PUFshield: A Hardware-Assisted Approach for Deepfake Mitigation Through PUF-Based Facial Feature Attestation

Presenter: Venkata K. V. V. Bathalapalli

Venkata K. V. V. Bathalapalli[1], Venkata P. Yanambaka[2], S. P. Mohanty[3], E. Kougianos[4]

University of North Texas, Denton, TX, USA.[1,3,4] and

Texas Woman's University[2].

Email: **vb0194@unt.edu[1], vyanambaka@twu.edu[2], saraju.mohanty@unt.edu[3], elias.kougianos@unt.edu[4],**

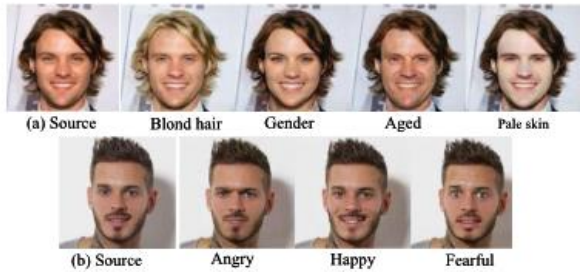
Outline

- Introduction to Deepfake
- Deepfake Techniques and Classification
- Deepfake Mitigation
- Introduction to PUF
- Proposed PUF-based Facial Feature Attestation Scheme
- Experimental Validation
- Conclusion & Future Research Directions

Deepfake

AI can be fooled by fake data

Attribute Manipulation



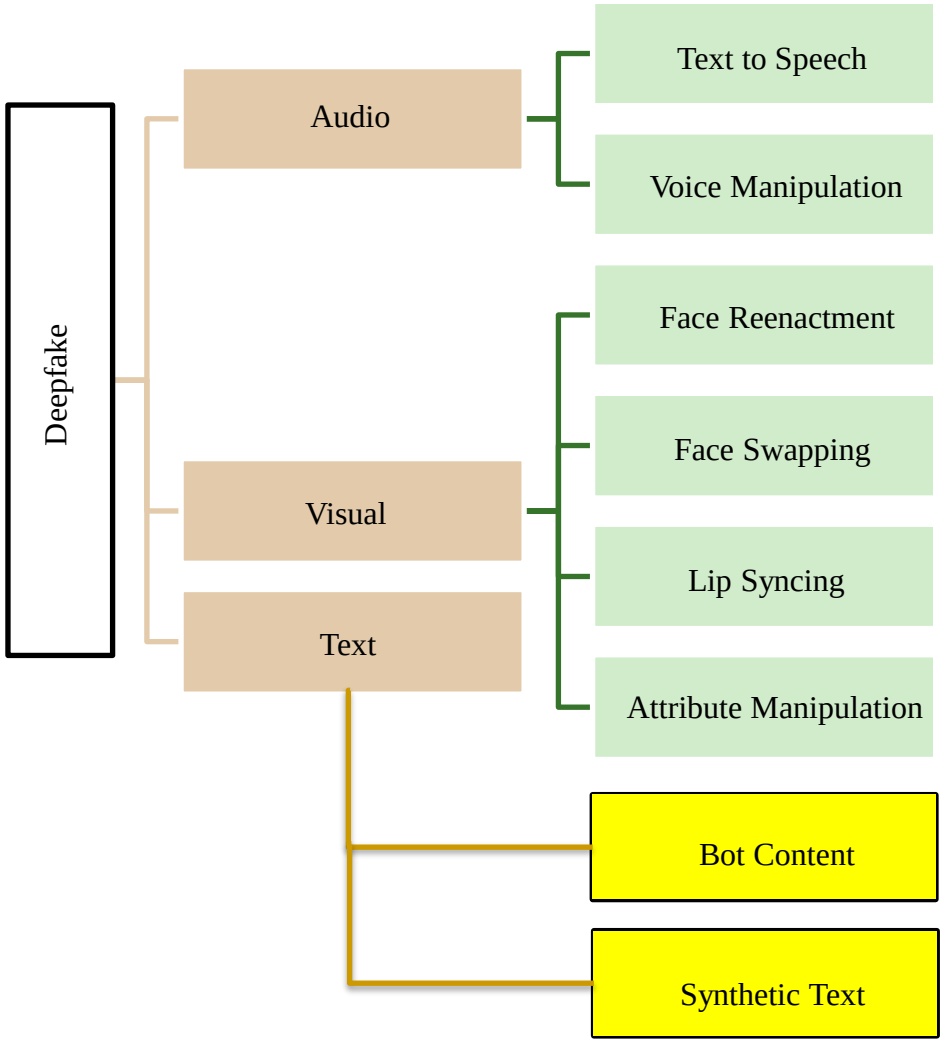
Identity Swapping

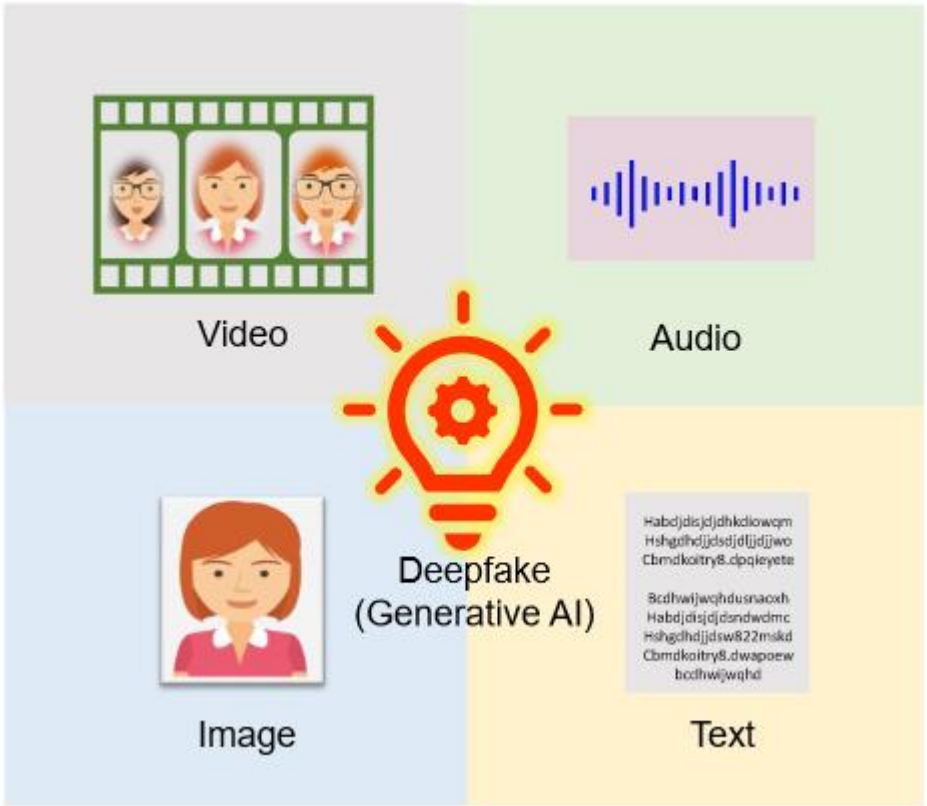
AI can create fake data (Deepfake)

