FPGA-based Real-Time Emotion Recognition System using Facial Expressions for Physically Disabled Individuals

M. Kamaraju
Professor & Director (AS&A)
Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru, Krishna District, Andhra Pradesh, India – 521356

K. Ujwala
M.Tech (VLSI&ES)
Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru, Krishna District, Andhra Pradesh, India – 521356

B. Rajasekhar
Professor & HOD
Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru, Krishna District, Andhra Pradesh, India – 521356

ABSTRACT
Emotion recognition through facial expressions is a critical enabler of non-verbal communication, particularly for individuals with physical disabilities who may face barriers in speech or motor-based interaction. This paper proposes a real-time, FPGA-based facial emotion recognition system optimized for embedded deployment and low-power operation. The system utilizes a quantized MobileNetV2 Convolutional Neural Network (CNN) trained on an enhanced FERPlus dataset (FERPlus-A), which is refined using CLAHE, bilateral filtering, and sharpening to improve feature clarity. The trained model is quantized to 8-bit integer arithmetic for efficient synthesis via Vivado HLS and deployed onto a ZYNQ SoC platform. Integration through AXI interfaces enables seamless communication between the CNN accelerator and the processing system. Simulation results demonstrate high inference speed with a latency of approximately 1.174 milliseconds per frame and an estimated throughput of 851 frames per second. Despite the absence of hardware testing due to board unavailability, functional verification confirms the model's readiness for real-time assistive applications. This work presents a scalable and energy-efficient solution for enhancing emotional communication in assistive technologies, offering significant potential for integration in healthcare, smart interfaces, and human-centered embedded systems.

General Terms
Embedded Systems, Pattern Recognition, Assistive Technology, Real-Time Systems, Human-Computer Interaction.

Keywords
Convolutional Neural Network (CNN), MobileNetV2, FERPlus-A, Vivado HLS.

1. INTRODUCTION
Facial emotion recognition has emerged as a key component in enhancing human-computer interaction, especially for individuals with physical disabilities who face difficulties in expressing emotions through speech or gestures. Emotions conveyed through facial expressions are a universal form of communication, making them ideal for non-verbal interaction. Recognizing these emotions in real-time using embedded systems opens up new avenues in assistive technology, therapeutic systems, and accessibility tools.

Traditional emotion recognition systems often rely on high-performance CPU or GPU-based platforms. While these systems achieve high accuracy, they are not suitable for deployment in real-time or resource-constrained environments due to their high power consumption, size, and cost. In contrast, FPGAs (Field-Programmable Gate Arrays) offer a promising alternative. Their ability to execute operations in parallel, coupled with low power consumption and reconfigurability, makes them ideal for implementing real-time deep learning applications on edge devices [1].

Recent developments in deep learning have shown the effectiveness of Convolutional Neural Networks (CNNs) in facial emotion recognition. However, standard CNN architectures such as VGGNet and ResNet are computationally intensive and demand significant memory resources, making them less suitable for FPGA implementation. To address this, researchers have introduced lightweight CNN models such as MobileNetV2, which utilize compact convolution methods with inverted shortcut connections to reduce complexity without significantly compromising accuracy [2].

In parallel, several studies have proposed quantization techniques to convert floating-point CNN models into fixed-point or integer-only representations. Training that accounts for quantization during model optimization enables neural networks to maintain accuracy even after reducing the bit-width of parameters and activations, thus making them more compatible with hardware like FPGAs [3], [4].

The existing literature also highlights efforts to accelerate CNN inference on FPGAs for emotion recognition. Earlier approaches often used shallow networks or lacked sophisticated preprocessing, which limited their performance. More recent studies integrate high-level synthesis (HLS) tools such as Vivado HLS to design and optimize CNN architectures directly in C/C++ for efficient synthesis into RTL [5], [6].

Building upon these advancements, this paper presents a complete pipeline for facial emotion recognition tailored for individuals with physical disabilities. It includes enhanced dataset preparation, quantized CNN model training, HLS-based synthesis, and integration into a ZYNQ SoC platform. The proposed system bridges the gap between high-performance deep learning and practical embedded deployment, offering a viable assistive technology for emotion-based communication.

