UNIVERSITÀ
POLITECNICA
DELLE MARCHE

Faculty of Engineering
Master's Degree in Biomedical Engineering

Implementation of impersonation attacks against machine learning-based biometric authentication systems

Candidate:
Mohammed Najie AL Dreay

Advisor:
Prof. Marco Baldi

Academic Year 2022-2023

UNIVERSITÀ
POLITECNICA
DELLE MARCHE

Faculty of Engineering
Master's Degree in Biomedical Engineering

Implementation of impersonation attacks against machine learning-based biometric authentication systems

Candidate:
Mohammed Najie AL Dreay

Advisor:
Prof. Marco Baldi

Academic Year 2022-2023

Università Politecnica delle Marche
Faculty of Engineering
Master's Degree in Biomedical Engineering
Via Brecce Bianche – 60131 Ancona (AN), Italy

Acknowledgments

I extend my deepest gratitude to those whose pivotal contributions led to the completion of this thesis:

Special thanks to Prof. Macro Baldi and Eng. Nibras Abo Alzahab for their unwavering support, guidance, and invaluable insights throughout the entire research journey.

To my family—my father, mother, Wasem, Rahma, and Gaith—I am profoundly thankful for your unwavering support, encouragement, and understanding during this academic odyssey.

Hala, my vivacious muse and comrade of my journey.

Mosab and Yazan, the pillars of support and kindness.

Wael and Aslan, friends from the days filled with joy.

To my friends in Syria, regardless of the miles, Mohammad Al-Alwan, Somar, and Mohammed Alhashem.

To friends abroad and success, Mohammed Omki, Muhammad Kalet, Ayham, and Alaa.

A salute to the Droubi Medical team—Mr. Danny, Mr. Khaled, Mr. Samer, and all colleagues.

Mohammad Mohammadi, my favorite roommate.

To countless others not mentioned explicitly but whose contributions, big and small, are deeply appreciated.

In gratitude,
Thank you, Grazie.

Ancona, December 2023

Mohammed Najie AL Dreay

Abstract

The research delves into the evolving field of iris recognition in biometric systems, emphasising the dangers of reconstruction attacks. It addresses the challenge of maintaining these systems' security and reliability in the face of sophisticated attack strategies.

The CASIA V1.0 dataset with 756 iris images was used in the study to build and test iris recognition models. Two primary models were created, one inspired by literature. Convolutional neural networks were used in these models, which were rigorously trained and validated. The project also investigated various reconstruction attack strategies, with a particular focus on reconstructing training data from released machine learning models using a reconstructor network.

The recognition models achieved high iris identification accuracy, with the first model achieving 96.43% and the second 92.86%. The reconstruction attack experiments, on the other hand, revealed significant differences in the biometric similarity of the reconstructed and actual iris images. To assess these differences, descriptive statistics and statistical analyses (including the Shapiro-Wilk test and paired t-tests) were used.

The study demonstrates that iris recognition systems can maintain a high level of security and data integrity even when subjected to sophisticated reconstruction attacks. The similarity between the reconstructed and actual iris images suggests that these systems are resistant to model inversion attacks, which increases trust in biometric security systems.

The study concludes that modern iris recognition models are not only highly accurate, but also have a high resistance to reconstruction attacks. This highlights the importance of these systems in secure biometric verification and identification processes, providing strong defence against potential security threats.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Iris recognition, a robust biometric identification approach, capitalizes on the steady and distinctive nature of the human iris. It guarantees accurate and safe user authentication in digital systems by utilizing both cutting-edge deep learning techniques and more conventional methods like Gabor filtering. Iris Recognition, a pioneer in biometric authentication, outperforms conventional credential-based methods and provides improved protection against identity theft and cyberattacks.

However, the advancement of deep learning reveals vulnerabilities known as "model stealing" is revealed. This vulnerability is driven by considerable data and computing demands. Despite safeguards, maintaining performance and addressing security risks need a careful balance. Face recognition is an area in which Artificial Neural Networks thrive, however the emergence of Deep Neural Networks in Machine-Learning-as-a-Service raises privacy issues due to the potential hazards of data exposure. As MLaaS becomes more ubiquitous, security worries increase and research faces new obstacles that highlight the need for privacy safeguards. Deep learning and other data-driven methods are examined in "Machine Learning for Image Reconstruction," a special issue that aims to bridge the gap between real and rebuilt images.

This research aims to evaluate the robustness of modern iris recognition systems against reconstruction attacks, emphasizing their significance in secure biometric applications. Through the utilization of convolutional neural networks and extensive experimentation, the study seeks to demonstrate the resilience of these models to sophisticated attack strategies, reinforcing their efficacy in biometric verification and identification processes.

Additionally, this research proposal aims to explore the application of state-of-the-art reconstruction algorithms on biometric data to enhance confidentiality, authentication, and data protection. By assessing the performance of various reconstruction algorithms on a large-scale biometric dataset, the study intends to uncover their potential benefits and limitations in terms of accuracy, efficiency, and privacy preservation. The findings will inform the integration of reconstruction algorithms into existing biometric systems, contributing to the development of more effective and secure applications in fields such as law enforcement, border control, and access control.

The rest of the thesis is structured as follows: The second chapter provides background information as well as a review of the literature on iris recognition. It covers the fundamentals of authentication, biometrics, and iris recognition. Moving on to Chapter 3, which provides an overview and a review of the literature on reconstructing attacks against neural network models. Following that, Chapter 4 describes the experimental work by describing the datasets, prepossessing, iris recognition models, and reconstruction attacks. Chapter 5 presents the experimental results, while Chapter 6 discusses those results and concludes the work by listing the thesis's main contribution.

Chapter 2

Iris Recognition and Authentication

2.1 Introduction

The iris recognition has emerged as a prominent biometric identification method, leveraging both machine and deep learning techniques. Unlike passwords and pins, which can be compromised, the uniqueness of the human iris remains consistent throughout one's life. As traditional security measures falter, the reliability and precision of iris-based systems offer a promising avenue for robust authentication. This section delves into the evolution and efficacy of Iris Recognition Systems using advanced computational methodologies [1]. The human iris provides a stable and unique biometric identifier, staying constant for a lifetime. With high randomness and distinctiveness, it surpasses other modalities like facial or fingerprint recognition. Near-infrared (NIR) sensing is preferred for capturing iris patterns, especially in melanin-rich eyes. John Daugman's pioneering Gabor filtering-based technique remains the gold standard for rapid and precise iris recognition [2].

Furthermore, iris recognition has become a forefront in biometric authentication, harnessing advancements in machine and deep learning. Traditional methods relied on hand-crafted features, but the introduction of models like AlexNet revolutionized the field. Deep learning offers an end-to-end approach, enabling precise extraction of the intricate patterns of the iris. This evolution ensures enhanced security and user identification in modern systems [3].

2.2 Biometric Authentication

The imperative to safeguard personal data and counter cyber threats necessitates robust user authentication in digital systems. Traditional methods involving possession or knowledge of credentials pose security risks, while biometric authentication, based on inherent physical traits, provides a secure and convenient alternative, ensuring aspects like universality, uniqueness, and circumvention resistance [4][5].

Biometric authentication, crucial for identification and non-repudiation in information security, addresses concerns like credit card fraud and identity theft. Unlike traditional methods, it establishes an unbreakable correspondence between individuals and their data, enhancing overall security[6]. Information security, encompassing

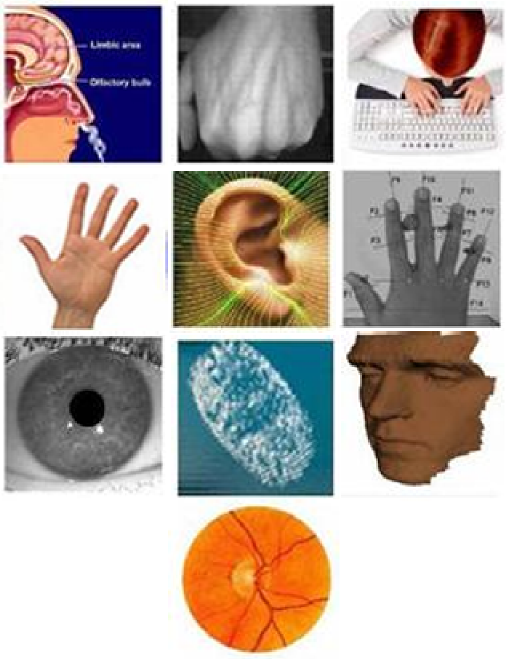


Figure 2.1: Some of the most widely used biometrics.

confidentiality, integrity, and availability, relies on various tools. Biometric systems address key aspects, supporting identification, authentication, and non-repudiation. With a rising focus on personal identification, especially due to concerns like credit card fraud, biometric authentication, in contrast to traditional methods, provides a more secure one-to-one correspondence between individuals and their data[7].

2.3 Biometrics

Advancements in technology have revolutionized identity verification, with automated systems implemented in diverse applications, spanning from shared computers to securing critical facilities like nuclear installations. Traditional methods, vulnerable to sharing or theft, have paved the way for biometric verification systems. Leveraging unique biological traits such as fingerprints, face, voice, iris, and keystroke patterns, these systems offer heightened security(shown in Figure 2.1). Despite their advantages, the inherent privacy risks associated with biometric data necessitate the development of privacy-preserving schemes, encompassing encryption-based, cancelable, multimodal, hybrid, and secure computation-based systems [8].