1. Deepfake refers to super realistic, but fake images, sounds, and videos generated by machine learning methods.
2. Deepfake leverages a Generative adversarial network (GAN) which enables the modification of human faces in a video or image.
3. Deepfakes can be classified as Audio, Visual and Text

Source: A. Malik, M. Kuribayashi, S. M. Abdullahi and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," in *IEEE Access*, vol. 10, pp. 18757-18775, 2022, doi: 10.1109/ACCESS.2022.3151186.

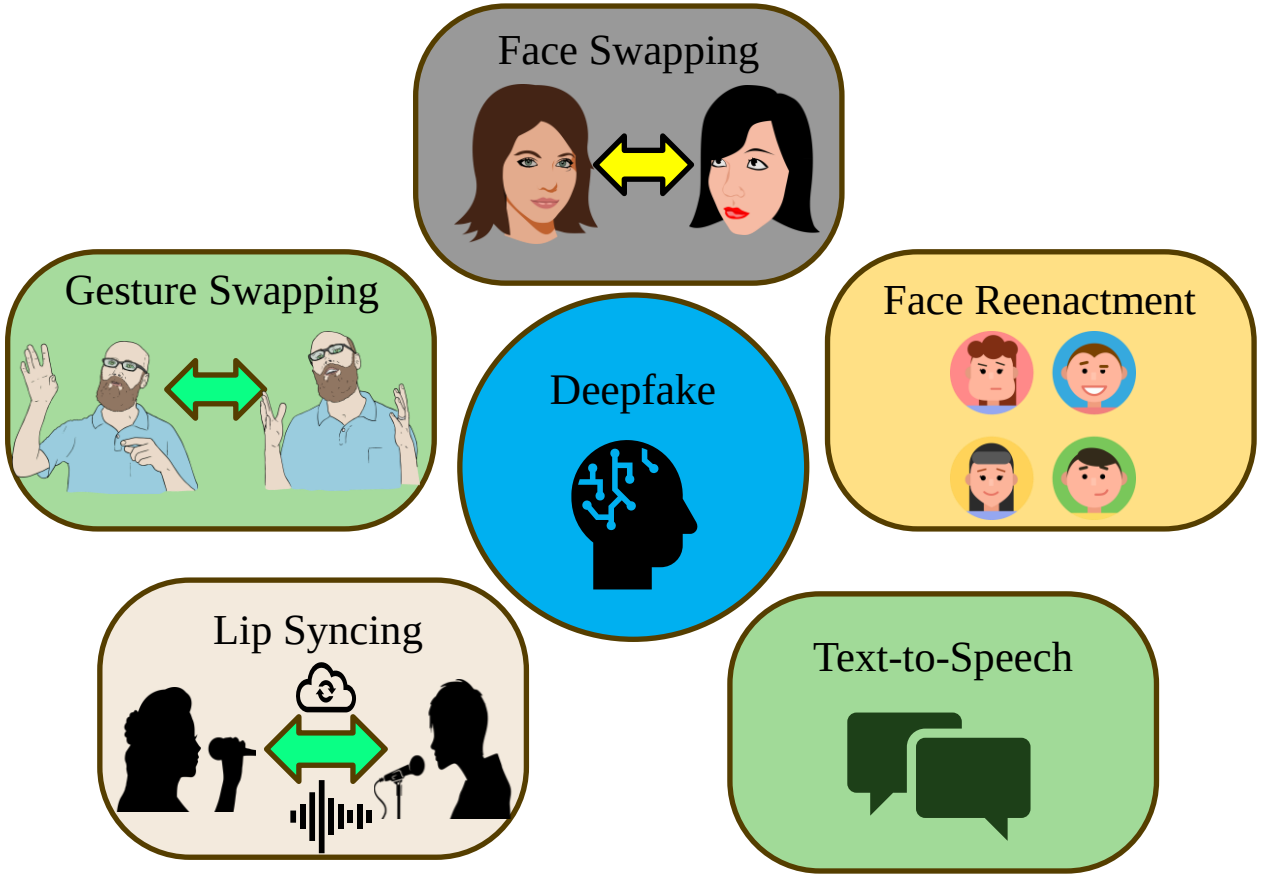
Deepfake Classification



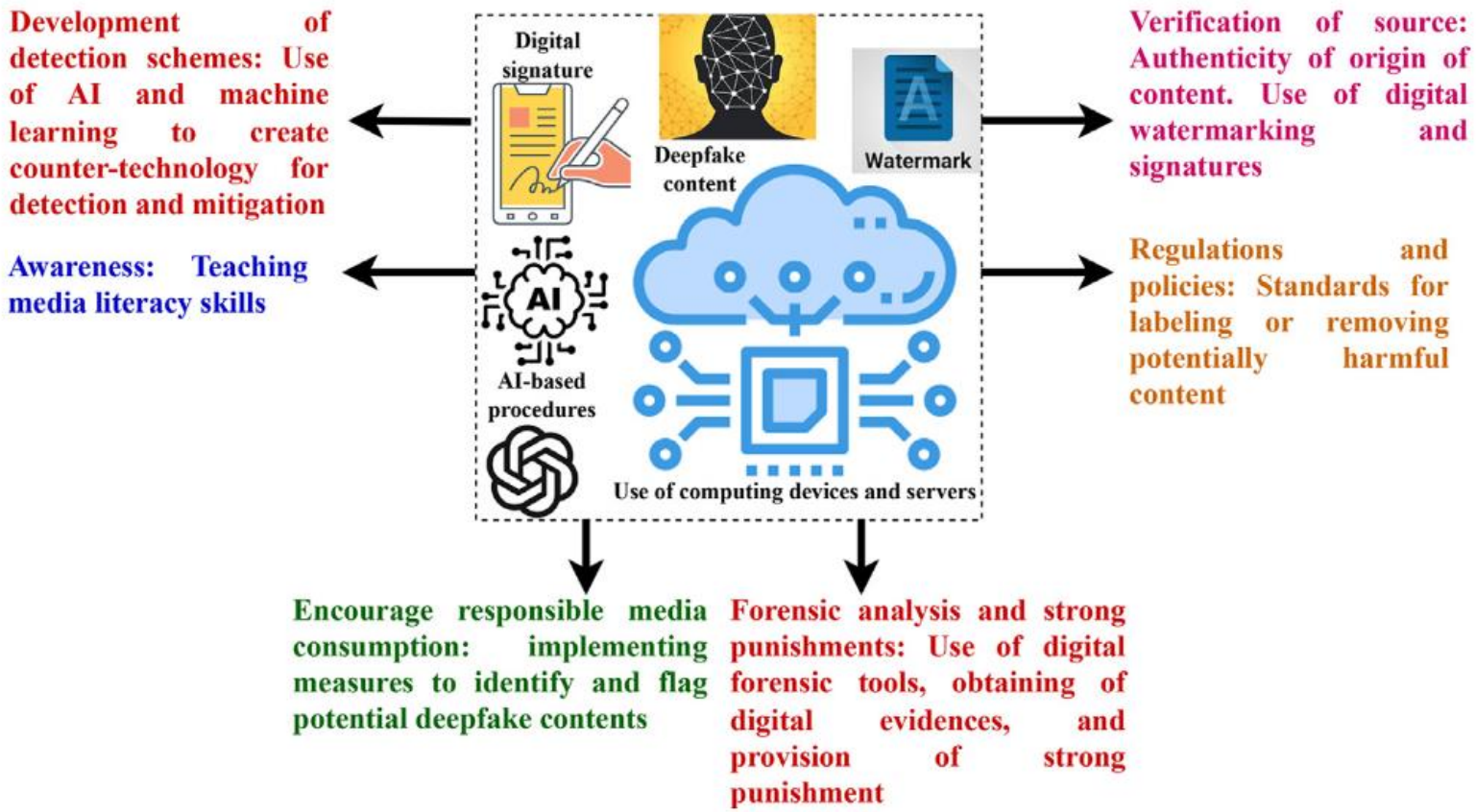


Source: A. Mitra, **S. P. Mohanty**, and E. Kougianos, "The World of Generative AI: Deepfakes and Large Language Models", *arXiv Computer Science*, arXiv:2402.04373, Feb 2024, 9-pages.

Deepfake Techniques

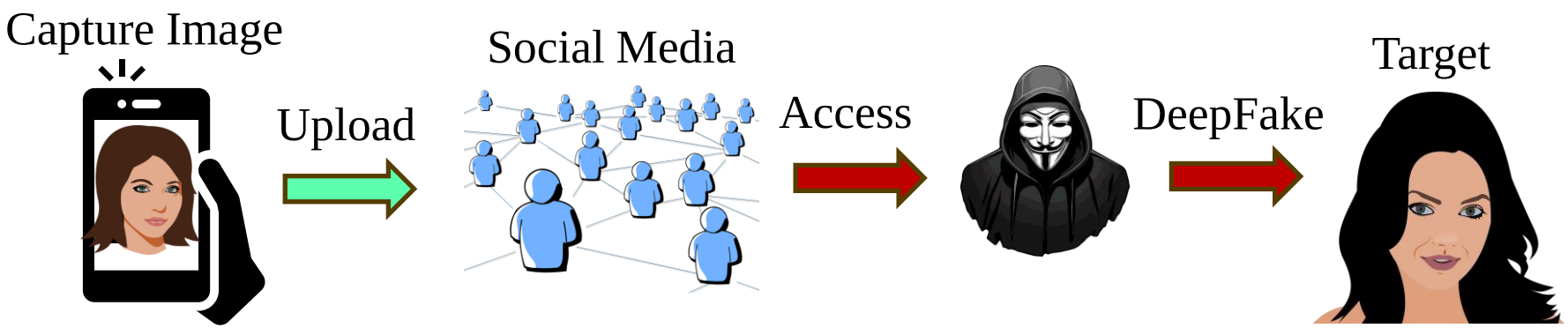


Deepfake Mitigation



Source: Wazid, M., Mishra, A. K., Mohd, N., & Das, A. K. (2024). A Secure Deepfake Mitigation Framework: Architecture, Issues, Challenges, and Societal Impact. *Cyber Security and Applications*, 100040.

Threat Model



Addressing visual Deepfake of individual content captured as a video/image is important and necessary to counter facial attribute manipulation which includes modifying facial attributes like eyes, nose, lips and replacing them with target's attributes.

Related Research

Work	Approach	Technique	Methodology	Tools	Features
Kato et.al [5]	Mitigation	Visual	Scapegoat Image Generation	StyleGAN2	Privacy and Anonymity
Zheng et.al [23]	Mitigation	Visual	PUF-based device and data hash	CMOS Image sensor	Image content authenticity
Krause et. al [8]	Detection	Audio	Language and phoneme focused	Logistic regression	Detection using mouth movements
Pishori et.al [15]	Detection	Visual	Eye Blink rate	CNN+RNN, OpenCV	Efficient through eye blink rate detection
Wang et.al [17]	Mitigation	Visual	GAN based secret message embedding in an image	GAN	Personal photo protection
Zhao et.al [22]	Detection	Visual	Image watermarking	Neural network with encoder and decoder	Effective image quality preservation
Ashok et.al [16]	Detection	Visual	Training XceptionNet using faceforenscis++ dataset	XceptionNet Model	Identifying Deepfake from Original content
Doan et.al [2]	Detection	Audio	Identifying silence, breathing, talking in an Audio	RawNet2	Biological sound-based detection
PUFshield (Current Work)	Mitigation	Visual	PUF-based Facial Feature Attestation	PUF, Dlib Facial detection and landmark prediction	Image and device integrity