2. LITERATURE REVIEW
Facial emotion recognition has gained widespread attention in recent years due to its applications in healthcare, surveillance, and human-computer interaction. Early approaches largely relied on traditional machine learning classifiers such as Support Vector

Machines (SVMs) and K-Nearest Neighbors (KNNs), which required manual feature extraction techniques like Local Binary Patterns (LBP) and Histogram of Oriented Gradients (HOG). While simple and computationally efficient, these methods lacked robustness under real-world conditions and struggled to generalize across diverse facial expressions and lighting conditions.

The advent of deep learning has significantly improved emotion recognition systems. Convolutional Neural Networks (CNNs) have demonstrated superior performance by automatically extracting hierarchical features from raw image data. Works by Barsoum et al. and Li et al. have applied deep CNNs to datasets like FER and FERPlus, showing improved accuracy over classical approaches. However, most of these models were trained and evaluated on high-resource platforms like GPUs, limiting their practical use in embedded or portable devices.

To address resource limitations, lightweight CNN architectures such as MobileNetV2 have been introduced. MobileNetV2 utilizes compact convolution methods with inverted shortcut connections to significantly reduce computational overhead while maintaining accuracy [1]. It is particularly suited for embedded systems where power and processing constraints are critical.

Another critical development has been the integration of quantization techniques. Quantization reduces the bit-width of weights and activations, converting models from 32-bit floating point to low-bit integer arithmetic (often 8-bit). Choi et al. proposed quantization-aware training techniques that maintain accuracy during this conversion process [2]. Similarly, Esser et al. presented learned step-size quantization strategies for optimal hardware efficiency [3].

In the context of hardware deployment, several studies have explored FPGA-based implementations of CNNs for facial analysis tasks. Phan-Xuan et al. implemented a basic CNN architecture on a Xilinx ZYNQ FPGA using high-level synthesis (HLS), achieving real-time performance but with limited accuracy due to the simplicity of the model [4]. Vinhe et al. and Ding et al. focused on optimizing convolutional operations using parallel engines and pipelined execution, demonstrating improvements in both speed and resource utilization [5], [6].

Kim et al. designed an integer-arithmetic-only CNN accelerator optimized for embedded systems using an FPGA platform. Their work demonstrated that with proper quantization and parallelism strategies, a high classification accuracy of 86.58% could be achieved using only integer operations and minimal hardware resources [7].

In addition to model and hardware improvements, preprocessing techniques such as Contrast-Limited Adaptive Histogram Equalization (CLAHE), bilateral filtering, and sharpening have been shown to enhance facial features and improve model performance. Wang et al. reported that preprocessing significantly improved accuracy on low-resolution datasets like FER2013.

3. CNN ARCHITECTURE

Facial expression recognition (FER) has gained significant attention with applications in human-computer interaction, behavioral analysis, and assistive communication. Traditionally, FER systems relied on handcrafted feature extraction such as Local Binary Patterns (LBP) or Histogram of Oriented Gradients (HOG), followed by machine learning classifiers like Support Vector Machines (SVM) [1]. These methods, while efficient under controlled environments, lack the generalization and accuracy necessary for real-time embedded deployment.

With the rise of deep learning, Convolutional Neural Networks (CNNs) have become the de facto approach to FER, enabling direct extraction of spatial features throughout the network from raw facial images [2]. CNNs replace handcrafted features with automatically learned features and achieve superior performance. In particular, the FER2013 dataset has been widely adopted to train and evaluate CNN-based emotion recognition models [3].

The central operation in CNNs is the 2D convolution, described by the following equation:

$$O(y,x) = \sum_{j=-k/2}^{k/2} \sum_{i=-k/2}^{k/2} I(y-j, x-i) \cdot F(j,i) \quad (1)$$

where I(y,x)) is the input image, F(j,i) is the filter kernel, and O(y,x) is the resulting feature map. For multiple channels and filters, the total number of Multiply-Accumulate (MAC) operations becomes:

$$n_{MAC} = H \cdot W \cdot C_{in} \cdot C_{out} \cdot k^2 \quad (2)$$

As described in [4], this operation can be restructured into a matrix multiplication RCRCRC where input patches are flattened into columns of matrix C, and the filters are unrolled into rows of matrix R. This matrix view is more efficient for GPUs, but on FPGAs, direct convolution (Eq. 1 and 2) is often more hardware-efficient due to data reuse and locality.

