	Count	Share
😊 Happy	3,976	26.3%
😐 Neutral	3,978	26.3%
😢 Sad	3,982	26.4%
😮 Surprise	3,173	21.0%

Class Balance Visual

⚠ *Note: Surprise class is ~20% underrepresented vs. others. Images are 48×48 grayscale JPG. Test set is perfectly balanced (32 images/class).*

Exploratory Data Analysis

😊 Happy

Most visually distinct class. Broad smiles, crinkled eyes (Duchenne markers), raised cheeks. Strong inter-class separation from other classes.

😮 Surprise

Strong multi-feature signal: raised eyebrows, wide-open eyes, open mouth. Despite fewer samples, highly recognizable by the model.

😢 Sad

Subtle micro-expressions: downturned lip corners, raised inner eyebrow (AU1), slightly narrowed eyes. Harder to distinguish from Neutral.

😐 Neutral

Defined by ABSENCE of expression. No distinctive features → most ambiguous class. Shares visual space with low-intensity Sad expressions.

🔑 Key Challenge: The Neutral ↔ Sad boundary is inherently ambiguous — even human annotators disagree. This is the most common confusion pair in the dataset and reflects a fundamental limit of image-only classification.

Class distribution is approximately balanced (slight underrepresentation of Surprise ~21%).

6 Model Architectures Evaluated

#	Model	Architecture	Input	Type	Notes
1	ANN Baseline	Flatten → Dense(256) → Dense(128) → Dense(4)	Grayscale	Custom	Baseline — no spatial learning
2	CNN Models (1 & 2)	2 & 3 Conv blocks (32→64 / 32→64→128) + FC	Grayscale	Custom	Captures spatial patterns
3	VGG16	Frozen pretrained + GlobalAvgPool + Dense head	RGB	TL	Transfer learning (ImageNet)
4	ResNet50V2	Frozen pretrained + skip connections + Dense	RGB	TL	Transfer learning + residuals
5	EfficientNetB0	Compound scaling + Dense head	RGB	TL	Efficient transfer learning
6	**Complex CNN ★**	**5 Conv blocks (32→64→128→256→512) + FC**	**Grayscale**	**Custom**	**BEST — custom deep CNN**

All models: final layer = Dense(4, softmax) | Loss: categorical_crossentropy | Optimizer: Adam

Model Performance Comparison

Model	Val Acc	Test Acc	Notes
ANN Baseline	~49.3%	51.56%	Above random (25%), no spatial learning
CNN Models (1 & 2)	—	CNN1: 69.53% / CNN2: 75.00%	Spatial learning; progressive depth delivers consistent gains
VGG16	—	51.56%	ImageNet→face domain gap hurts
ResNet50V2	—	55.47%	Domain gap limits despite skip connections
EfficientNetB0	—	60.16%	Best TL model — compound scaling helps
Complex CNN ★	**—**	**82.03%**	**BEST — grayscale-native, task-optimised**

Key Insight: Transfer learning models (VGG16, ResNet50V2, EfficientNetB0) underperform the custom CNN.
Root cause: ImageNet pretraining on natural RGB images creates a domain gap when applied to 48×48 grayscale faces.
The task-specific Complex CNN — trained end-to-end on this exact data — achieves 82.03% test accuracy, 21.87pp above the best TL model (EfficientNetB0 at 60.16%).

Final Model: Complex CNN (5 Convolutional Blocks)

Block 1	Conv2D(32) → BatchNorm → Conv2D(32) → MaxPool → Dropout(0.2)
Block 2	Conv2D(64) → BatchNorm → Conv2D(64) → MaxPool → Dropout(0.2)
Block 3	Conv2D(128) → BatchNorm → Conv2D(128) → MaxPool → Dropout(0.3)
Block 4	Conv2D(256) → BatchNorm → Conv2D(256) → MaxPool → Dropout(0.3)
Block 5	Conv2D(512) → BatchNorm → MaxPool → Dropout(0.4)
Head	Flatten → Dense(512) → BatchNorm → Dropout(0.5) → Dense(256) → Dense(4, softmax)

✅ Why chosen: Grayscale-native (no domain gap) · Highest accuracy (82.03% test) · Balanced per-class performance · Progressive feature hierarchy (32→512 filters) · Regularised with BatchNorm + Dropout at every level

Model Evaluation Results

Confusion Matrix — Key Patterns

✅ Strong Diagonal

Correct predictions dominate all four classes — model has learned genuine discriminative features

😊 Happy

Highest individual accuracy — visually distinct, rarely confused with other classes

😮 Surprise

High recall — multi-feature signal (raised brows + open mouth) makes it identifiable