Biometrics, rooted in historical practices, has transitioned from law enforcement to widespread civilian use. Upholding criteria like universality, distinctiveness, permanence, and collectability, it prioritizes practical considerations such as accuracy, user acceptance, and resilience against fraudulent methods [9]. The last decade has witnessed the extensive adoption of biometric technologies for people authentication, driven by their resistance to loss or theft. However, challenges persist in addressing variations in biometric measurements and determining information content, particularly in entropy-based measures [10].

Biometric authentication, prevalent in diverse applications like immigration cards and facial recognition in casinos, involves automated methods of verifying living persons based on physiological or behavioral traits. Coined as "anthropometric authentication," this field distinguishes itself by focusing exclusively on living human subjects, excluding forensic techniques and non-living subjects [11].

2.4 Iris Biometric

The iris, the colored ring of tissue around the pupil, is controlled by muscles to regulate light entering the eye, exhibiting a unique and relatively constant pattern of furrows, ridges, and pigment spots, with its appearance believed to be randomly determined during fetal development (figure 2.2 shows an example) and varying between individuals and eyes[12]. The iris, recognized for its highly randomized appearance, serves as an accurate biometric due to its data-rich structure, genetic uniqueness, stability over time, and physical protection. The iris code computation relies on quality iris images focused on the customer's iris and properly positioned, employing specialized filter banks to extract information and ensure accuracy through circular bands conforming to iris and pupil boundaries[13]. Iris is a strong biometric in terms of recognition performance, both theoretically and empirically[14]. Over the past two decades, iris technology has evolved and integrated into various devices. The typical processing flow of an iris image-based biometric system involves an enrollment phase for database creation and a testing phase for real-time or pre-stored image recognition, segmented for detailed analysis[15].

2.5 Literature review

We conducted a systematic literature review to gain insights into the state-of-the-art iris recognition models. Initially, we detail the methodology employed for this review, followed by a presentation of key findings.

2.5.1 Method

The IEEE Xplore search engine was used to search the selected research, the keywords for this task were: "iris recognition," "machine learning," "deep learning," "neural networks," and "biometric authentication". the titles and abstracts of the search results have been reviewed. In this review, only studies from 2019 to 2023 were included.

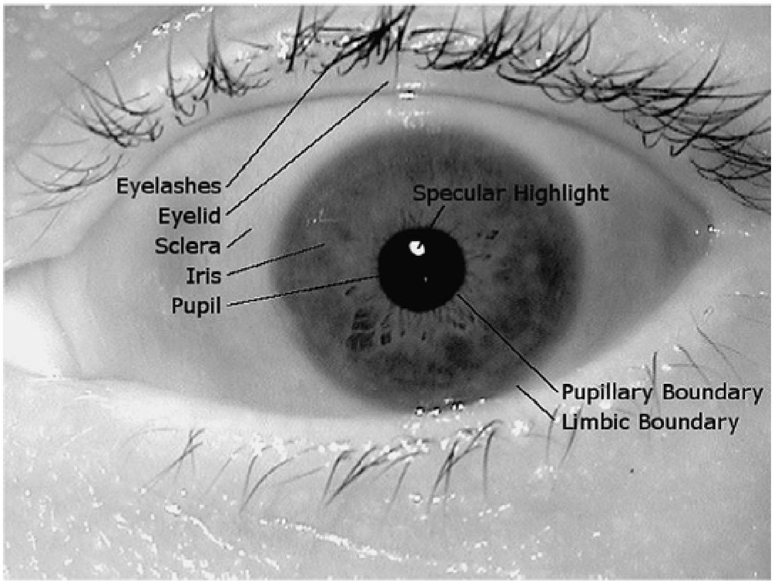


Figure 2.2: Image from the Iris Challenge Evaluation Dataset.

2.5.2 Results

The search above method resulted in the identification of six research articles published between 2019 and 2023. These articles are relevant to finding that ML model exists for biometric classification (or equivalently, authentication) resistant against attack to retrieve the training data. By employing this approach, recent and up-to-date research findings were included in the review.

2.5.2.1 Vijaykumar V et al, 2022

Nowadays, the need for reliable and efficient authentication systems is crucial. Artificial intelligence, specifically machine learning, has revolutionized biometric authentication systems. A powerful approach within machine learning is using deep convolutional neural networks (CNNs) for visual representation. Iris recognition, a biometric technique based on the unique iris patterns of individuals, offers a robust and effective solution for authentication.

This research [16] proposes a robust iris recognition strategy based on a CNN using a Kalman Filter. The suggested system surpasses existing iris recognition methods on public iris databases like Ubiris.v2, CASIA, and MMU V1.0, achieving an accuracy of over 99 percent in experimental findings.

Iris recognition technology is widely regarded as one of the most secure and reliable biometric identification methods. It leverages advanced techniques, such as iris scanning, to capture high-resolution iris images and compare them against stored patterns in a database. Convolutional Neural Networks (CNNs) have emerged as a prominent deep learning method for iris recognition. CNNs, shown in figure 2.3, employ convolutional layers to extract relevant features from input images and classify

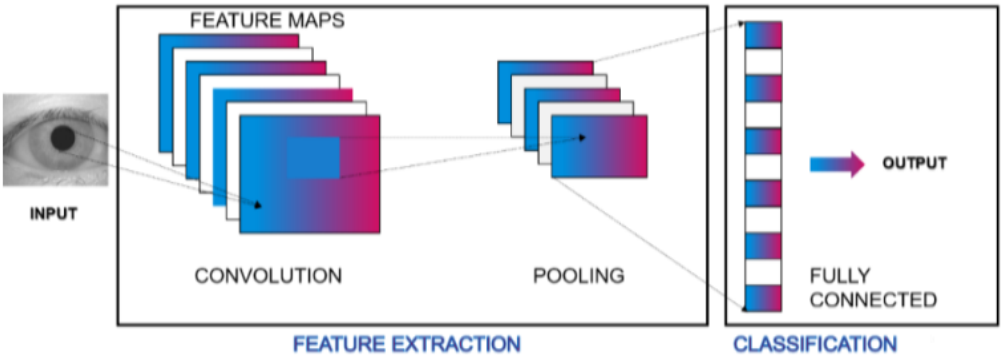


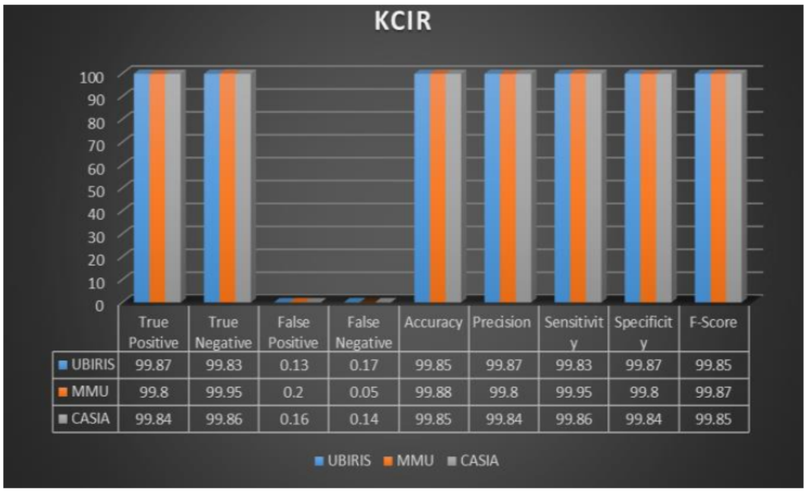
Figure 2.3: Convolutional Neural Network.used by Vijaykumar V et al.(2022)

them. The proposed KCIR model automates feature extraction and classification using CNNs, eliminating the need for domain-specific knowledge. Training the CNN involves employing the backpropagation algorithm and the Adam optimization approach to adjust weights and learning rates.

CNNs are classification algorithms that leverage weight sharing techniques and computational layers to process high-dimensional data. They consist of convolution, non-linearity, and pooling layers, followed by a fully connected layer and a logistic regression activation function. CNNs offer advantages such as reduced parameters, faster computation, and the ability to recognize boundaries, textures, and structures in various environments. The depth and complexity of CNN's hidden layers vary, with low-level layers handling basic aspects and high-level layers identifying complex features. Increasing the number of hidden layers enables recognition of distinct objects with similar properties.

Kalman filters, known for their adaptability to changing systems, are memory-light and highly efficient, making them ideal for real-time and embedded applications. The data used in this study was sourced from the UBIRIS v2, MMU, and CASIA v1 datasets. The training and testing were performed with a split of 70% and 30%, respectively. The results indicate that the Adam optimization technique, which iteratively adjusts network weights based on training data, effectively reduces the cost determined by cross-entropy. Cross entropy is employed to measure the distance between output probabilities and the ground truth values. The training was conducted over 150 epochs.

In conclusion, this paper proposes a robust iris-based biometric identification method that, as shown in figure 2.4, achieves accurate and precise individual recognition using deep learning. The proposed approach involves preprocessing the input image, performing improved segmentation to locate boundaries, normalizing the segmented images using Daugman's Rubber Sheet model, extracting effective features through a CNN architecture, and employing a Neural Network (NN) classifier to achieve over 99 percent accuracy. The experiments were conducted on the UBIRIS v2,



	True Positive	True Negative	False Positive	False Negative	Accuracy	Precision	Sensitivity	Specificity	F-Score
■ UBIRIS	99.87	99.83	0.13	0.17	99.85	99.87	99.83	99.87	99.85
■ MMU	99.8	99.95	0.2	0.05	99.88	99.8	99.95	99.8	99.87
■ CASIA	99.84	99.86	0.16	0.14	99.85	99.84	99.86	99.84	99.85

Figure 2.4: Evaluation metrics for all the datasets for the proposed model. Used by Vijaykumar V et al. (2020)

MMU, and CASIA datasets. Future work includes comparing the proposed method with other contemporary algorithms and analyzing time and space complexities.