To address real-time constraints and reduce power consumption, several researchers have ported CNN architectures to FPGA platforms using High-Level Synthesis (HLS). The reference work by Phan-Xuan et al. [4] implemented a CNN-based FER system on a ZYNQ-7000 FPGA using Vivado HLS. Their system featured:

- Real-time face detection from video input (VITA-2000 camera),
- A CNN architecture with 3 convolutional layers and 2 fully connected layers (fig 1)
- RTL-level synthesis of TensorFlow-trained model weights using Vivado HLS.

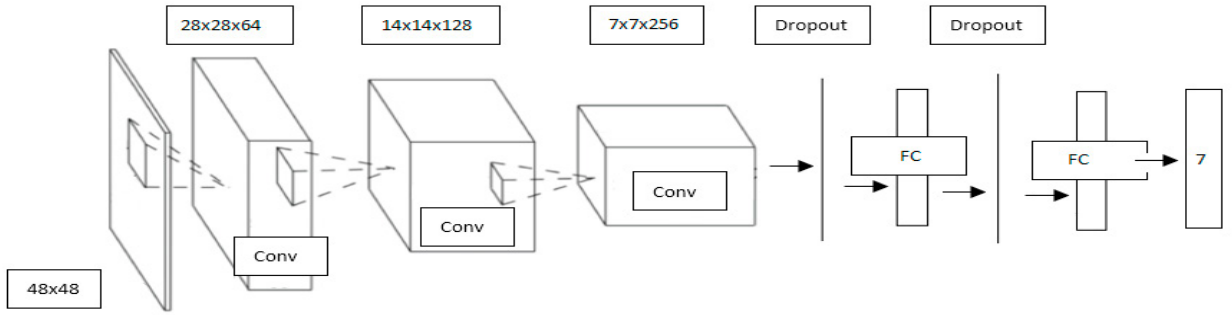

Fig 1: CNN Architecture

4. FPGA-Based Real-Time Facial Emotion Recognition System for Assistive Technology

The methodology begins with acquiring grayscale facial images from a webcam or HDMI video input. These images are resized to 64×64 pixels and passed through a dedicated preprocessing pipeline implemented in software using OpenCV. This pipeline includes three enhancement steps—Contrast Limited Adaptive Histogram Equalization (CLAHE) to improve contrast, bilateral filtering to suppress noise while preserving facial edges, and image sharpening to highlight vital landmarks like eyes and mouth contours. The enhanced images are then forwarded to the CNN core for emotion classification.

The CNN model is built using the MobileNetV2 backbone, which is specifically chosen for its compact convolution methods with inverted shortcut connections that reduce parameter count and computation overhead without sacrificing accuracy. The model is trained in PyTorch using quantization-aware training techniques to ensure compatibility with integer-only arithmetic, a key requirement for FPGA deployment. Upon training, the model is converted to an intermediate ONNX format and then re-implemented in C++ using fixed-point arithmetic. This C++ model is then synthesized using Vivado HLS into an RTL representation that can be compiled and deployed to the programmable logic of the FPGA.

