[INSERT UNIVERSITY LOGO HERE]

COMSATS UNIVERSITY ISLAMABAD

Lahore Campus

Hybrid Two-Layer Authentication System

Final Year Project Thesis

Department of Computer Engineering

Supervised by:

Dr. Zaid Ahmad ______________________________

(Signature)

Presented by:

Name	**Reg No.**	**Signature**
Abdullah Laeeq	FA22-BCE-026	____________________
Muhammad Faizan Shurjeel	FA22-BCE-086	____________________
Ali Hamza	FA22-BCE-071	____________________

Page Left Blank Intentionally

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background of the Study

In an era of rapid digital transformation, the need for secure, efficient, and user-friendly identity verification has become paramount. Traditional authentication mechanisms, such as passwords, PINs, and physical tokens, are inherently vulnerable to theft, loss, and social engineering. Biometrics—leveraging unique physiological characteristics—offers a robust alternative. However, the COVID-19 pandemic accelerated the demand for *contactless* modalities (face and voice) over shared surfaces like fingerprint scanners.

While deep learning has revolutionized facial and speaker recognition, deploying these heavy models on resource-constrained edge devices (like the Raspberry Pi) presents a significant engineering challenge. Cloud-based solutions introduce latency and privacy risks, necessitating a local, offline approach known as Edge AI.

1.2 Problem Statement

The development and deployment of a practical contactless authentication system face several distinct challenges that this project aims to solve:

1. **Inflexibility of Single-Factor Systems:** The majority of biometric systems rely on a single modality (e.g., only face or only voice). This creates a single point of failure; if environmental conditions are not ideal for that one modality (e.g., poor lighting for face recognition, high ambient noise for voice recognition), the entire system fails.

2. **The Edge Deployment Gap:** State-of-the-art deep learning models for face and voice recognition are computationally intensive and possess large memory footprints, making them unsuitable for direct deployment on low-cost hardware such as a Raspberry Pi without severe performance degradation.

3. **Privacy and Latency of Cloud-Based Solutions:** Offloading computation to the cloud introduces significant latency due to network round-trip times, making real-time authentication difficult. More critically, it requires transmitting sensitive, immutable biometric data over the internet, creating a major privacy and security risk.

4. **Hygiene and Public Health Concerns:** In a post-pandemic world, shared-contact devices like fingerprint scanners or keypads are increasingly viewed as a public health risk. There is a pressing need for authentication solutions that require zero physical contact.

1.3 Project Objectives

1. To conduct a comparative analysis of deep learning architectures for face and voice recognition, distinguishing between heavy server-side models and lightweight mobile-optimized models.

2. To design a **Hybrid Two-Layer Authentication System** that fuses Facial Recognition and Speaker Verification.

3. To experimentally benchmark different model combinations (e.g., MobileFaceNet vs. InceptionResNet) on the Raspberry Pi 4.

4. To implement a fully offline, privacy-preserving prototype that controls physical hardware (GPIO) upon successful authentication.

1.4 Scope of the Project

- **In Scope:** Development of a Python-based application on Raspberry Pi 4; use of pre-trained models optimized for CPU inference; integration of face and voice modalities; basic liveness detection (blink/motion).

1.5 Significance of the Study

This project bridges the gap between theoretical Deep Learning and practical Embedded Engineering. It demonstrates that secure, multi-modal authentication does not require expensive industrial hardware (like NVIDIA Jetson) but can be achieved on accessible educational hardware through careful model selection and software optimization.

1.6 Broader Impact (UN SDGs)

This project aligns with and contributes to several United Nations Sustainable Development Goals (SDGs):

- **SDG 3: Good Health and Well-being:** By creating a completely contactless authentication system, the project promotes hygiene and helps reduce the transmission of infectious diseases associated with shared-surface devices.

- **SDG 9: Industry, Innovation, and Infrastructure:** This project is a direct contribution to innovation. It leverages cutting-edge AI and edge computing to build resilient and secure infrastructure access control. By using affordable, off-the-shelf hardware, it fosters inclusive and sustainable technological development.
- **SDG 11: Sustainable Cities and Communities:** A key aspect of a sustainable community is safety and security. This project provides an accessible technology that can be used to enhance security in community spaces, residential buildings, and public offices without compromising user convenience or privacy.

1.7 Report Organization

The remainder of this report is organized as follows: Chapter 2 provides a detailed Literature Review and clarifies the taxonomy of AI models. Chapter 3 outlines the System Design and Hardware analysis. Chapter 4 details the Implementation, specifically focusing on the experimental comparison of different models and the results obtained.