2.5.2.2 ZHUANG Y et al, 2020

This article [17] focuses on the importance of designing an efficient user authentication system that can accurately detect personal identity. Iris recognition, a widely researched biometric identification technology, is gaining popularity due to increased awareness of personal privacy. Leveraging artificial intelligence, specifically convolutional neural networks (CNNs), presents an opportunity to enhance iris recognition and protect private data. CNNs are well-suited for image processing and pattern recognition, making them a practical algorithm. This study focuses on developing a highly precise and efficient iris recognition system based on CNNs. A dataset of iris samples from 20 individuals, including both eyes, is used to train the deep recognition system. The model initially shows signs of underfitting and limited convergence with insufficient training epochs. However, as the number of training epochs increases shown in figure 2.5, the model achieves a testing accuracy of 99%.
In summary, The proposed CNN-based iris recognition system, as shown in Figure 2.6, demonstrates remarkable accuracy in predicting the identities of up to 20 individuals. The system incorporates two convolutional layers, each accompanied by a corresponding pooling layer. The initial convolutional layer employs six filters, resulting in six distinct feature maps. The subsequent convolutional layer receives a 2x2 subsampled form of these six feature maps from the first pooling layer. Following convolution with twelve filters in the second layer, the outputs are further processed in the second pooling layer for an additional round of subsampling. Consequently, the network transforms an input image from a size of 28x28 into 12 feature maps, each with a size

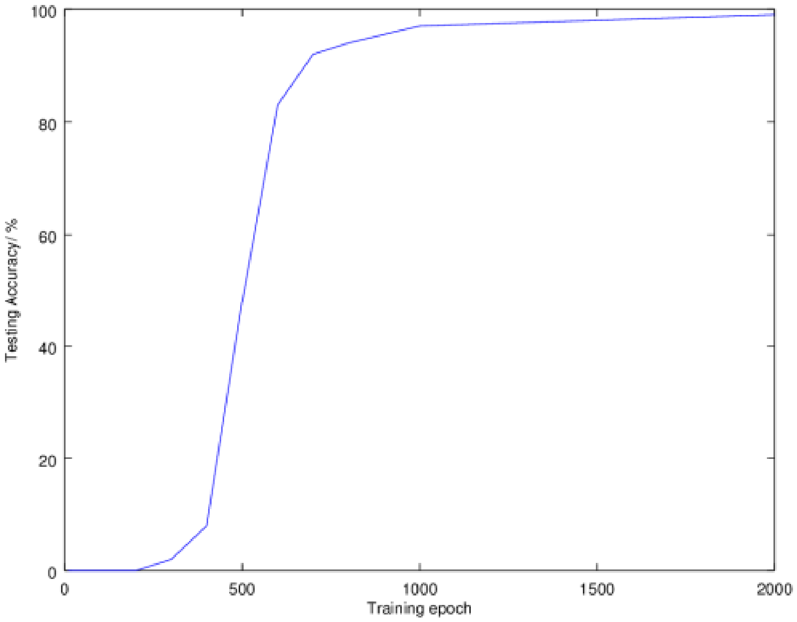


Figure 2.5: Plotting the testing accuracy against the number of training epoch, notice the accuracy started to plateau near 1000 epochs. used by ZHUANG Y et al.(2020)

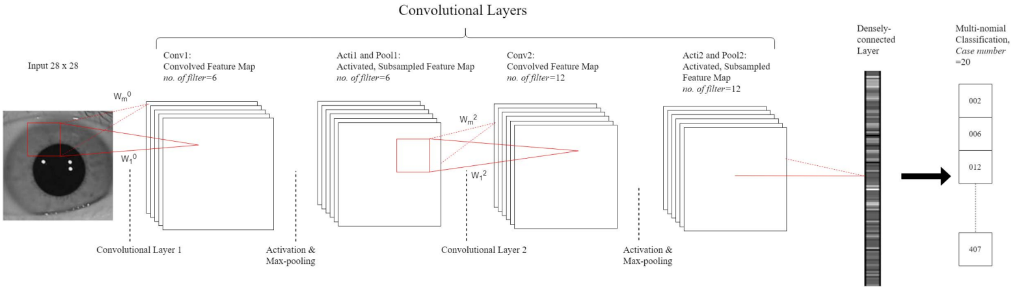


Figure 2.6: The architecture of the CNN-based iris recognition system.used by ZHUANG Y et al.(2020)

of 4x4. These maps are then fed into the fully connected layer, enabling the system to accurately determine the correct identification. However, there are some challenges to address. These include the relatively small number of iris pairs used for training, the computational complexity resulting in long training times, and the black-box nature of the model, which lacks transparency regarding its inner workings. To overcome these challenges, future research should focus on increasing the number of diverse iris samples while maintaining accuracy, improving computational efficiency through methods like hybridization in the network architecture, and reducing processing complexity.

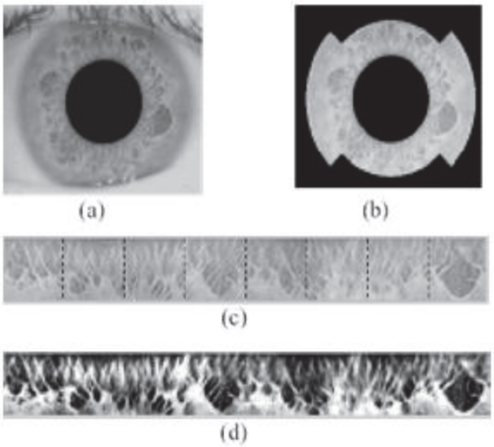


Figure 2.7: (a) Creative iris image; (b) After performing localization; (c) Eight unpacked texture images; (d) Eight unpacked texture images after enhancement and denoising. Used by Thakkar S et al. (2020)

2.5.2.3 Thakkar S et al, 2020

This paper [18] introduces a novel method for iris recognition by leveraging Gabor filters and a supervised neural network. The authors compare their approach against related works in the field and achieve an impressive Correct Recognition Rate (CRR) of 99.9998 % on the CASIA iris databases, surpassing the performance of previous methods. The study emphasizes the importance of biometric identification in the context of advancing technology and the need for robust security systems. Notably, the paper highlights the efficacy of features extracted from the neural network and Gabor filters in iris recognition tasks shown in figure 2.7. To facilitate testing and performance analysis, the authors construct a dedicated image database, although specific details regarding the dataset, such as the number of individuals or images, are not provided.

This study presents an algorithmic framework that combines Gabor filters and deep learning for iris recognition, resulting in improved accuracy compared to existing approaches. The proposed algorithm effectively extracts comprehensive and distinct iris features, demonstrating its potential for reliable and accurate biometric authentication. The achieved CRR of 99.9998% on the CASIA iris databases indicates the high performance and superiority of the proposed method. The results highlight the value of incorporating Gabor filters and neural networks in iris recognition systems, further contributing to the advancement of biometric identification technologies.

2.5.2.4 Vizoni MV et al, 2020

This paper [19] presents a novel method for person authentication based on ocular deep features extracted using a Convolutional Neural Network (CNN). The motivation for this research stems from the limitations of biometric systems that rely on the whole face, which can result in poor performance in certain cases. By focusing specifically on the ocular region, the proposed method aims to improve the accuracy and robustness of biometric identification.

The method consists of two main stages as shown in figure 2.8: feature extraction and identification (authentication) of individuals. In the feature extraction stage, ocular images are processed through a pre-trained CNN network, which applies successive convolutions and samplings to analyze patterns. Importantly, the classification stage of the CNN is not utilized, and the feature vector is obtained before classification. Instead of using the deep features directly, the difference between the probe and gallery deep feature vectors is employed. This pairwise strategy involves comparing feature vectors obtained from the same individual to generate genuine comparison patterns, while impostor comparison patterns are generated by comparing feature vectors from different individuals. By converting the multi-class authentication problem into a binary classification problem, the classifier can determine whether a pair of ocular images is genuine or impostor.

In the identification stage, the extracted feature vectors are used for biometric authentication. The objective is to verify if the biometric characteristics obtained from the probe image match those stored in the database (gallery image). To enhance the robustness and performance of the authentication system, a pairwise approach is adopted, which models the relationships between feature vectors. The difference feature vector, obtained by subtracting the probe and gallery feature vectors, is presented to a Support Vector Machine (SVM) classifier. The SVM classifier is trained to determine if the comparison is genuine (same individual) or impostor (different individuals). Probability values provided by the SVM are used as scores for calculating Receiver Operating Characteristics (ROC) curves and Equal Error Rates (EERs), allowing for the evaluation of the system's performance.