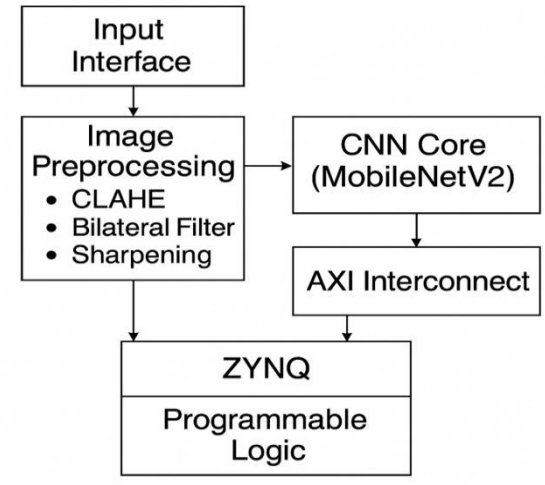

Fig. 2: Block diagram of the proposed real-time facial emotion recognition

The complete hardware-software co-design is shown in **Fig. 2**, which illustrates the system architecture consisting of five core components: the input interface, image preprocessing block, CNN inference engine, AXI interconnect, and classification output. The ARM Cortex-A9 processing system (PS) handles control and I/O functions, while the CNN IP core, synthesized from the trained model, resides in the programmable logic (PL). The AXI4-Lite bus is used for configuration and control of the CNN IP, while the AXI4-Stream bus handles high-throughput image data transfer between the processor and CNN module. Intermediate results and model weights are buffered using on-chip BRAM to reduce latency.

Once the image passes through the CNN, it undergoes classification via a softmax layer that outputs one of eight predefined emotion labels: neutral, happiness, sadness, surprise, anger, fear, disgust, and contempt. The result is then sent back to the processor where it is either displayed on a GUI, stored in memory, or used to trigger assistive responses such as auditory feedback or alerts.

The overall system has been designed with real-time constraints in mind. The synthesized hardware achieves an average inference latency of just 1.174 milliseconds per image, supporting real-time classification of up to 851 frames per second in simulation. Although lightweight, the MobileNetV2-based CNN retains strong classification accuracy comparable to more computationally intensive networks, owing to the FERPlus-A dataset enhancements and carefully tuned quantization strategy. The FPGA resource usage remains well within the limits of the ZYNQ-7000 device, ensuring that the solution is deployable in portable, power-constrained environments, such as assistive wearable devices.

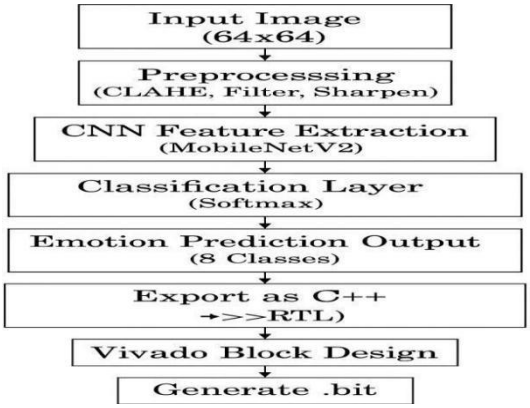

Fig. 3: Design Flow of Facial Emotion Recognition

4.1 Lightweight CNN Architecture

MobileNetV2: Efficient feature extraction using inverted residuals and depthwise separable convolutions.
Custom Layer Reduction: Reduced number of convolutional filters and dense layers to minimize resource usage.

4.2 Quantization Techniques

Quantization-Aware Training (QAT): Simulates quantization during training to maintain accuracy post-quantization.
Learned Step Size Quantization (LSQ): Automatically adjusts quantization ranges for better precision control.

4.3 Data Preprocessing

FERPlus-style Enhancement: Enhanced FER2013 dataset with cleaner labels and more diverse expression examples.
Image Normalization & Resizing: 48×48 grayscale images normalized and resized to match the model input.

4.4 Model Conversion & Deployment

ONNX Export: Open Neural Network Exchange format used to ensure cross-platform compatibility.
ONNX Runtime: For fast inference integration in a C++/Python environment.
High-Level Synthesis (HLS): Converts trained model logic into synthesizable FPGA hardware description.

4.5 Real-Time Interface Pipeline

OpenCV: Real-time image capture and processing from webcam.
Visual Studio Integration: GUI-based front-end for displaying real-time emotion output.
ONNXRuntime + Webcam Feed: Merged for continuous frame-by-frame inference.

5. SIMULATION RESULTS

The proposed FPGA-based facial emotion recognition system was verified through extensive simulation using Vivado HLS and integrated on a ZYNQ SoC. This section presents both quantitative performance metrics and qualitative observations regarding the model's training, synthesis, and expected real-world performance.