Novel contributions

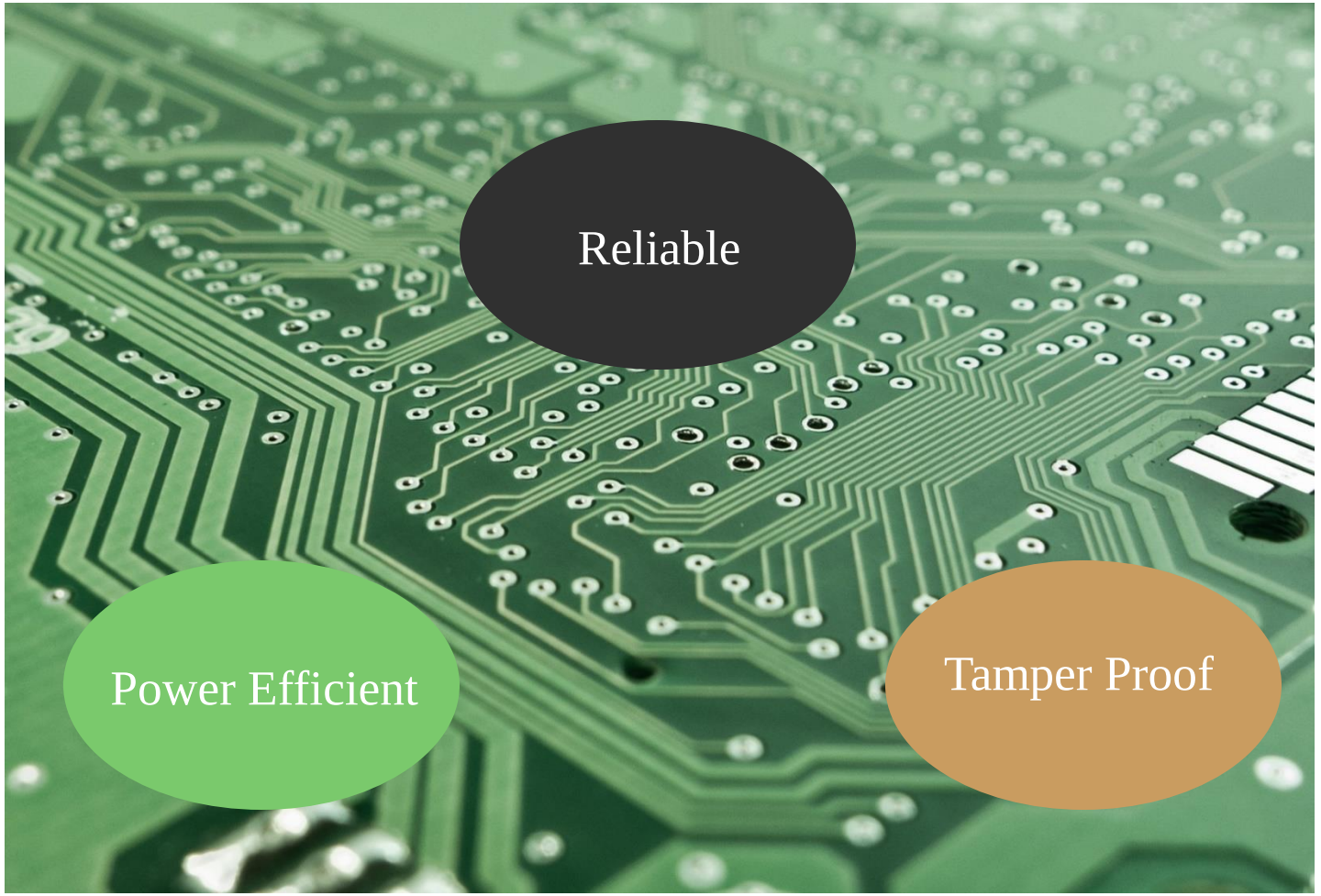
- A secure digital content integrity verification scheme through hardware enabled attestation.
- Presenting a state-of-art PUF-based approach for digital content attestation.
- A state-of-art solution for countering facial attribute manipulation to prevent visual Deepfakes.
- A device security framework providing PUF-based digital fingerprint for the camera capturing image/video.
- An approach to counter Deepfakes countering facial attribute manipulation.

Physical Unclonable Function (PUF)-Introduction

Why PUFs?

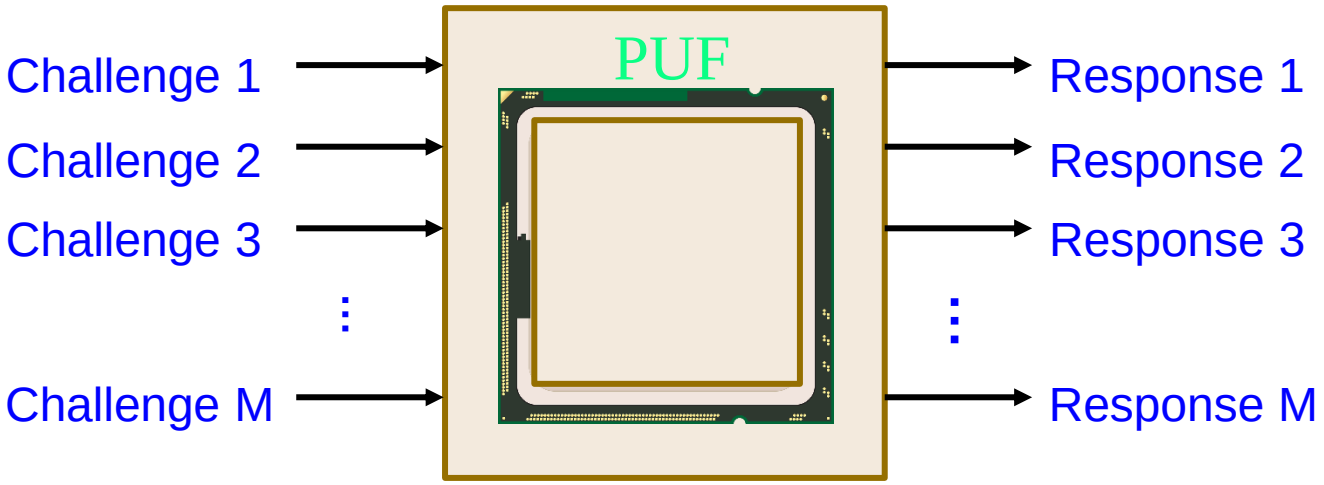
- Hardware-assisted security.
- Key not stored in memory.
- Not possible to generate the same key on another module.
- Robust and low power consuming.
- Can use different architectures with different designs