The experimental protocol involved assessing five different pre-trained CNN architectures (Resnet50, VGG16, VGG19, Xception, and VGG-Face) for feature extraction. Feature vectors were extracted from the fully connected layer before classification, resulting in feature vectors of either 1000 or 4096 dimensions. Genuine and impostor difference feature vectors were generated by subtracting feature vectors from the same individual or different individuals, respectively. An SVM classifier with an RBF kernel was trained using the genuine and impostor difference feature vectors. During testing, the difference feature vectors obtained from the test set were submitted to the trained SVM classifier.

The results demonstrate the superiority of the proposed method compared to traditional distance functions applied directly to feature vectors. The ROC curves clearly show in figure 2.9 that the proposed method outperforms the direct application of cosine and Euclidean distances. Among the five CNN architectures used for feature extraction, VGG-Face achieved the best performance, with an EER of 3.18%. This indicates the effectiveness of the proposed method in achieving accurate and reliable biometric identification.

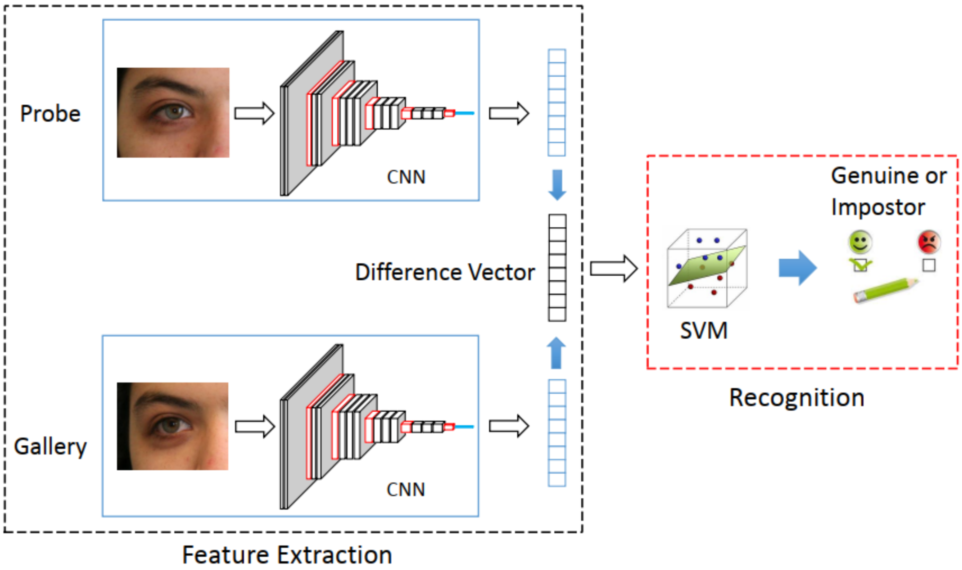


Figure 2.8: Diagram of the proposed method. used by Vizoni MV et al.(2020)

In conclusion, the proposed method for ocular region recognition-based authentication demonstrates its effectiveness in improving biometric authentication systems. By utilizing deep features and adopting a pairwise strategy, the method achieves state-of-the-art performance. The results suggest that ocular characteristics have great potential for biometric authentication, either as standalone features or in combination with other biometric modalities. This work contributes to the advancement of biometric authentication systems and highlights the importance of exploring ocular biometrics further.

2.5.2.5 Sudhakar T et al, 2019

This study [20] introduces a novel cancelable biometric system with the objective of enhancing the security and privacy of biometric systems. The system utilizes deep learning techniques to extract iris features and subsequently converts them into cancelable biometric templates through the application of random projection. Optimal biometric authentication is achieved by employing a support vector machine (SVM) after conducting a comparative analysis of alternative classifiers. The proposed system demonstrates improved template security and enhanced identification accuracy. Figure 2.10 illustrates the overall architecture of the cancelable biometric system, which consists of four main phases: feature extraction, transformation, fusion, and testing. During the feature extraction phase, features of both the right and left irises are extracted using a convolutional neural network (CNN). This process involves five steps, including rescaling, pixelating, normalization, CNN training and testing, and feature extraction based on CNN. It is worth noting that feature extraction is identical for both the training and testing phases. In the transformation phase,

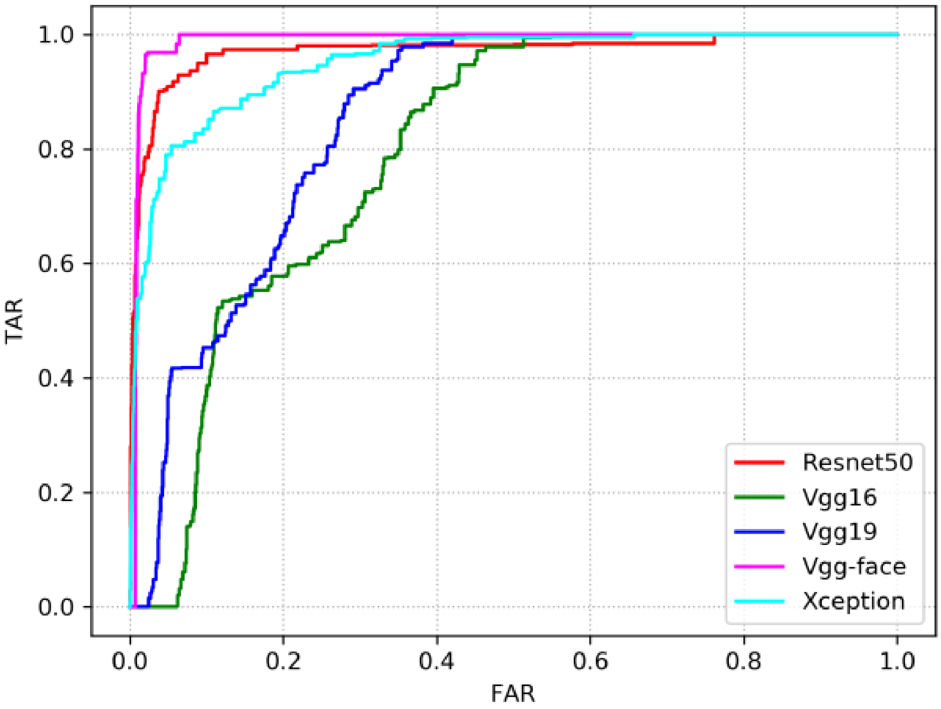


Figure 2.9: ROC curves obtained by the five CNNs assessed in this work as a means of obtaining deep ocular features. As one can observe VGG-Face obtained the best result. used by Vizoni MV et al.(2020)

random projection is applied to the feature matrix to generate the cancelable template. The iris was chosen as the biometric trait due to its stability over time, difficulty in replication, clear differentiability, and the improved reliability achieved by utilizing both irises (multi-instance). The proposed method offers computational efficiency and requires less memory, as each template has a dimension of only (1 x 256), thereby eliminating the need for dimension reduction techniques such as Principal Component Analysis (PCA). This efficiency is made possible by utilizing maxpool, which effectively reduces the iris dimensions from (1 x 16384) to (1 x 256). Each feature set is then transformed into a new feature space using a user key and orthogonal projection. The cancelable matrices of the user's right and left irises are fused through multiplication, resulting in the final template during the fusion phase. This approach provides a high level of resistance to inversion, as imposters are unable to recreate the original biometric using the key or template, or both. In situations involving suspicious activity, the previous template can be revoked, and a new template is issued to the user based on a new key. During the testing phase, the user must present scans of both irises along with their user key for access verification. The key serves as a code to a random matrix used during training. SVM is employed to classify test data into specific classes, thereby verifying the user for access authorization. The effectiveness of the proposed methodology is extensively validated using two multi-instance iris databases.

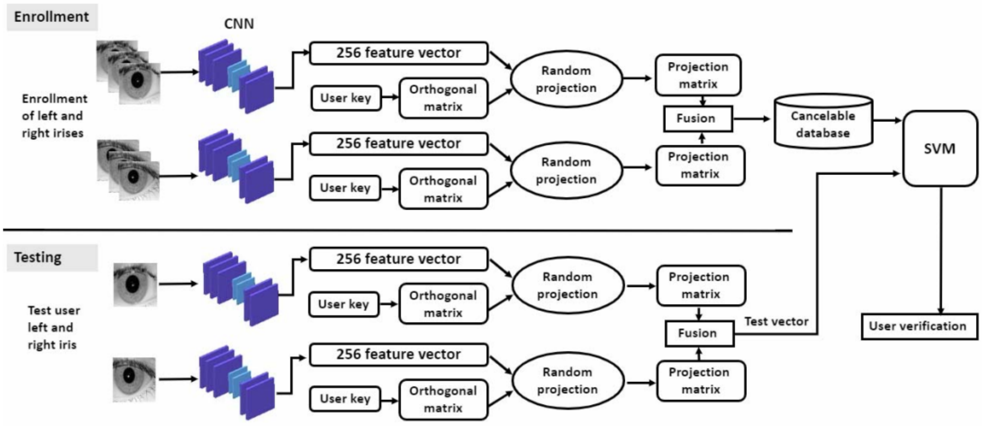


Figure 2.10: Overall architecture of the proposed cancelable biometric system. used by Sudhakar T et al.(2019)

The primary aim of utilizing random projection (RP) is to preserve the Euclidean distance and statistical properties before and after the projection. The proposed technique involves three steps, as depicted in Figure 2.11: a) The generation of an orthogonal matrix based on a user key, which remains consistent for both the left and right irises of an individual. b) The feature matrix is multiplied by the orthogonal matrix, resulting in a projection matrix. c) The projection matrices of the left and right irises are multiplied together to obtain the cancelable template.