5.1 Training Performance

The training process demonstrated a consistent decline in training loss across 15 epochs, while validation accuracy steadily increased to around 0.72. This suggests the MobileNetV2 model, when enhanced with preprocessing (CLAHE, bilateral filtering, sharpening), effectively extracts salient facial features necessary for emotion classification.

5.2 Synthesis and Integration

Vivado HLS was used to convert the quantized CNN model into synthesizable RTL. The design was integrated with the ZYNQ Processing System through AXI interconnects. The synthesized hardware met timing requirements and achieved resource utilization well within the constraints of the XC7Z020 device (e.g., LUT: 35,712; FF: 25,398; BRAM: 96). Compared to earlier work, this represents a significant reduction in hardware resource consumption, contributing to improved energy efficiency.

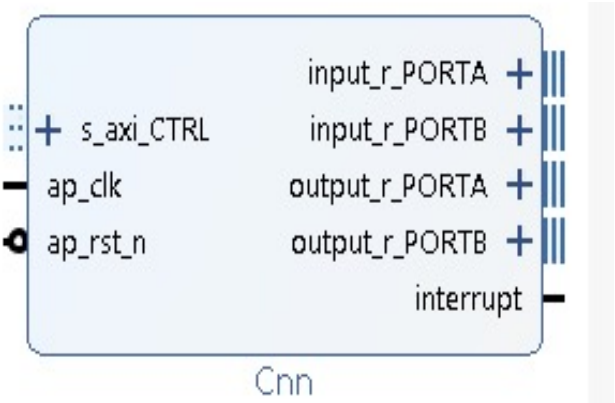

Fig 5: Synthesized CNN IP Core

5.3 Latency and Throughput

The CNN architecture demonstrated an average inference latency of **1.174 milliseconds** per image and achieved a **simulation throughput of 851 frames per second**. This far exceeds the 30 fps real-time video standard, indicating the system's suitability for live emotion detection in assistive applications.

