



GRADUATION PROJECT

The Impact of Using a Deep Learning Based Facial Expression Recognition

Prepared by: Chadi Omar Elminawi | ID No: 252001554

Supervisor: Dr. Amjad Qatish

College of Informatics — Midocean University
Master of Science in Artificial Intelligence

Presentation Outline

01

Problem & Motivation

02

Research Objectives & Contribution

03

Related Work & Research Gap

04

System Requirements & Design Methodology

05

Proposed Model: Dataset & CNN Architecture

06

Experimental Results & Evaluation

07

Conclusion & Future Work

Problem Definition

Visually impaired individuals face major difficulty reading facial expressions one of the most important channels of non-verbal communication leading to poor understanding, weaker communication, and social isolation, with a negative impact on mental health and social participation.

1

Most existing assistive systems focus on navigation and object detection, with few addressing real-time emotional understanding.

2

High-accuracy CNN models are often computationally heavy, making deployment on resource limited devices difficult.

3

Limited research exists on lightweight, accurate, and affordable systems deployable on embedded platforms.

The Need: a low-cost, real-time intelligent system that translates expressions into understandable audio feedback.

Research Objectives

Main Objective: Develop an intelligent real-time emotion recognition system using deep learning that classifies facial expressions and provides auditory feedback during social interactions for visually impaired users.

- 1 Build an automated computer-vision pipeline (OpenCV) for real-time face detection and ROI cropping
- 2 Design and train a custom CNN to classify facial expressions into 7 core categories
- 3 Address dataset class imbalance via dynamic class-weighting techniques
- 4 Improve performance on hard samples through a two-phase training strategy
- 5 Convert recognized emotions into natural spoken phrases using TTS technology
- 6 Build an accessible, interactive web interface (Gradio) for live camera evaluation
- 7 Optimize the for high accuracy and low latency suitable for portable deployment

Thesis Contribution

1

End-to-End Integration

Real-time face detection (Haar Cascade), ROI extraction, CNN classification, and emotional voice feedback generation (gTTS) in one unified pipeline.

2

Advanced Training Strategy

Class-weight balancing combined with a novel two-phase fine-tuning approach that isolates and retrain on hard/misclassified images.

3

Context-Aware Audio Output

Direct mapping of emotion classifications into expressive spoken sentences, e.g. "I am feeling very happy today!".

4

Web & Live-Camera Deployment

Interactive Gradio interface with WebRTC live camera capture for real-time validation and testing.

Related Work & Research Gap

Key Related Studies

Kusuma et al. (2020) Wearable-device FER system converting expressions to audio 78–80% accuracy in real-world conditions

Shehada et al. (2023) Lightweight FER with partial transfer learning on Raspberry Pi Up to 82.1% accuracy

Kirkvik (2022) Compared sequential CNNs with ResNet50 and VGG16 Transfer learning reduces training time

The Research Gap

Most accurate models are computationally heavy and hard to deploy on low-cost hardware, current assistive systems rarely achieve real-time performance in uncontrolled environments, and little research addresses fully integrated systems (detection + classification + audio feedback) for embedded platforms a gap this thesis addresses through an optimized two-phase CNN pipeline.

User Needs & Engineering Requirements

User Needs

- Use an off-the-shelf, general-purpose computing platform
- High-accuracy facial-expression classification
- Easy to use: capture a snapshot and analyze it
- Play back the identified emotion via a speaker
- Reliable across different people and lighting conditions
- Cost-effective in hardware and software

Key Engineering Requirements

$\geq 85\%$ Facial-expression classification accuracy

Choosing an off-the-shelf platform reduced development complexity while meeting deep-learning performance needs, and diverse training data (FER2013) reduced bias and improved accuracy across classes.

Computing Platform: Google Colab, Training to Deployment

Google Colab was adopted as a single unified platform covering the entire project lifecycle: GPU acceleration for model training, then hosting the Gradio interface directly within the same notebook with a shareable public link for instant testing no separate server or device required.