These steps collectively ensure that the statistical characteristics and Euclidean distance are maintained throughout the process, and the resulting cancelable template retains the essential features for biometric recognition.

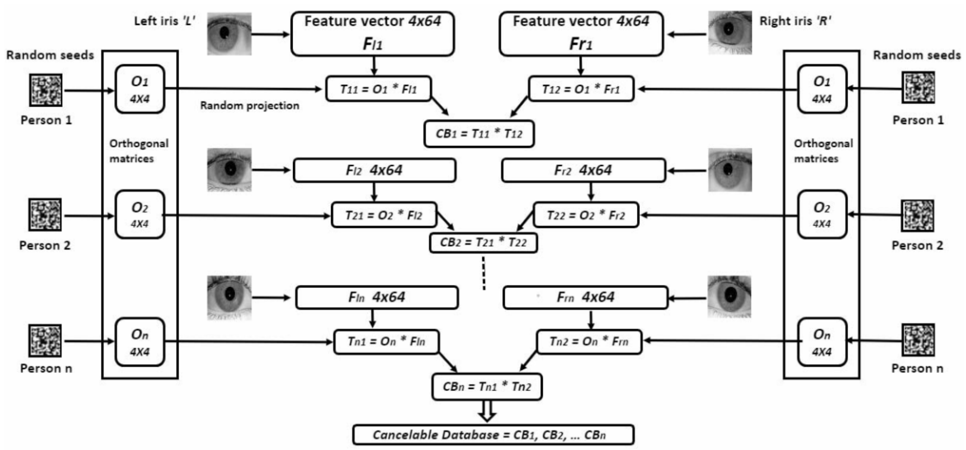


Figure 2.11: Generation of cancelable template using proposed random projection method. used by Sudhakar T et al.(2019)

In conclusion, the study explores a multi-instance cancelable biometric system that integrates deep learning, random projection, and machine learning. The system exhibits desirable properties such as non-invertibility, cancelability, differentiability, and computational efficiency. The implementation of a powerful deep neural network architecture contributes to high accuracy results. Additionally, the use of random projection significantly enhances the recognition rate and provides biometric transformation. Comparative analysis of various machine learning classifiers highlights optimal performance for user verification. The study prioritizes privacy preservation, improved security, and cancelability, which are key objectives in cancelable biometrics.

2.5.2.6 Boyd A et al, 2019

In this study [21], the authors investigate the effectiveness of deep learning-based feature extraction for iris recognition. The primary research question addressed is whether it is more beneficial to train models from scratch on a large iris image dataset or to fine-tune existing models to adapt them to the iris recognition domain. The ResNet-50 architecture is selected as the base model for experimentation.

The ResNet-50 architecture is a deep convolutional neural network as shown in figure 2.12 consisting of 53 convolutional layers. It is a fully convolutional model, meaning it is independent of input image dimensions and can accommodate larger input sizes. To explore the different training strategies, the authors utilize five sets of weights. Three sets are trained specifically for iris images, while the remaining two are obtained from the ImageNet and VGGFace2 datasets. These weights serve as benchmarks for comparison. Different weight initialization methods are employed for the trained networks.

To handle the large number of iris classes, the final classification layer of ResNet-50 is substituted, and a global average pooling layer is introduced. This modification enables the model to generate compact feature vectors suitable for iris recognition.

Features are extracted from each convolutional layer of the ResNet-50 architecture. In this study, features are obtained from all 53 individual convolutional layers, resulting in feature vectors with sizes ranging from 16,384 to 524,288. To ensure consistency and avoid dominance by larger scaled features, these feature vectors undergo Min-Max scaling. The aim is to generalize the extracted features for iris recognition, as the networks are not directly trained for classification on the specific CASIA-Iris-Thousand database used in this study.

Given the high dimensionality of the feature vectors, dimensionality reduction is applied before classification. Principal Component Analysis (PCA) is employed on each layer, projecting the features onto a new subspace with 2000 dimensions. This step aims to retain the most important features while avoiding overfitting. The feature vector size is further reduced by selecting the number of features that capture 90% of the feature variance. The "randomized" Singular Value Decomposition (SVD)

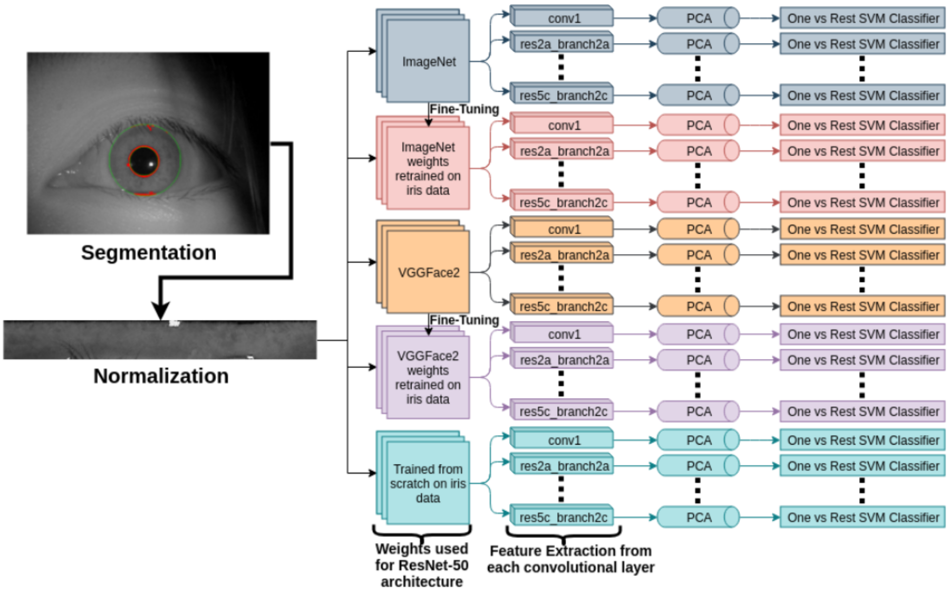

Segmentation
Normalization
ImageNet
Fine-Tuning
ImageNet weights retrained on iris data
VGGFace2
Fine-Tuning
VGGFace2 weights retrained on iris data
Trained from scratch on iris data
conv1
res2a_branch2a
res5c_branch2c
PCA
One vs Rest SVM Classifier
Weights used for ResNet-50 architecture
Feature Extraction from each convolutional layer

Figure 2.12: The figure illustrates the experimental setup, where iris images are segmented to create training and testing datasets. ResNet-50 models trained on ImageNet, VGGFace, fine-tuned versions, and a network trained from scratch are used to generate feature vectors for classification on the CASIA-Iris-Thousand dataset. used by Boyd A et al.(2019)

solver is utilized for efficient computation of PCA.

For classification, a one-versus-rest Support Vector Machine (SVM) with a linear kernel is employed. The classification training set consists of 70% of the CASIA-Iris-Thousand database, while the remaining 30% forms the testing subset. Stratified splitting ensures a balanced distribution of samples across classes in both sets. Each layer is associated with its own one-versus-rest classifier, and accuracy is evaluated based on correct classifications in the test set.

The evaluation of the proposed approach is presented in Figure 2.13, which illustrates the classification accuracy for each convolutional layer. The fine-tuned networks consistently outperform other networks, with similar performance observed between the fine-tuned ImageNet and VGGFace2 weights. The fine-tuned models achieve high classification accuracy, reaching up to 99%, demonstrating their effectiveness as feature extractors for iris recognition.

Comparison with a previous study reveals that four out of the five networks in this work surpass the highest recorded recognition rate achieved using the DenseNet architecture. The fine-tuned ImageNet network achieves the highest recognition rate of 99.38%. This highlights the effectiveness of fine-tuning pre-trained networks for iris recognition and indicates that deeper networks, like ResNet-50, offer improved performance due to the increased number of layers available for feature extraction.

In conclusion, the study provides insights into the optimal training strategy for iris recognition using deep learning-based feature extraction. The findings demonstrate

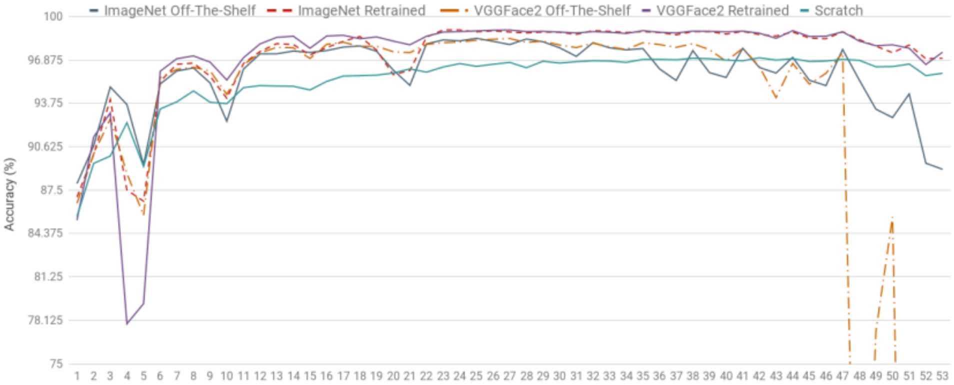


Figure 2.13: This plot shows the classification accuracy for each convolutional layer of the five networks tested on the CASIAIris-Thousand dataset. The x-axis is the convolutional layer number. Out of frame: results of VGGFace2 off-the-shelf for layers 48, 51, 52 and 53 which were 47.4%, 25.75%, 39.87%, and 53.81% respectively. used by Boyd A et al.(2019)

the superiority of fine-tuning off-the-shelf weights compared to both off-the-shelf models and training from scratch. The proposed methodology, utilizing the ResNet-50 architecture and fine-tuning techniques, proves effective in achieving high accuracy for iris recognition. The availability of fine-tuned ResNet-50 models trained on a large iris image dataset further supports reproducibility and provides a valuable resource for future research in this field.