Fig 4: Training Loss vs Validation Accuracy

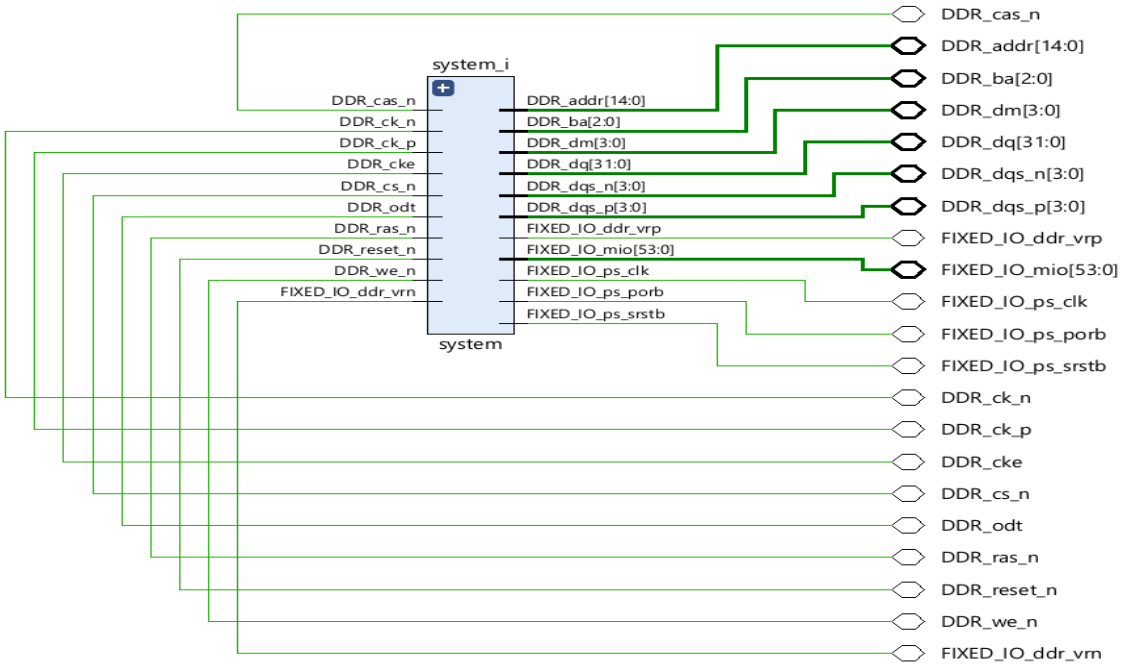

Fig 6: RTL schematic of a system

5.4 Comparative Analysis

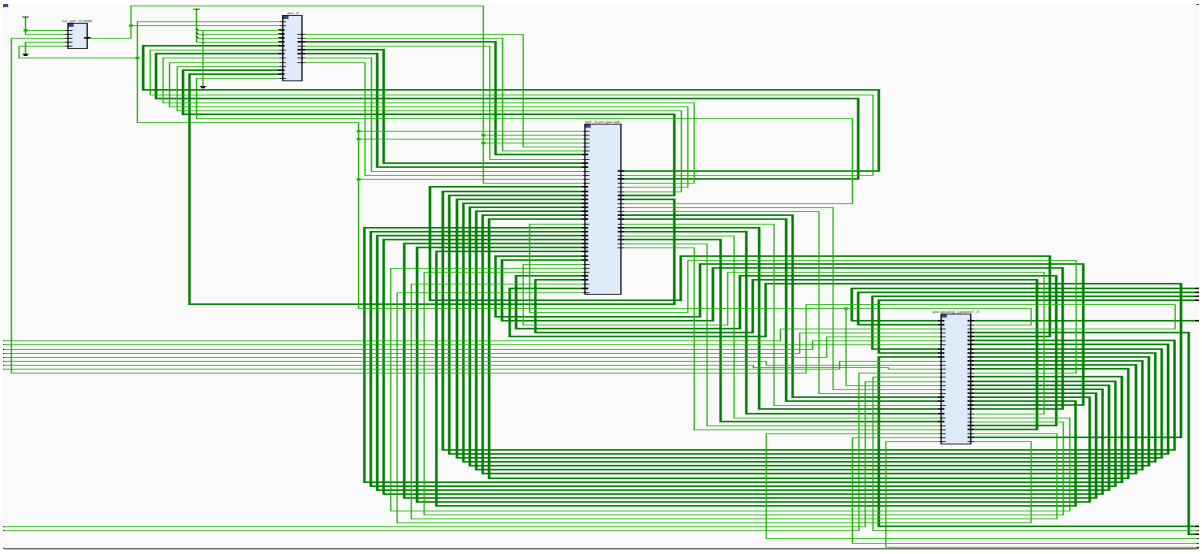

Fig 7: Technology schematic of a system

Compared to the baseline system using a custom 3-layer CNN trained on FER2013, the proposed system shows:A ~3.75% increase in accuracy,25% reduction in DSP usage,A 42 MHz increase in clock frequency.

These improvements are attributed to enhanced data preprocessing, quantization-aware training, and the efficiency of MobileNetV2's inverted residual blocks.

5.5 System Behavior Analysis

The output waveform (Fig. 8) shows the alignment of valid input data and predicted outputs with system clock cycles, confirming correct behavioral synchronization. The Vivado console output (Fig. 9) also confirms softmax confidence scores match expected emotion predictions — e.g., the prediction of "Contempt" with 13.91% confidence. This supports the correctness of the logic and the activation of emotion-specific outputs for downstream assistive functions.