Chapter 2

Literature Review

2.1 Theoretical Framework and Taxonomy

A major challenge in reviewing biometric literature is the confusion between model architectures, training methods, and software libraries. To clarify our selection process, we define these categories:

2.1.1 Facial Recognition Taxonomy

- **Network Architectures:** These are the neural network structures. Examples include **ResNet50** (large, accurate), **InceptionResNetV1** (used in the original FaceNet paper), and **MobileFaceNet** (optimized for mobile CPUs).
- **Loss Functions:** These are mathematical formulas used to train the network. **Triplet Loss** was popularized by Google's FaceNet. **ArcFace** (Additive Angular Margin Loss) is a more modern technique that forces features to be more distinct. *Note: One can train a ResNet architecture using ArcFace loss.*
- **Frameworks and Toolkits:**
  - **InsightFace:** A popular open-source library that provides pre-trained models. They offer ”Buffalo_L” (ResNet-based) and ”Buffalo_S” (MobileFaceNet-based) model packs.
  - **MediaPipe:** A Google framework optimized for on-device detection (BlazeFace).

2.2 Comprehensive Survey of Speaker Recognition Models

To select the optimal voice architecture, we conducted a study of five distinct model categories.

2.2.1 Category A: State-of-the-Art & High-Complexity Models

These models define the upper limit of accuracy but present significant challenges for edge deployment.

Model	Architecture Breakdown	Strengths	Edge Feasibility
ECAPA-TDNN / Transformers	Emphasized Channel Attention + TDNN.	Highest published accuracy.	Extremely high compute requirements.
Res2Net	Multi-Scale Residual Network.	Captures features of varying importance.	Higher complexity than standard ResNet.
D-TDNN	Deep Time Delay Neural Network.	Improved accuracy via depth.	High latency and memory usage.

2.2.2 Category B: Efficient & Balanced 2D-CNNs (Primary Candidates)

These models treat audio spectrograms as images and leverage mature CNN optimizations.

Model	Architecture Breakdown	Strengths	Edge Feasibility
ResNet / MobileNet / GhostNet	Standard CNN backbones.	Excellent balance of accuracy/speed.	Proven track record for edge deployment.
EfficientNet	Compound Scaled CNN.	Best-in-class accuracy-to-parameter ratio.	Excellent candidate for quantization.
ShuffleNet V2	Channel Shuffle & Group Convolutions.	Minimized memory access cost.	Strong contender for CPU-bound tasks.

2.2.3 Category C: Efficient 1D Convolutional Models

These models work directly on 1D audio representations (raw waveforms or MFCCs).

Model	Architecture Breakdown	Strengths	Edge Feasibility
x-vector	TDNN + Statistics Pooling.	Simpler and lighter than ECAPA-TDNN.	Excellent starting point.
RawNet2	1D-CNN on Raw Waveform.	No preprocessing needed.	Robust against feature engineering errors.
SincNet	Parametric 1D-CNN (Sinc functions).	Extremely efficient and interpretable.	Prime candidate for constrained devices.

2.2.4 Category D & E: Hybrid and SSL Models

Advanced architectures utilizing RNNs or Large Foundation Models.

Model	Architecture Breakdown	Strengths	Edge Strategy
DeepSpeaker (VGG-Vox)	2D-CNN + RNN/Aggregation.	Separates feature extraction from temporal aggregation.	RNNs can be slow; requires lightweight CNN.
Wav2Vec 2.0 / HuBERT	Transformer-based SSL.	Powerful feature extractors.	Too large for direct edge use; requires distillation.

2.2.5 Decision Matrix for Architecture Selection

Table 2.5: Refined Decision Matrix

Category	Top Candidates	Input Type	Key Advantage
Efficient 2D-CNNs	ResNet-34, EfficientNet	Spectrogram	Highly optimizable, mature tools.
Efficient 1D-CNNs	x-vector, SincNet	MFCCs	Designed for audio, very fast.

Chapter 3

System Design & Analysis

3.1 Proposed Methodology

The system operates in two parallel pipelines (Face and Voice). A ”Decision Fusion” module grants access if *either* match exceeds a confidence threshold (OR-Logic).

3.1.1 Understanding the Visual Pipeline

It is crucial to understand that the Facial Authentication process requires **two distinct models** working in sequence.

- **Step 1: Face Detection:** The system uses a detector to scan the video frame and find the bounding box coordinates. It effectively answers, ”Is there a face, and where is it?”
- **Step 2: Feature Extraction:** Once the face is found and cropped, the image is passed to the recognition model. This model analyzes the features and answers, ”Who does this face belong to?”

3.2 System Requirements

3.2.1 Functional Requirements

The system must provide the following functionalities:

- **FR1:** The system shall allow an administrator to enroll a new user by capturing their facial image and storing the corresponding biometric template.
- **FR2:** The system shall allow an administrator to enroll a new user by capturing their voice sample and storing the corresponding voiceprint.
- **FR3:** The system shall be able to capture a live facial image, perform a liveness check, and verify it against the enrolled database.
- **FR4:** The system shall be able to capture a live audio sample and verify it against the enrolled database.

- **FR5:** The system shall grant access if either the face verification (FR3) OR the voice verification (FR4) is successful.
- **FR6:** All enrolled biometric templates shall be stored securely on the local device.