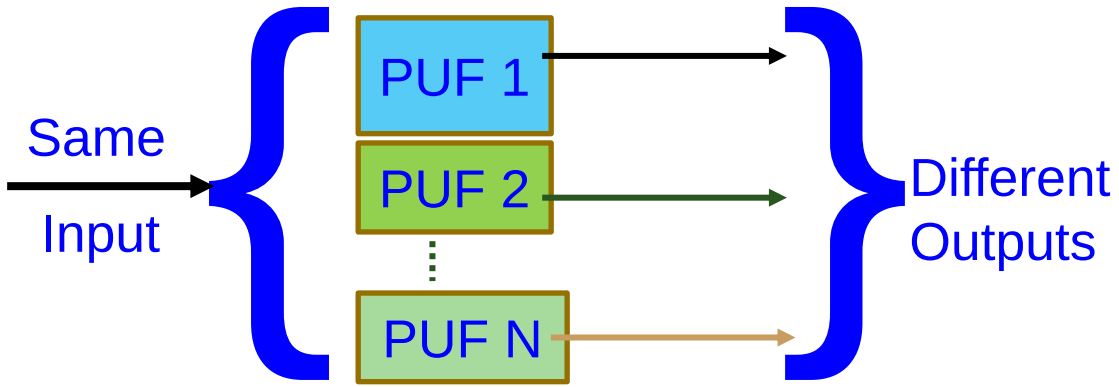
PUF: A Hardware-Assisted Security Primitive



- A secure fingerprint generation scheme based on process variations in an Integrated Circuit
- PUFs don't store keys in digital memory, rather derive a key based on the physical characteristics of the hardware; thus secure.
- A simple design that generates cryptographically secure keys for the device authentication

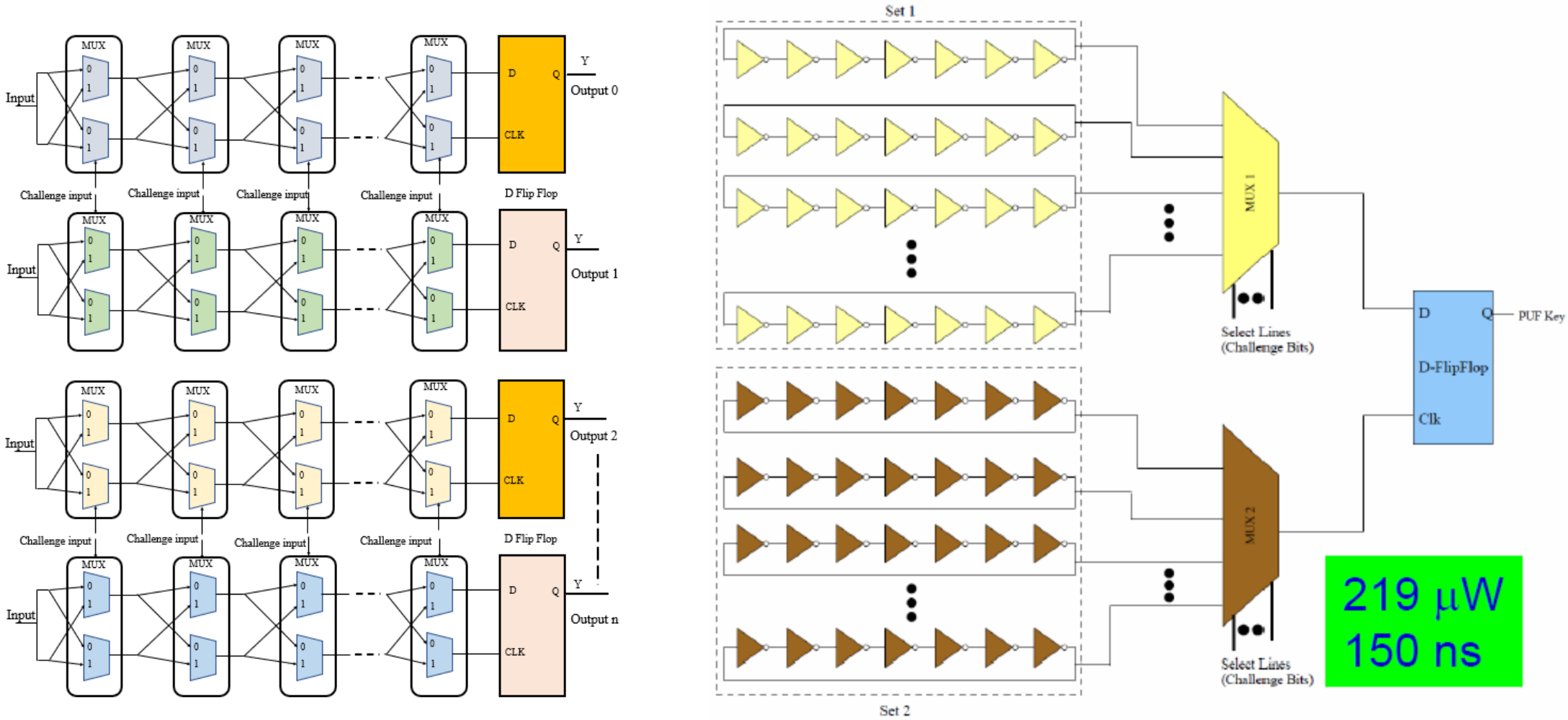
PUF Key Generation and Working





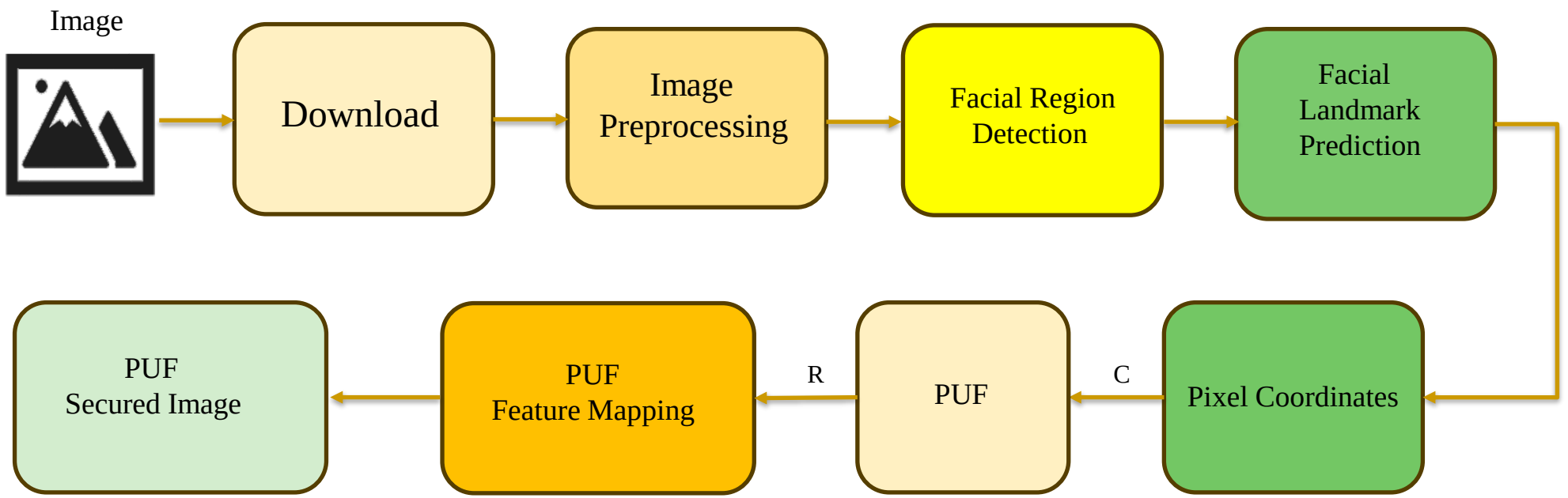
Source: International Symposium on Smart Electronics Systems (iSES) 2019 Demo (PUFchain: Hardware-Integrated Scalable Blockchain)