1

Google Colab - Training

- High-performance GPU acceleration
- Pre-installed Python libraries
- Seamless Google Drive integration for model saving

2

Google Colab - Gradio Deployment

- Gradio interface launched directly inside the notebook
- Shareable public link (share=True), no separate server
- Live in-browser image capture via JavaScript

This fully cloud-based approach removes the need for a local server or dedicated device, letting the system be tested instantly from any browser via the public link.

Dataset: FER2013

An open-source benchmark dataset introduced during the ICML 2013 Challenges, containing 35,887 grayscale images at 48×48 pixels, spread across 7 emotional-expression classes, reflecting real-world diversity in age, gender, pose, and lighting.

35,887

Images

48×48

Grayscale px

7

Expression classes

The Seven Expression Classes

Angry

Disgust

Fear

Happy

Sad

Surprise

Neutral

Angry	Angry	Fear	Sad	Neutral
Fear	Sad	Happy	Happy	Fear
Angry	Neutral	Neutral	Neutral	Happy
Surprise	Happy	Fear	Neutral	Sad
Sad	Fear	Angry	Angry	Happy

Samples from the FER2013 dataset

Proposed Architecture: Custom Two-Phase CNN

Rather than relying on heavy pre-trained models (e.g., VGGNet), a custom sequential CNN was built (3 convolutional blocks + 1 dense classification block) designed to be lightweight, fast, and effective for real-time inference.



PHASE 1 Initial Training

- Data Augmentation
- Balanced class weights
- Up to 60 epochs (automatic early stopping)

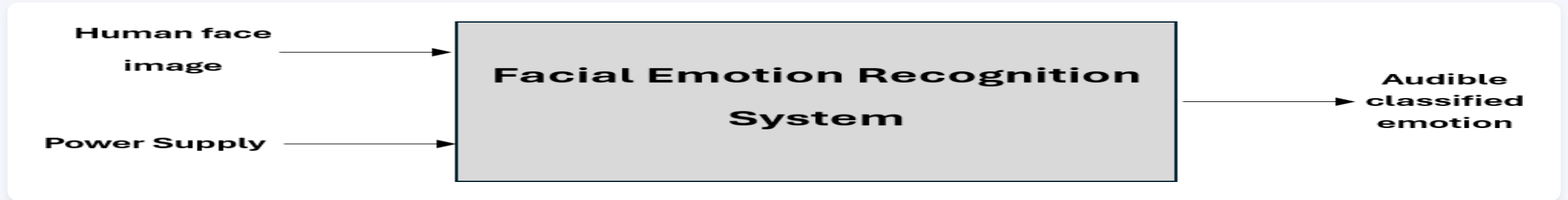
PHASE 2 Hard-Images Fine-Tuning

- Extract misclassified images from validation
- Retrain to refine decision boundaries
- Up to 50 more epochs (automatic early stopping)

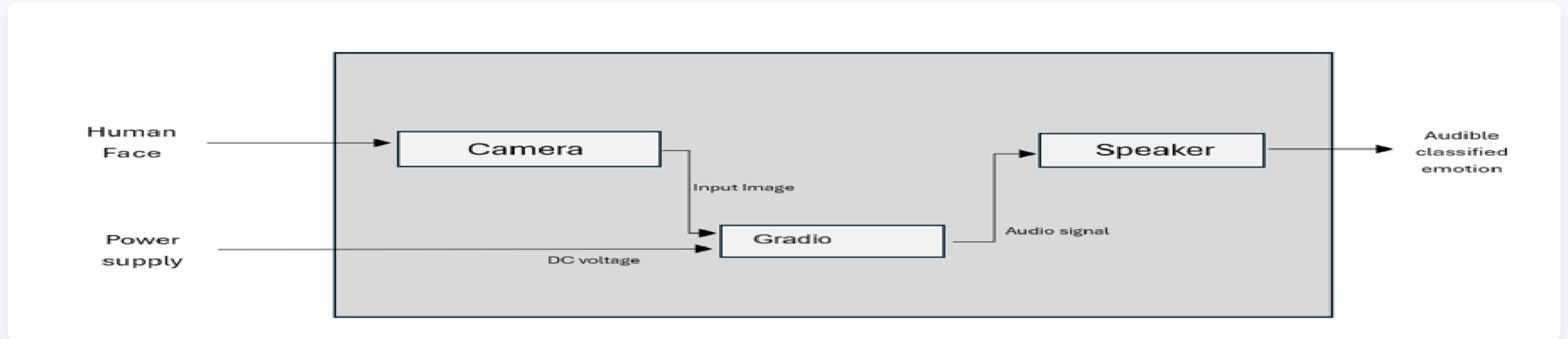
The seven classes: Angry · Disgust · Fear · Happy · Sad · Surprise · Neutral