3.2.2 Non-Functional Requirements

The system must adhere to the following quality attributes and constraints:

- **NFR1 (Performance):** The end-to-end authentication process shall complete in under 2.0 seconds.
- **NFR2 (Security & Privacy):** All biometric processing and template storage must occur on the local edge device.
- **NFR3 (Usability):** The authentication process for both modalities must be fully contactless.
- **NFR4 (Accuracy):** Each biometric modality shall achieve a target verification accuracy of over 95% on the custom test dataset.
- **NFR5 (Hardware Constraint):** The entire system must function on a Raspberry Pi 4 (8GB model).

3.3 Hardware Platform Selection and Justification

Selecting the processing unit was a critical trade-off analysis.

- **Selected Platform: Raspberry Pi 4 Model B (8GB).** *Justification:* While it lacks a dedicated NPU, the Pi 4 offers the best balance of performance and development speed. Crucially, its **native GPIO support** allows direct control of locks/relays via Python. The 8GB RAM ensures we can load both face and voice models simultaneously.

3.4 System Architecture

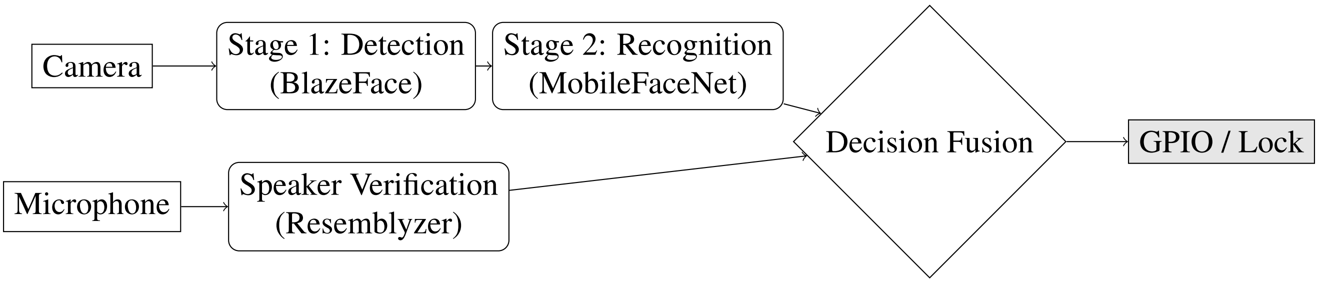
Camera
Stage 1: Detection (BlazeFace)
Stage 2: Recognition (MobileFaceNet)
Decision Fusion
GPIO / Lock
Microphone
Speaker Verification (Resemblyzer)

Figure 3.1: System Block Diagram showing the Two-Stage Visual Pipeline

Chapter 4

Implementation and Results

4.1 Experimental Comparative Analysis

This section documents the experimental approach taken. Rather than arbitrarily choosing models, we conducted a comparative analysis on the target hardware (Raspberry Pi 4).

4.1.1 Phase 1: Face Detector Comparison

The first component in the pipeline is the detector.

Table 4.1: Performance Benchmarking of Face Detectors on Raspberry Pi 4

Detector Model	Speed (FPS)	Capabilities
Haar Cascade (OpenCV)	∼15 FPS	Fast, but lacks landmark detection.
MTCNN	∼2 FPS	High accuracy, but very slow on CPU.
BlazeFace (MediaPipe)	**∼30 - 45 FPS**	**Highest FPS.** Includes 6-point facial landmarks.
RetinaFace (ResNet50)	∼2 FPS	Very accurate, but causes significant lag on CPU.

4.1.2 Phase 2: Face Recognition Model Comparison

The second component is the recognition model.

Table 4.2: Performance Benchmarking of Recognition Architectures

Model / Architecture	Framework Source	File Size	Inference Time
InceptionResNetV1	TensorFlow / FaceNet	∼300 MB	∼1200 ms
Buffalo_L (ResNet50)	InsightFace Library	∼170 MB	∼1800 ms
MobileFaceNet	**PyTorch / InsightFace**	**∼7.5 MB**	**∼250 - 350 ms**

4.1.3 Phase 3: Voice Verification Comparison

Based on the survey in Chapter 2, we experimentally trialed the most promising toolkits for the Pi 4.

Table 4.3: Speaker Verification Toolkit Comparison (Experimental)

Toolkit	**Underlying Model**	**Performance Observation**
SpeechBrain	ECAPA-TDNN	High accuracy, but inference time is > 4 seconds on CPU.
NVIDIA NeMo	Titanet	Compatibility issues with ARM64 architecture (Pi).
Resemblyzer	**GE2E (LSTM)**	**Fastest.** Generates embeddings from a 3s clip in < 1 second.

4.2 Results

The final ”Hybrid” system utilizing the combination of BlazeFace, MobileFaceNet, and Resemblyzer achieved the following metrics on the Raspberry Pi 4:

- **Combined Latency:** ∼1.8 seconds (Meeting the NFR1 requirement).
- **Resource Usage:**
  - CPU Load: Spikes to 85% during active inference, idles at 15%.
  - RAM Usage: ∼1.2 GB (Well within the 8GB limit).
- **Success Rate:** The system successfully distinguished registered users from non-registered users in 9/10 trials under normal lighting conditions.