PUF Designs



Source: iSES 2019 Demo (PMsec: PUF-Based Energy-Efficient Authentication of Devices in the Internet of Medical Things (IoMT))

PUFshield: Proposed Deepfake Mitigation Technique



Facial Landmarks Coordinates in Dlib

Facial Landmarks	Pixel Coordinates
Left Eye	36-41
Right Eye	42-47
Left Eyebrow	17-21
Right Eyebrow	22-26
Jaw	0-16
Nose Bridge	27-30
Lower Nose	31-35
Outer Lip	48-59
Inner Lip	60-67

Working Flow of PUFshield:

Step 1 : Capture Image

Step 2 : Perform Image Preprocessing

Image → 600 X500

Image → Gray Scale

Step 3 : Perform Facial Region (RoI) Detection

Histogram of Gradients → RoI

Step 4 : Access PUF at the Camera

Step 5 : Obtain Facial Landmarks Pixel Coordinates

Step 6 : Facial Landmarks→ PUF→R1

Extract for a set of 8 coordinates at a time

Extract for all 68 facial landmarks R1-----R17

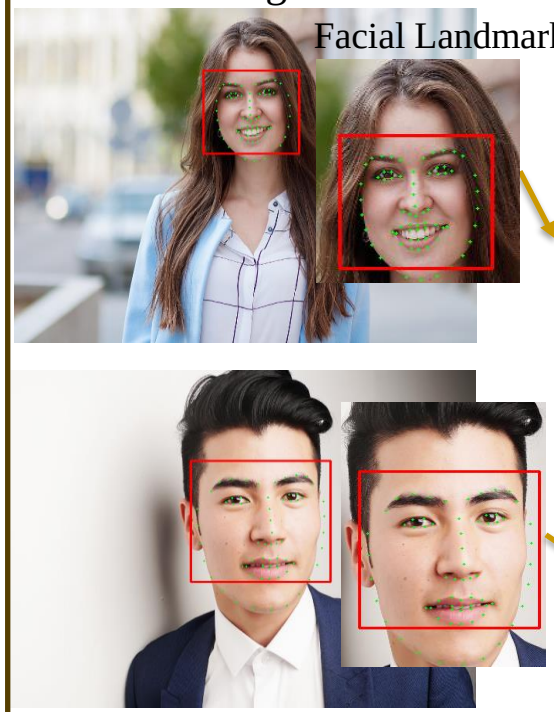
Perform XOR Operation of all facial coordinates

Step 7 : Final image fingerprint is final XORed output

Experimental Validation of PUFshield

Images

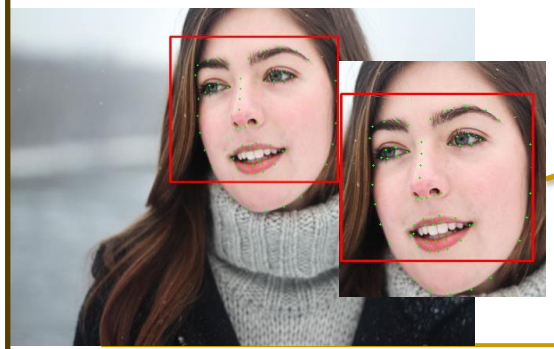
Facial Landmarks



Facial Landmark Coordinates

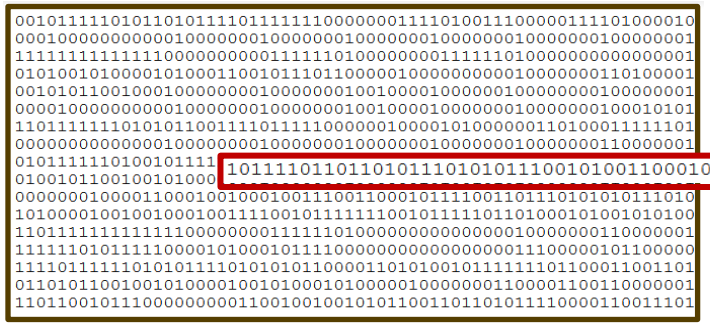
[119, 235, 124, 266, 131, 297, 142, 328, 157, 357, 179, 383, 210, 402, 239, 417, 274, 422, 307, 413, 333, 396, 356, 373, 371, 344, 376, 311, 378, 277, 381, 245, 379, 212, 146, 199, 161, 182, 184, 175, 209, 175, 232, 182, 273, 179, 294, 169, 318, 166, 342, 171, 359, 187, 254, 193, 257, 209, 259, 226, 262, 243, 236, 270, 249, 271, 263, 273, 276, 269, 289, 267, 175, 208, 190, 201, 204, 199, 221, 206, 205, 208, 190, 209, 290, 202, 305, 193, 320, 193, 335, 200, 321, 202, 306, 202, 211, 327, 229, 312, 251, 301, 267, 304, 281, 299, 301, 308, 321, 320, 304, 340, 284, 350, 270, 353, 254, 352, 232, 344, 220, 327, 252, 313, 268, 314, 281, 311, 312, 321, 283, 333, 269, 336, 253, 334]