System Design

Level 0

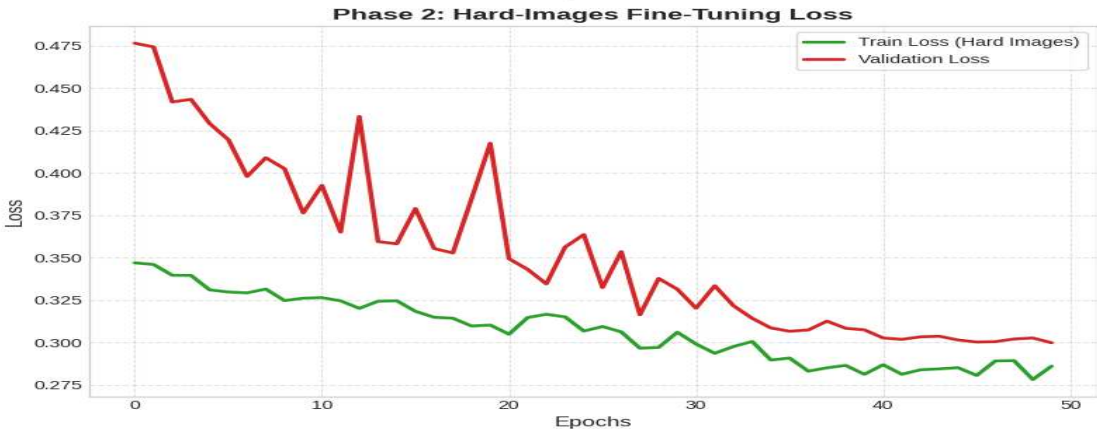
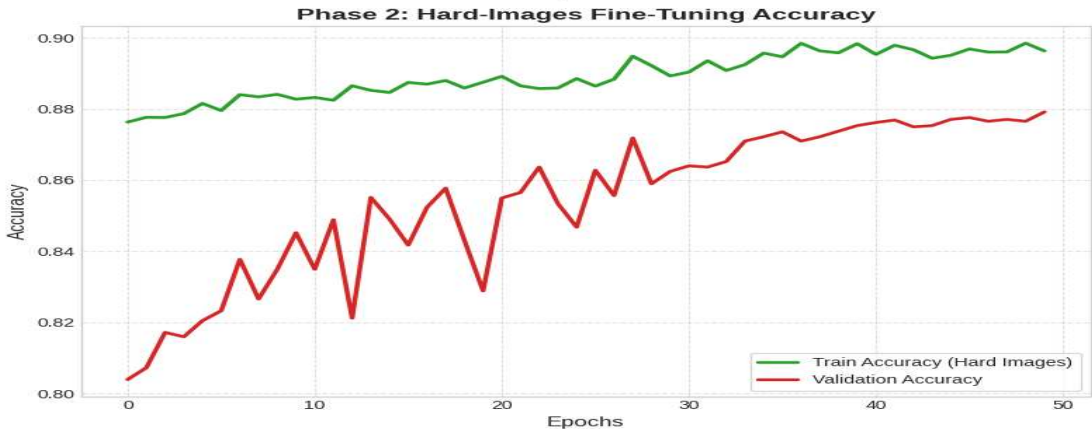