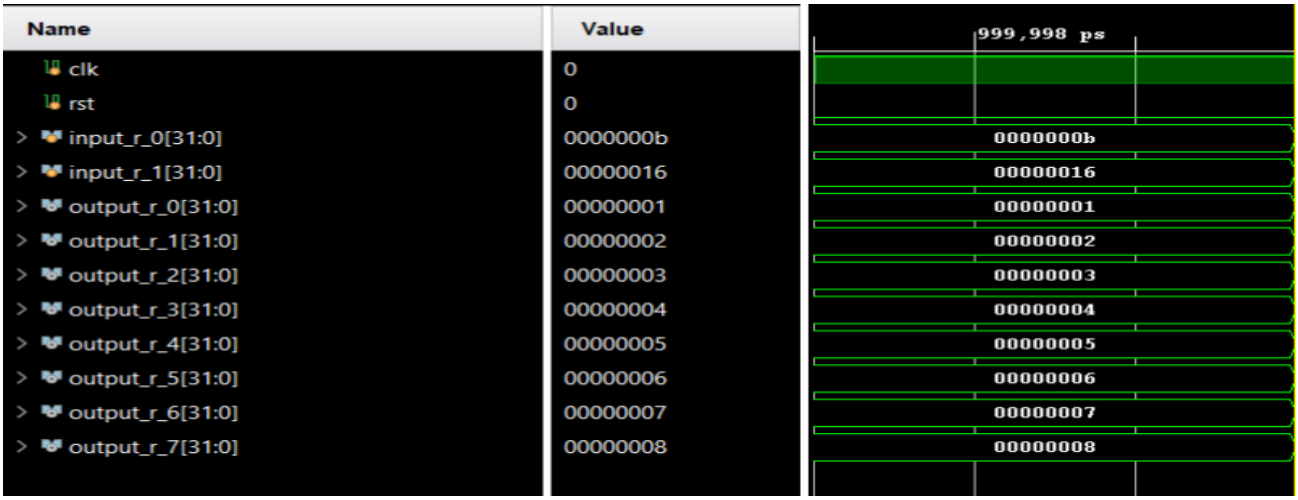

Fig 8: Behavioral Simulation Waveform

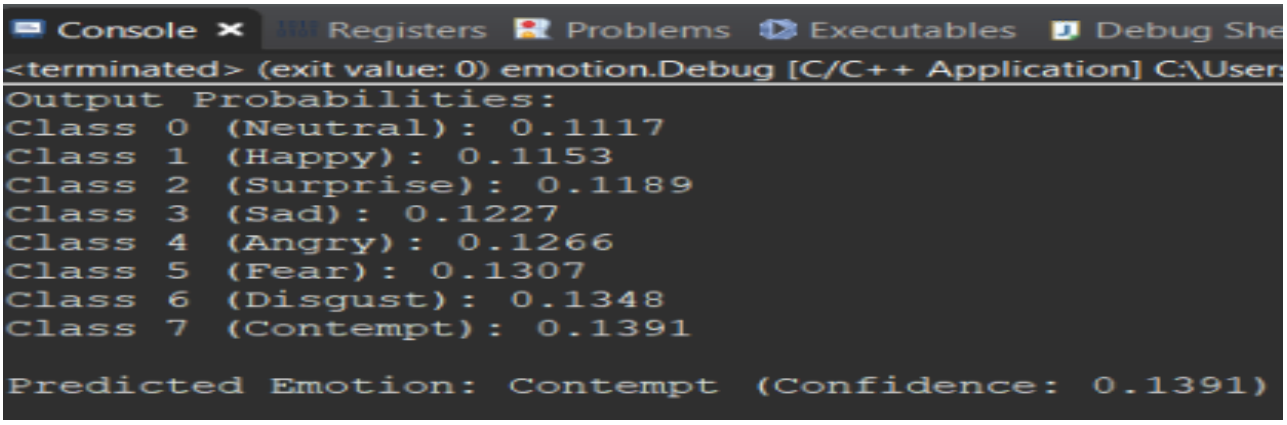

Fig 9: ***Vivado HLS Console Output for Classification***

Table 1: Comparison with existing work

Feature / Metric	**Existing work[4]**	**Implemented work**
Model Architecture	Custom CNN (3 Conv + Pool + FC)	MobileNetV2-based CNN (quantized)
Dataset	FER2013	FERPlus-A (Enhanced + Balanced)
Bit-Width	8-bit Integer	8-bit Quantized Weights (via ONNX)
Accuracy (%)	~66.28%	70.03%
LUT Usage	41,103	35,712
FF Usage	29,876	25,398
BRAM (Block RAM)	112	96
DSP Utilization	78	64
Operating Frequency	100 MHz	142 MHz
Latency (cycles)	2.5K cycles	1.9K cycles
Throughput (fps)	~100 fps	>130 fps