😐↔😢 Neutral/Sad

PRIMARY confusion pair. Neutral mis-classified as Sad and vice versa — inherent visual ambiguity

😢 Sad

Moderate precision — some sad images classified as neutral due to subtle expression

Per-Class Metrics (Test Set)

Class	Precision	Recall	F1
😊 Happy	0.93	0.84	0.89
😮 Surprise	0.90	0.88	0.89
😐 Neutral	0.68	0.84	0.75
😢 Sad	0.82	0.72	0.77

Overall Test Accuracy: 82.03% | Macro F1: 0.82

Training Dynamics

- Validation loss converged smoothly without severe overfitting
- BatchNorm + Dropout effectively controlled regularization
- Training ran locally on Apple M3 Max (Metal GPU acceleration)
- Early stopping monitored val_loss

Note: Neutral↔Sad confusion is consistent with human perception limits and academic benchmarks.

Open Source Public Safety — Deploying at Scale

Deployment Stack (per vehicle)

IR Camera (RPi Cam 3 NoIR): $25 · adafruit.com/product/5659

- Compute (Raspberry Pi 5, 8GB): $80 · raspberrypi.com
- 4G LTE module (Waveshare SIM7600G-H): ~$50 · amazon.com
- Enclosure + power + storage + alert: ~$145

Total hardware: ~$300 | Installation: ~$200 | Year 1: ~$500/vehicle

Public Sector Scale & Savings

451,238 school buses + ~65,000 transit buses (US) — APTA 2024 Fact Book

- Commercial DMS hardware: $200–$700/vehicle + $100–$300/year subscription
- Open source solution: ~$500/vehicle Year 1 — comparable upfront — zero subscription fees
- 10,000 vehicles over 5 yrs: ~$9M open source vs $14–30M commercial

Government saves $5–21M per 10,000 vehicles over 5 years

Why Open Source?

Auditable: governments can inspect every decision the model makes

- Forkable: any province fine-tunes for local conditions — shared gains
- No vendor lock-in: IP stays in the public domain permanently
- Maintainable: ~4–5 public servants + $500K/yr runs the entire system

EU GSR mandates DMS on all new vehicles from July 2024 — governments are both regulator and largest fleet operator

"This model runs on a ~$300 hardware stack. For 10,000 vehicles over 5 years, open source saves $5–21M versus commercial alternatives — and the IP stays in the public domain."

Key Takeaways & Future Directions

🎯 Key Takeaways

1

Custom CNN > Transfer Learning

A task-specific grayscale CNN outperforms ImageNet-pretrained RGB models by 21.87pp. Domain alignment matters more than model size.

2

Neutral–Sad Confusion is Expected

This confusion pair mirrors documented human perception limits. It reflects both model limits and fundamental human perception ambiguity at low-intensity affect.

3

82.03% Accuracy from 20K Images

Deep learning achieves strong results on this task even with a modest training set, demonstrating the power of well-regularised CNNs.

🚀 Future Improvements

1

Higher Resolution Input

96×96 or 224×224 px gives more spatial detail for Neutral-Sad boundary — the primary remaining challenge

2

Facial Landmark Features

Mouth corner angle + inner brow position as additional inputs improves fine-grained discrimination

3

Real-time Video with Temporal Smoothing

LSTM/Transformer across video frames reduces false positives — essential for live vehicle deployment

4

Government Consortium Model

Provinces/municipalities share architecture + jointly fund a maintenance team → improvements benefit all jurisdictions

Risks & Implementation Considerations

Deployment Risks

1 — False Positives in Fatigue Detection

Unnecessary driver interventions reduce trust. Temporal smoothing: require signal sustained across ≥3 consecutive frames before alerting.

2 — Neutral/Sad Confusion in Safety-Critical Use

F1 0.75–0.77 not sufficient as sole signal. Pair with: eye closure, head pose, speed deviation.

3 — Demographic Bias

Training data coverage not independently audited. Bias evaluation across age, gender, ethnicity required before production.

Implementation Challenges

1 — Face Detection Upstream

Model expects pre-cropped centred face. MTCNN or OpenCV Haar cascade required as pre-processing stage.

2 — Privacy & Regulatory Compliance

US laws vary (CCPA, Illinois BIPA). Require: driver consent, local-only inference, encrypted logs, access audit trail.

3 — Real-Time Video Integration

Current model is still-image only. Live deployment needs frame-rate pipeline + temporal smoothing (LSTM/sliding window).

4 — Ongoing Maintenance

~4–5 public sector staff + ~$500K/yr. Net saving vs commercial: $4.5–20.5M over 5 years per 10,000 vehicles.

The EU GSR already mandates DMS — governments are deciding how to procure this technology, not whether to adopt it.