[242, 205, 243, 230, 246, 257, 251, 282, 260, 306, 275, 326, 292, 342, 314, 353, 337, 355, 357, 348, 373, 331, 386, 310, 396, 288, 402, 263, 404, 240, 405, 216, 404, 194, 260, 179, 271, 165, 287, 160, 304, 163, 320, 168, 342, 166, 355, 159, 369, 155, 383, 157, 391, 168, 333, 188, 335, 205, 336, 222, 338, 240, 320, 255, 329, 257, 337, 258, 344, 256, 351, 253, 279, 195, 287, 189, 298, 188, 307, 194, 299, 197, 288, 198, 352, 191, 360, 184, 370, 183, 377, 188, 371, 192, 362, 193, 300, 290, 313, 282, 326, 278, 335, 281, 345, 278, 356, 281, 366, 285, 357, 298, 346, 306, 335, 308, 325, 308, 312, 302, 305, 290, 326, 288, 336, 289, 345, 287, 360, 287, 345, 291, 335, 293, 326, 292]

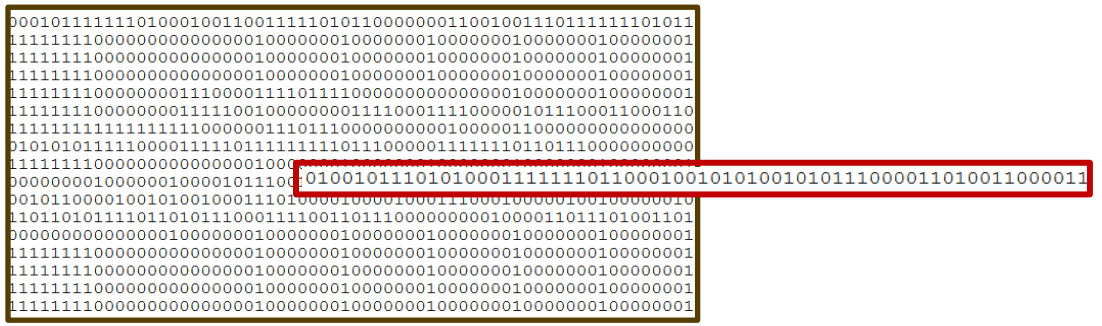


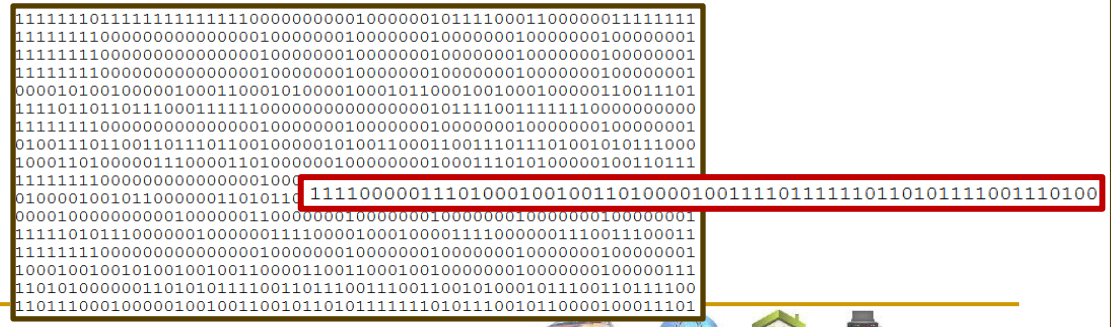
[245, 134, 244, 159, 247, 184, 253, 210, 263, 235, 275, 259, 290, 281, 309, 296, 331, 300, 355, 294, 377, 279, 395, 257, 409, 231, 417, 201, 421, 170, 423, 137, 421, 107, 241, 99, 244, 84, 257, 78, 271, 77, 284, 82, 309, 74, 328, 63, 349, 58, 371, 63, 386, 76, 299, 104, 299, 119, 298, 134, 297, 150, 293, 176, 299, 177, 305, 176, 313, 173, 321, 170, 254, 124, 259, 114, 270, 111, 283, 117, 272, 122, 261, 125, 331, 107, 339, 97, 352, 95, 365, 101, 355, 106, 342, 108, 287, 224, 291, 211, 301, 203, 310, 203, 319, 199, 337, 202, 358, 210, 343, 231, 328, 242, 318, 245, 308, 245, 296, 239, 290, 222, 303, 211, 312, 210, 321, 208, 353, 211, 324, 229, 315, 232, 305, 232]