2.5.3 Comparison tables

Table 2.1 outlines the study characteristics within the review, encompassing details such as the study year, dataset, pre-processing steps, feature extraction methods, data split approaches, optimization functions, loss functions, recognition performance metrics, activation functions, and the number of epochs and batches used.

Table 2.2 presents a comprehensive comparison of performance metrics across the included studies, offering insights into the variations and outcomes observed in different research findings.

Table 2.1: Study characteristics of the articles included in the review

Study, yr	Dataset	Pre-processing	Feature Extraction	Data split	Optimization function	Loss function	Recognition	Performance Metrics	Activation Function	Epoch-Batch
Vijaykumar V et al, 2022	three datasets: The UBIRIS v2 database contains 5799 photos from 350 people, MMU contains 460 pictures from 46 subjects, and CASIA V1.0. contains 756 pictures from 108 subjects.	improved segmentation technique is applied to locate the inner and outer boundaries of the iris in the input image. Next, the segmented iris images are normalized using Daugman's Rubber Sheet model. This normalization technique transforms the iris images into a standardized form.	utilized a Convolutional Neural Network (CNN) architecture to extract effective features from preprocessed iris images	Train (70%), test (30%)	Adam optimizes network weights, reducing cost based on CrossEntropy.	/	using a Neural Network (NN) classifier	Accuracy, Precision, Sensitivity, Specificity, F-score, false positives/negatives..etc	/	Epoch= 150
ZHUANG Y et al, 2020	20 individuals, CASIA-irisV4, 800 images	segmentation of the images in the red color channel to contain only the eyeball	two layers of CNN layer, each with a pooling layer	separated randomly into training and testing set at a ratio of 7:1	Gradient descent	/	using CNN	prediction accuracy and misidentified iris	/	Epoch= 2000, minibatch= 50
Thakkar S et al, 2020	CASIAV4	localization, normalization, enhancement, and noise removing.	dividing the iris image into 8 sub-images of 64x64 pixels, applying 20 Gabor filters on each sub-image to get 160 filtered sub-images, finding the average absolute deviation (AAD) from each filtered sub-image as features, and making a 1D array of 160x1 as a feature vector from AAD values.	Train (80%), test (20%)	/	/	involves using an array of AAD values as input to a trained neural network for determining the authenticity of the user and identifying the person's index if genuine. The neural network is a supervised model trained with various iris feature vectors as inputs.	Correct Recognition Rate (CRR)	/	/
Vizoni MV et al, 2020	261 individuals, UBIPr dataset, 10950 images	The paper does not mention any specific preprocessing method used in the experiments.	by a pre-trained CNN network and images go through successive convolutions and samplings. The output of each convolutional layer is input to the next layer. The feature vector is obtained before the network classifies the samples, and this vector varies its size according to the architecture of the network used.	/	/	/	involves pairwise comparison of feature vectors obtained from the probe and gallery images, with the SVM classifier used to determine the genuineness or impostor status of the comparison based on the calculated difference feature vector.	Receiver Operating Characteristic (ROC) and the Equal Error Rate (EER).	/	/
Sudhakar T et al, 2019	two datasets: The IITD dataset consists of iris images of 176 males and 48 females and Multi-media University (MMU) dataset consists of iris images of 45 individuals.	comprises five steps: a) Rescaling, b) Pixelating, c) Normalization, d) CNN training and testing, and e) Feature extraction based on CNN.	extraction of features from the dense layer of CNN and then Random Projection (RP) is applied to generate a cancelable template	Data is split into five equal segments to train and test the CNN.	RMSprop optimizer and annealer speed up convergence to global minimum.	Categorical cross-entropy	The linear Support Vector Machine (SVM) is implemented and evaluated using 5-fold cross-validation.	Accuracy, loss, Equal Error Rate (EER), Precision, Recall, and F1-scores	use the Relu activation function and employ Flatten, Dense, and SoftMax layers.	Epoch: 20
Boyd A et al, 2019	two datasets: in-house iris data consists of 373,629 full iris images and the CASIA-Iris-Thousand database contains 20,000 images from 1000 subjects.	OSIRIS tool segments iris images, generating normalized images (64×512).	from the convolutional layers of the ResNet-50 architecture, producing feature vectors of varying sizes. Min-Max scaling is applied. To address the issue of large-scale feature vectors, dimensionality reduction is performed using PCA, projecting the features onto a 2000-dimensional subspace.	Train (70%), test (30%)	the ResNet-50 model was fine-tuned on the iris classification database.	the ResNet-50 model was fine-tuned on the iris classification database, it is a suitable loss function for multi-class classification.	a one-versus-rest SVM classifier with a linear kernel was trained using the extracted features from the ResNet-50 architecture.	Receiver Operating Characteristic (ROC) curves and the true positive rate at 0.1% false match rate and the peak recognition rate can be generated.	/	/

Table 2.2: Comparison between performance matrices results among included studies

Study, yr	**Accuracy (%)**	**TPR (%)**	**Loss (%)**	**Precision (%)**	**Sensitivity (%)**	**Specificity (%)**	**F-score (s)**	**FAR (%)**	**EER (%)**	**CRR (%)**	**Notes**
Vijaykumar V et al, 2022	99.85	/	/	99.87	99.83	99.87	99.85	/	/	/	UBIRIS dataset
	99.88	/	/	99.8	99.95	99.8	99.87	/	/	/	MMU dataset
	99.85	/	/	99.84	99.86	99.84	99.85	/	/	/	CASIA dataset
ZHUANG Y et al, 2020	99	/	/	/	/	/	/	/	/	/	/
Thakkar S et al, 2020	/	/	/	/	/	/	/	/	/	99.99	/
Vizoni MV et al, 2020	/	/	/	/	/	/	/	/	7.11	/	Resnet50
	/	/	/	/	/	/	/	/	30.23	/	VGG16
	/	/	/	/	/	/	/	/	23.97	/	VGG19
	/	/	/	/	/	/	/	/	12.9	/	Xception
	/	/	/	/	/	/	/	/	3.18	/	VGG-Face
Sudhakar T et al, 2019	96.6	/	10.36	98	/	/	/	95	12	/	Left Iris IITD
	98.03	/	7.3	98	/	/	/	95	12	/	Right Iris IITD
	96.57	/	13.25	96	/	/	/	90	15	/	Left Iris MMU
	95.66	/	13.24	96	/	/	/	90	15	/	Right Iris MMU
Boyd A et al, 2019	97.03	97.93	/	/	/	/	/	/	/	/	layer 42^1
	98.43	98.93	/	/	/	/	/	/	/	/	layer 25^2
	98.41	98.93	/	/	/	/	/	/	/	/	layer 27^3
	99.03	99.38	/	/	/	/	/	/	/	/	layer 23^4
	99.03	99.27	/	/	/	/	/	/	/	/	layer 27^5
(1) network trained from scratch, (2) off-the-shelf ImageNet weights, (3) off-the-shelf VGGFace2 weights,											
(4) network that used weights that were fine-tuned from ImageNet weights, (5) network that used weights that were fine-tuned from the VGGFace2 weights											

Chapter 3

Attacks based on Reconstruction techniques

3.1 Introduction

The landscape of deep learning models is evolving, marked by their insatiable appetite for substantial data and computational resources. This progress, however, unveils a vulnerability known as 'model stealing,' allowing adversaries to replicate these models without direct access, thereby jeopardizing their integrity [22]. Traditional defense mechanisms, while in place, often grapple with the delicate balance between preserving performance and fending off evolving security threats [22].

At the core of artificial intelligence, Artificial Neural Networks (ANNs) replicate the intricate structures of the human brain, demonstrating their prowess in applications like face recognition and AI gaming. Major industry players such as Google, Microsoft, and Amazon harness the power of Deep Neural Networks (DNNs) for large-scale data processing, ushering in the era of Machine-Learning-as-a-Service (MLaaS). This innovative paradigm brings unparalleled ease of integration but simultaneously raises a critical dilemma – the risk of exposing sensitive training data and proprietary model intricacies. Despite ongoing defensive measures, the domain of neural network (NN) privacy research is still nascent, underscoring the need for a comprehensive exploration and policy realignment to effectively address these evolving privacy concerns [23] [24].

As machine learning seamlessly embeds itself into various facets of our daily lives through Machine Learning-as-a-Service (MLaaS) platforms provided by industry behemoths like Amazon and Google, the spotlight intensifies on the imperative to address security implications. While the current research thrust predominantly focuses on enhancing the performance of training algorithms, an emerging need compels the community to grapple with the security challenges, especially regarding the privacy sensitivity of trained models. Real-world applications, acutely aware of the potential leakage of sensitive information, are increasingly adopting privacy measures. This has resulted in the surge of adopting oracle access or black-box access within MLaaS systems as pragmatic solutions, striking a balance between safeguarding privacy and ensuring usability [25]. This special issue delves into the realm of "Machine learning for image reconstruction," emphasizing the crucial role of data-driven approaches in image reconstruction. Bridging the gap between existing images and reconstructed internal structures, machine learning, especially deep learning, is rapidly becoming a prominent method in the field, showcasing its potential as a new frontier in image reconstruction [26].