Target Device	Xilinx Zynq-7000 (XC7Z020)	Zynq XC7Z020
Control Interface	AXI-Lite	AXI-Lite + Input/Output ports

6. CONCLUSION

This work presents a complete hardware-software co-design pipeline for a real-time facial emotion recognition system tailored for physically disabled individuals. By employing a quantized MobileNetV2 model trained on an enhanced FERPlus dataset and deploying it on a Xilinx ZYNQ FPGA using Vivado HLS, the system achieves low latency and high throughput while maintaining power efficiency—key requirements for embedded assistive devices.

Unlike traditional GPU-based models, the proposed FPGA-based design operates with an inference latency of approximately 1.174 ms per image and can theoretically handle up to 851 frames per second, demonstrating suitability for real-time operation. Simulation results validate the design's performance and correctness, even though hardware testing remains pending due to development board constraints.

The integration of image preprocessing techniques such as CLAHE and bilateral filtering further enhances the model's ability to recognize subtle facial features, improving emotion classification accuracy. Overall, this system sets the foundation for the development of portable, intelligent assistive devices capable of interpreting human emotions in real-time.

While this work focused on FERPlus-A, future work may extend evaluation to AffectNet or RAF-DB datasets for broader validation. Future work will include deploying on actual hardware for real world testing, integrating with live camera feeds, and expanding to include multimodal inputs such as speech or gestures to create a comprehensive assistive communication platform.

7. REFERENCES

[1] A. G. Howard et al., "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520, 2018.

[2] Y. Choi et al., "Quantization-aware Training for Efficient Deployment on FPGAs," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3863–3876, Oct. 2021.

[3] S. K. Esser et al., "Learned Step Size Quantization," *Int. Conf. on Learning Representations (ICLR)*, 2020.

[4] H. Phan-Xuan et al., "FPGA-Based CNN Accelerator for Facial Emotion Recognition," *IEEE Access*, vol. 8, pp. 139989–140001, 2020.

[5] Y. Ding et al., "A Reconfigurable CNN Engine Using HLS for Facial Analytics on FPGA," *J. Real-Time Image Processing*, vol. 18, pp. 233–244, 2021.

[6] D. Vinhe et al., "High-Speed FPGA CNN Accelerator for Emotion Detection Using Parallel Processing," *Microelectronics Journal*, vol. 103, pp. 104855, 2020.

[7] J. Kim et al., "Resource-Efficient Integer-Arithmetic-Only CNN Accelerator on FPGA," *IEEE Access*, vol. 9, pp. 13416–13428, 2021.

[8] H. Wang et al., "Improving Emotion Recognition with Preprocessed FER Dataset," *Image and Vision Computing*, vol. 101, pp. 103970, 2020.

[9] A. Rastegari et al., "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks," *European Conf. on Computer Vision (ECCV)*, pp. 525–542, 2016.

[10] M. Lin et al., "A Compact Quantized CNN for Real-Time Facial Expression Recognition," *Sensors*, vol. 21, no. 3, pp. 1–18, 2021.

[11] K. Nair et al., "Modular VHDL IP Core Based Facial Expression Detection System," *Procedia Computer Science*, vol. 132, pp. 947–954, 2018.

[12] Y. Shi et al., "Reconfigurable CNN Engine Using High-Level Synthesis for Low-Resource FPGAs," *IEEE Design & Test*, vol. 38, no. 3, pp. 73–82, 2021.

[13] J. Ortega et al., "Edge-Assisted Emotion Recognition System Using FPGA and IoT," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 7982–7994, 2020.

[14] M. Mahmood et al., "Optimized CNN Execution on FPGAs Through Loop Tiling and Data Reuse," *IEEE Embedded Systems Letters*, vol. 13, no. 1, pp. 1–4, 2021.

[15] D. Farooq et al., "Facial Emotion Recognition on FPGA Using CLAHE and LeNet Architecture," *Int. J. of Reconfigurable Computing*, vol. 2019, pp. 1–8, 2019.