Level 1 — Detailed System Components



Performance Results: Accuracy & Loss

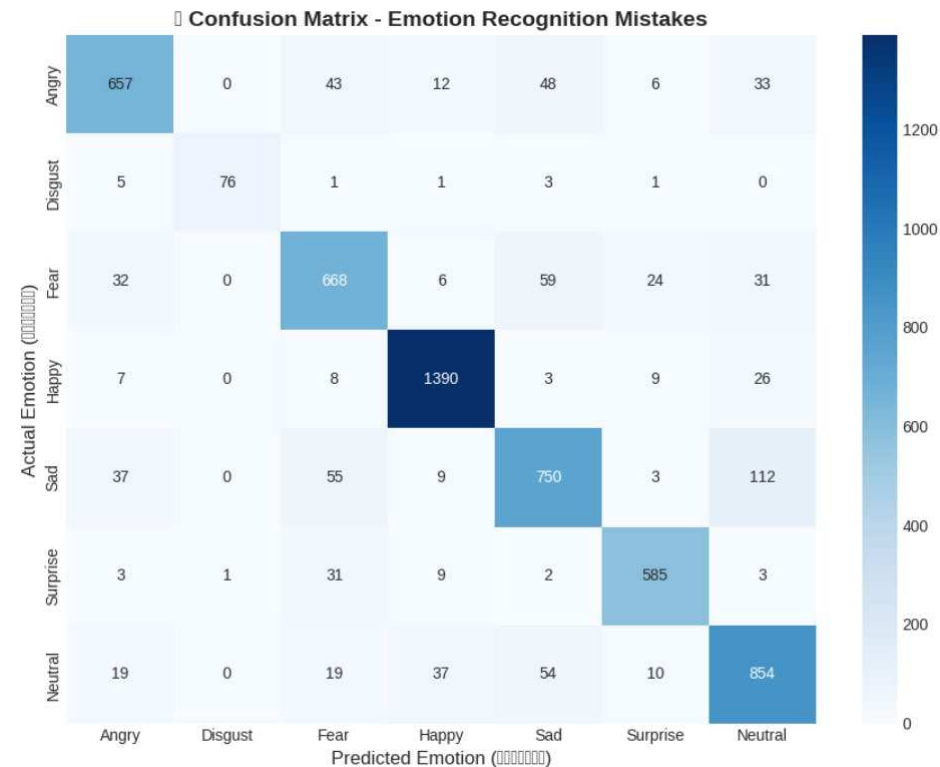
Model Performance Analysis (Graduation Project)



Evaluation: Confusion Matrix

A comprehensive quantitative evaluation was performed on an unseen FER2013 test subset, covering all seven classes, to extract True/False Positive and Negative rates (TP, TN, FP, FN) and analyze model performance in detail.

- Strong performance across most classes, with some overlap between visually similar expressions (e.g., Fear and Surprise)
- Stable extraction and classification despite variation in lighting and pose
- A solid foundation for real-time social-interaction assistive applications



Confusion matrix of the custom CNN model

Interactive System Interface (Gradio)

An easy-to-use web interface captures a live camera frame, automatically detects and crops the face via OpenCV, passes the image to the trained CNN model (FER2013_Final.h5), displays prediction probabilities for each expression, and immediately plays the generated gTTS voice response.



Conclusion & Key Contributions

1

Efficient Real-Time Performance

Fully local (edge computing) processing reduces response latency and preserves data privacy.

2

Reliable Accuracy

An optimized CNN architecture with strategic Dropout layers balances accuracy and computational complexity.

3

Operational Robustness

Stable performance despite changes in lighting, pose, and partial facial occlusion.

4

Social Inclusion

Context-aware audio feedback gives visually impaired users greater independence in social interaction.

Future Work

- 1 Transition to hybrid models (Vision Transformers) to capture contextual relationships between facial landmarks
- 2 Use lightweight pre-trained architectures (MobileNetV3 / EfficientNet-Lite) via transfer learning
- 3 Expand toward a multimodal system fusing vocal tone and linguistic content with facial analysis
- 4 Deploy models on dedicated AI accelerators such as NVIDIA Jetson Nano and Google Coral TPU
- 5 Conduct long-term field studies with real users under strict privacy and consent protocols

Thank You

Questions & Discussion

Chadi Omar Elminawi · 252001554

Supervised by Dr. Amjad Qatish — College of Informatics, Midocean University