3.2 Literature review

To have an overview of the state-of-the-art of reconstructing attacks in the field of biometric authentication, a systematic literature review was conducted. First, we explain the method followed for this literature review and then a summary of findings is presented.

3.2.1 Method

The IEEE xplore and Google scholar search engine was employed to identify research relevant to data reconstruction from neural networks. Our primary search terms were "Reconstructing Training Data", "Neural Networks", "Federated Learning", "Dataset Distillation", and "Dataset Reconstruction Attacks". Furthermore, we looked for papers that discussed the challenges and implications of adversaries using these techniques. We primarily centered on research that focused on traditional deep learning models, excluding those that delved into alternative architectures. Special attention was given to works that explored ensemble inversion, informed adversaries, and the boundaries of training data reconstruction.

3.2.2 Results

The search methodology described previously yielded 7 papers that were published in the years 2022 and 2023.

3.2.2.1 Loo N et al, (2023)

Loo N et al. (2023) [27] presented a significant enhancement to the dataset reconstruction attack framework, diving deeper than the foundational methods by Haim et al. (2022) [28]. The new model, which operates on neural networks trained under MSE loss, leverages the inherent properties of the Neural Tangent Kernel (NTK) to guarantee a consistent and robust reconstruction. This advanced technique is anchored on the equation:

$$\mathcal{L}_{Reconstruction} = \|\Delta\theta - \alpha^\top \nabla_\theta f_{\theta_0}(X_T)\|_2^2 \tag{3.1}$$

Through rigorous evaluations on well-established datasets such as MNIST and CIFAR-10, it was evident that wider networks boasted enhanced capability in reconstructing more expansive datasets. An intriguing aspect of this study is the

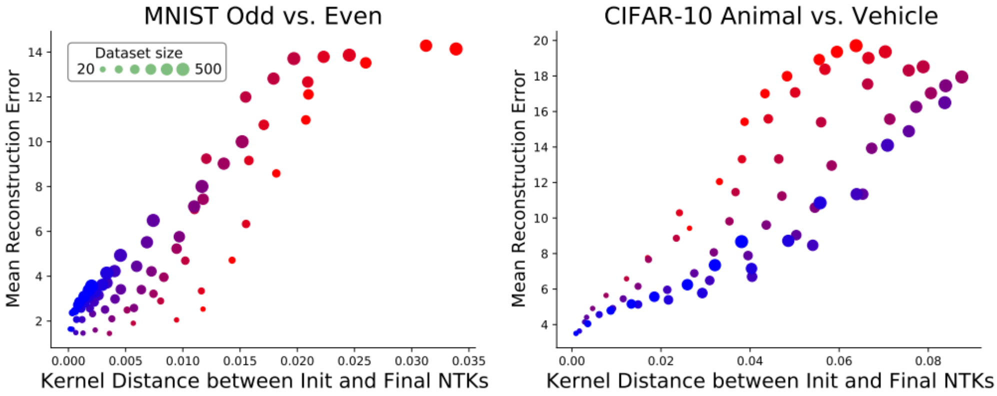


Figure 3.1: The mean reconstruction error, depicted through the average value on the reconstruction curve, displays a strong correlation with the evolution of the finite-width NTK over the course of training. Dot size signifies the dataset size, while the model width is distinguished by color variations. used by Loo N et al, (2023).

establishment of a novel connection between dataset reconstruction and distillation. They postulated that training on a distilled dataset could serve as a defense mechanism against the reconstruction of the original dataset, shielding it from potential attacks.

However, a noteworthy limitation of their approach was its focus on 2-layer networks, suggesting that more intricate architectures might remain vulnerable and require further exploration. Their findings, including the varying reconstruction quality across different network widths and the correlation between reconstruction quality and kernel distance, can be visually appreciated in Figure 3.1. This figure ideally would encapsulate the essence of their method, be it the attack mechanism, the dataset distillation method, or the demonstrated reconstruction curve.

3.2.2.2 Haim N et al, (2022)

Haim N et al. [28] delve into the capability of neural networks to reveal details about their training data. They explore the premise that the parameters of a trained classifier might contain sufficient information to reconstruct a substantial chunk of its training data.

Their Dataset Reconstruction Scheme capitalizes on the implicit biases of gradient-trained neural networks to recover training data. For the experiments, the primary architecture used is the Multi-layer Perceptron (MLP) comprising three fully-connected layers.

Their approach pivots on a uniquely designed reconstruction loss function $L_{reconstruct}$:

$$L_{\text{reconstruct}}(\{x_i\}_{i=1}^{m}, \{\lambda_i\}_{i=1}^{m}) = \alpha_1 L_{\text{stationary}} + \alpha_2 L_{\lambda} + \alpha_3 L_{\text{prior}} \tag{3.2}$$

Empirical results depict the method's efficacy in reconstructing training samples,

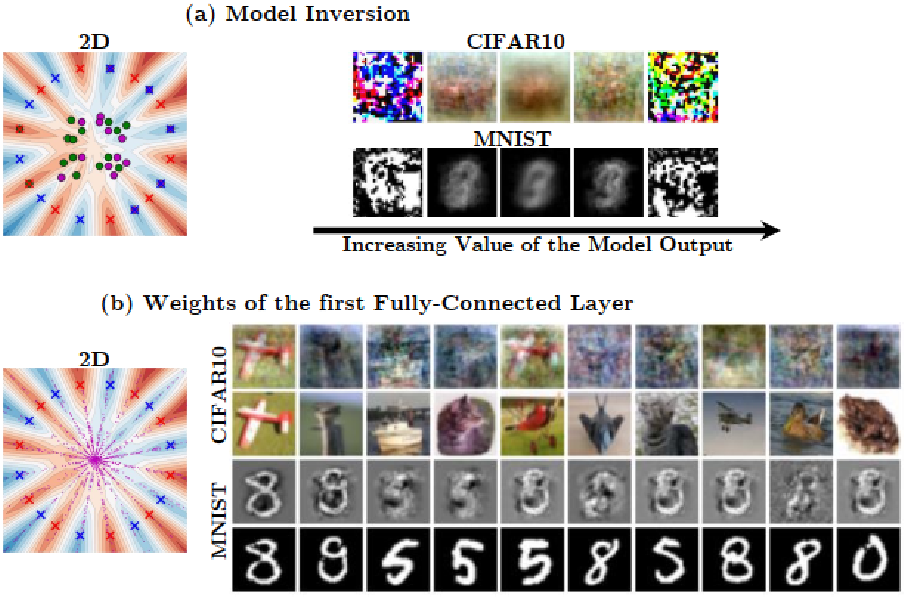

(a) Model Inversion
2D
CIFAR10
MNIST
Increasing Value of the Model Output
(b) Weights of the first Fully-Connected Layer
2D
CIFAR10
MNIST

Figure 3.2: Comparison to other reconstruction schemes. The top showcases Model Inversion results while the bottom highlights the neural network's first layer weights. The ordering of CIFAR and MNIST images is based on their output values. used by Haim N et al, (2022).

albeit with some noise. The technique appears most potent for data samples on the decision "margin" of the neural network.

A critical comparison of their method with alternatives like "Model Inversion" and "Weights Visualization" is elucidated in Figure 3.2. The top of the figure exhibits the performance of Model Inversion on 2D, CIFAR10, and MNIST datasets, while the bottom displays the first layer's weights. Notably, in CIFAR and MNIST, the images are ordered by their output values.

The paper by Haim N et al. [28] underscores the capacity within neural networks to retain and possibly reveal their training data inadvertently.

3.2.2.3 Balle B et al, (2022)

This paper [29] delves into the potential risks associated with the ability of machine learning models to inadvertently memorize and consequently reveal details about their training data. Using a rigorous threat model, the authors developed a sophisticated attack strategy that leverages an informed adversary. This adversary, familiar with some portions of the training data, can successfully recreate other unseen data points by just observing the parameters of the trained model. Particularly, they employed a reconstruction attack, termed "RecoNN" (Figure 3.3 provides an overview), that maps model parameters back to training images, which proved effective against standard classifiers used on datasets like MNIST and CIFAR-10.One of the most compelling visual proofs of this vulnerability was observed in Figure 3.4, where

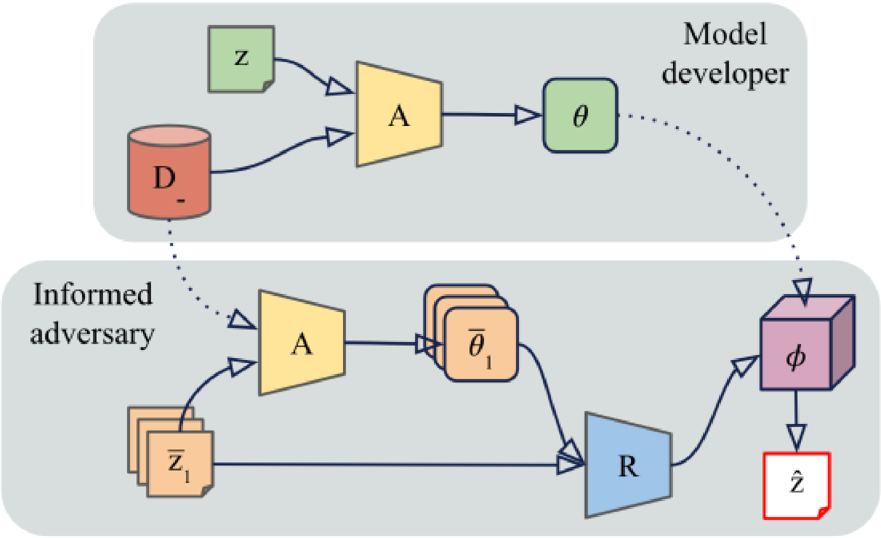


Figure 3.3: Overview of RecoNN-based attack. used by Balle B et al, (2022)

reconstructions of six random targets from the test set were vividly showcased. However, the research also demonstrated that differential privacy (DP), a widely recognized privacy preservation technique, can be employed to counteract such reconstruction attacks. The paper introduced a novel definition, "reconstruction robustness," linking it with Renyi Differential Privacy (RDP). Crucially, they establish that even when differential privacy is applied with relatively high values of its privacy parameter ϵ, it can still provide significant protection against reconstruction.