PUF Attestation



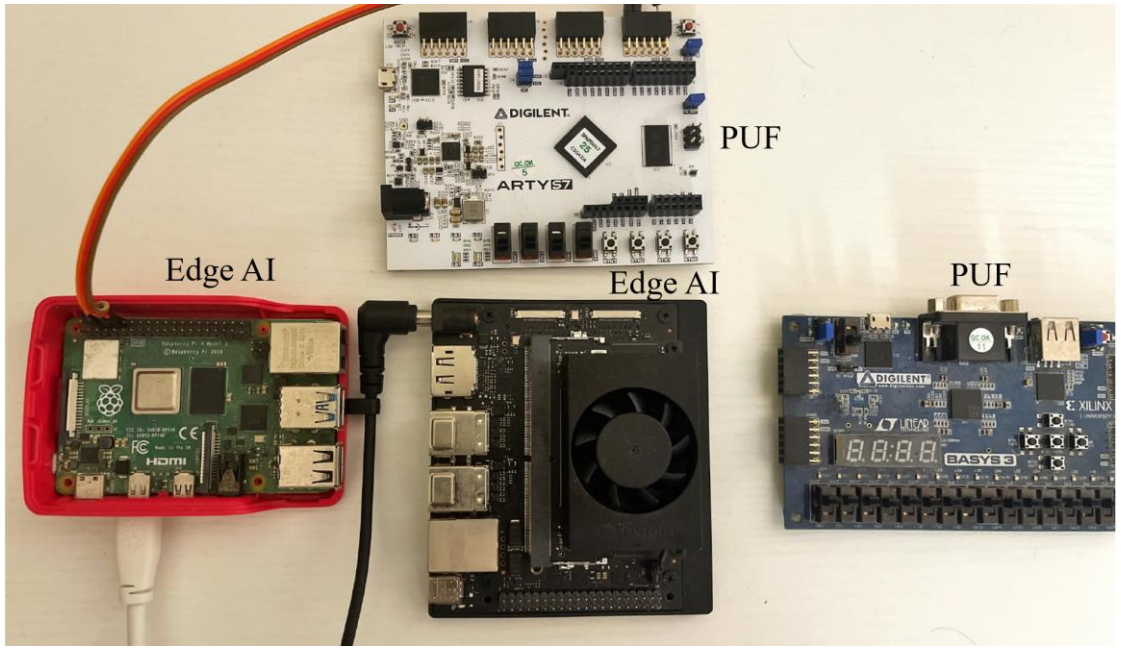
Final PUF Keys





Performance Analysis

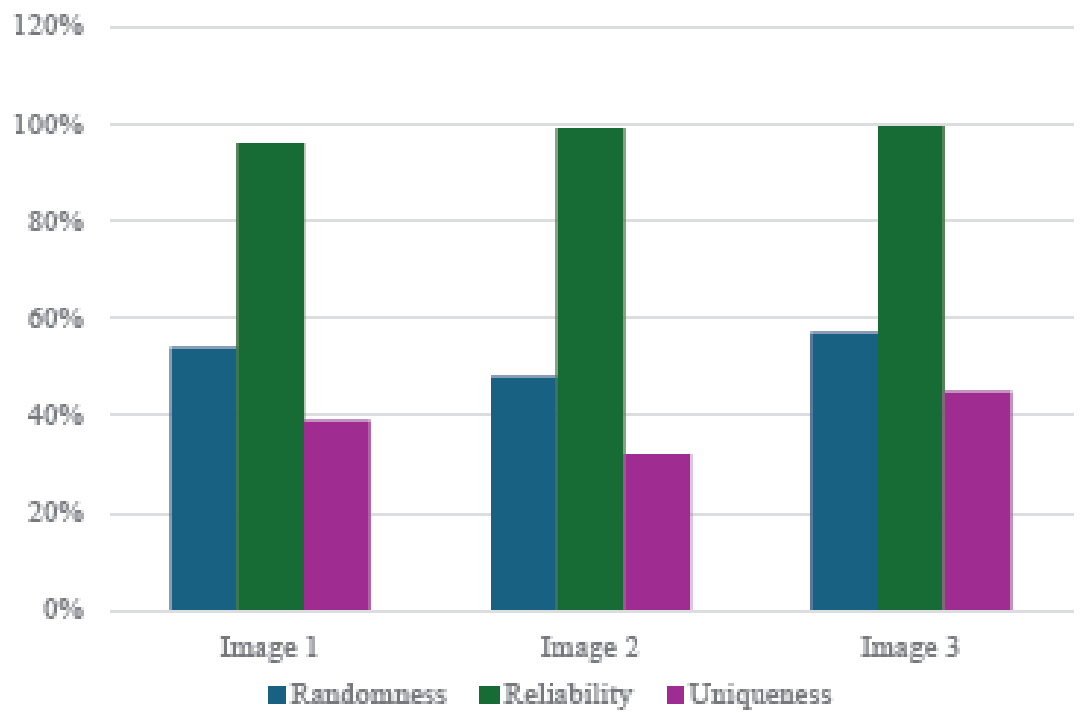
Prototype



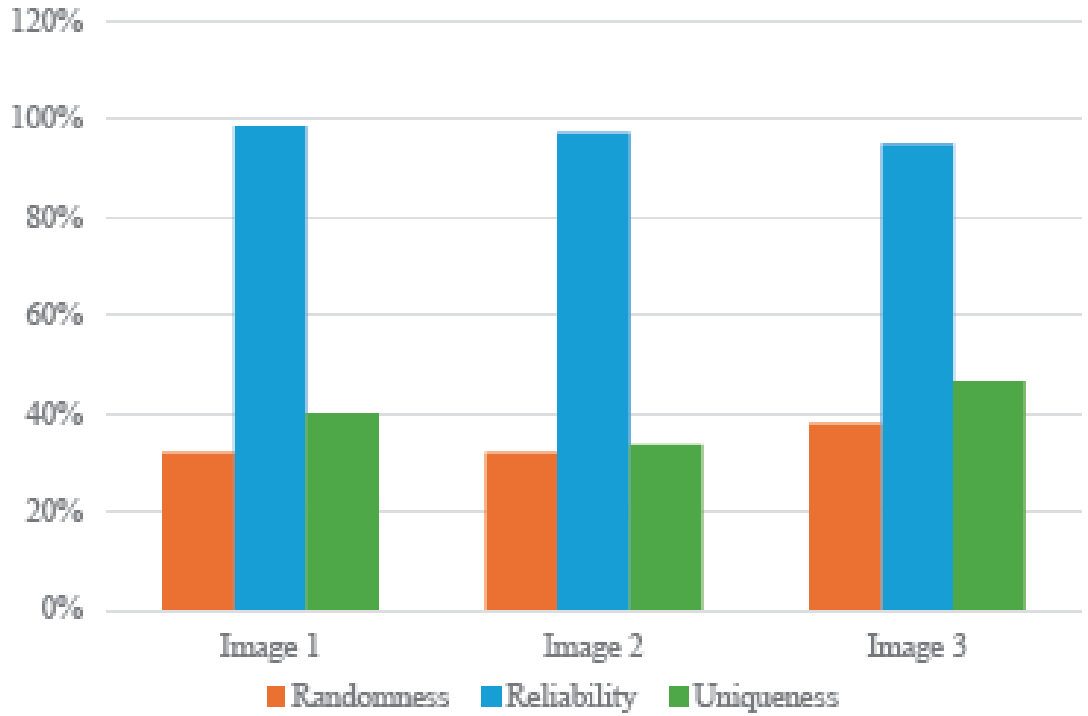
Computational Time Analysis

Content	Parameter	Results
Image 1	Facial detection Facial Landmark Prediction	60 ms 3 ms
Image 2	Facial detection Facial Landmark Prediction	57 ms 2 ms
Image 3	Facial detection Facial Landmark Prediction	56 ms 3 ms
All images	Attestation Time	300 ms

Image Attestation Metrics



(a) Artix-7 FPGA



(b) Spartan-7 FPGA

Conclusion and Future Research

- This research work presented and validated a state-of-art Deepfake mitigation technique that utilizes the potential of PUF for secure facial feature mapping and attestation.
- The proposed work experimentally validated the PUF-based facial feature attestation process for an image. This work can effectively counter Deepfake particularly facial attribute manipulation technique.
- The metrics evaluation results and computational time and power analysis on various hardware clearly demonstrates the potential of the proposed PUFshield.
- As a direction for future research, countering other techniques of visual Deepfakes such as face swapping, lip syncing in video and audio Deepfakes using PUF can be potential areas for PUF-based Deepfake mitigation.

Thank You!