This Equation highlights the relationship between reconstruction robustness and differential privacy:

$$\mathbb{P}_{Z\sim\pi,\theta\sim M(D\_\cup\{Z\})}[\ell(Z,R(\theta))\leq\eta]\leq\gamma \tag{3.3}$$

Essentially, this equation describes the probability bounds for achieving a specific reconstruction error, serving as a bridge between reconstruction robustness and established differential privacy parameters. As a countermeasure, the paper suggests the application of differential privacy during model training, highlighting that even with larger values of the privacy parameter ϵ, effective mitigation against reconstruction attacks can be achieved. The results underscore the necessity for developers to be aware of these vulnerabilities and adopt practices like differential privacy to ensure data confidentiality and security.

3.2.2.4 Guan J et al, (2022)

In this paper[30], the authors delve into an innovative correlation-based fingerprinting framework meticulously designed to counteract and detect a myriad of model theft attacks. This novel approach is a timely response to the pitfalls and challenges

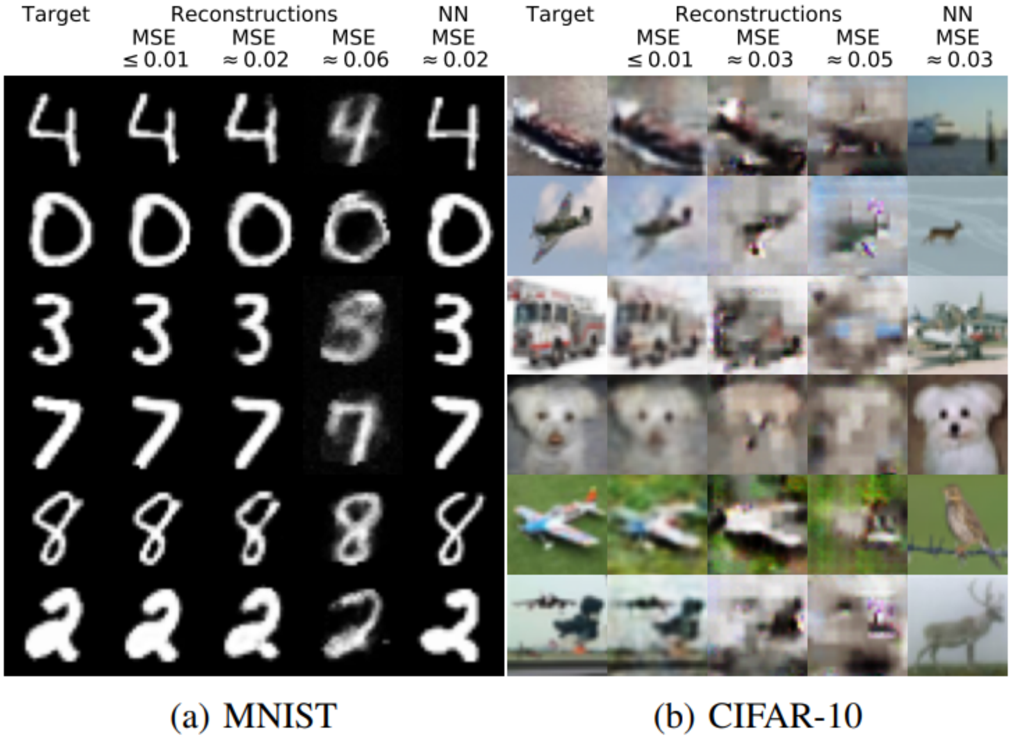


Figure 3.4: displays reconstructions of six random test set targets. It compares original targets, default attack reconstructions, and those by the NN oracle. used by Balle B et al, (2022).

encountered with traditional intellectual property (IP) protection methods. Current strategies are often marred by accuracy losses, considerable time overheads, and susceptibility to specific theft mechanisms. In contrast, the proposed framework, encompassing the SAC-w and SAC-m techniques, adopts a pioneering stance. It harnesses the unique properties of misclassified normal samples or CutMix Augmented samples. Central to this approach is the computation of the correlation difference between these samples. This pivotal step facilitates the identification of potentially stolen models, offering a detection mechanism that operates without intruding into the model's primary training process.

A critical equation that encapsulates the theft detection strategy is the probability-based model extraction, which can be represented as

$$L = \alpha.KL(f^T_{stolen}(x), f^T_{source}(x)) + (1-\alpha).CE(f_{stolen}(x), l_{source}) \tag{3.4}$$

The comprehensive experiments, as reflected in Figure 3.5, illuminate the preeminent performance of the SAC-m technique in the domain of fingerprinting, remarkably doing so while necessitating a substantially reduced sample pool. Additionally, the strategic incorporation of cosine similarity in the analytical process has further bolstered the model's theft detection capabilities.

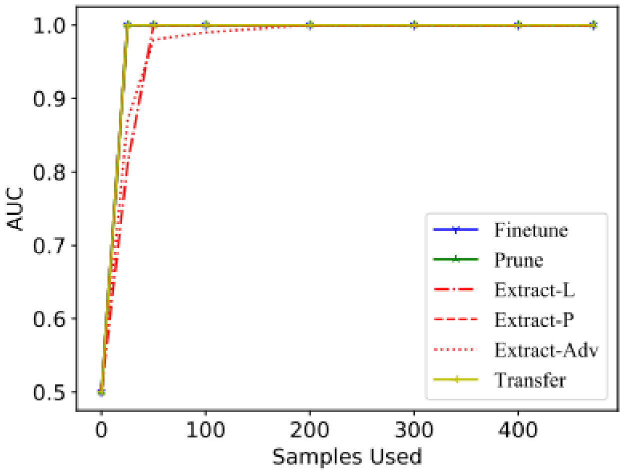


Figure 3.5: SAC-w's AUC change with different amounts of misclassified samples. used by Guan J et al, (2022).

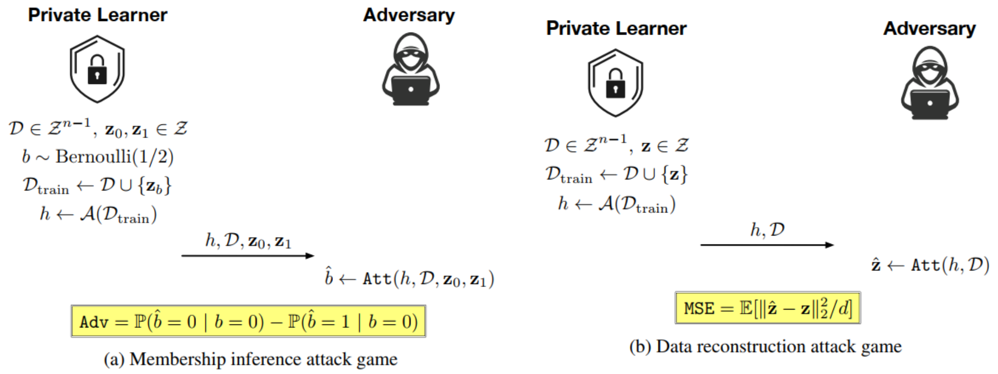
In summation, this research offers a solution to the dilemma of model theft that looms large in machine learning. By merging accuracy conservation with time efficiency, it promises a transformative impact on the domain. However, these technological advancements also beckon introspection on broader socio-economic fronts. The burgeoning acceleration of machine learning as a service, fueled by such breakthroughs, might inadvertently reshape the job landscape, hinting at potential socio-economic challenges ahead.

3.2.2.5 Guo C et al, (2022)

This paper [31] introduces a novel framework for quantifying data leakage in machine learning models, particularly emphasizing the Fisher information matrix's role in assessing privacy. The Fisher information matrix serves as a statistical measure, revealing how much information an observable variable unveils about an underlying parameter. The groundbreaking insight in this paper is the expression of the Fisher information matrix as an integral, linking the gradient of the log-likelihood function to the empirical distribution of training data. This newfound identity empowers the accurate measurement of potential information leakage from the model's parameters concerning sensitive training data.

To enhance privacy analysis, the paper introduces two pioneering privacy accounting methods: Renyi differential privacy (RDP) and Fisher information leakage (FIL). RDP offers a flexible trade-off between privacy and utility, expanding the horizons of differential privacy. FIL, inspired by statistical privacy principles, takes the spotlight by providing more robust mean squared error (MSE) lower bounds compared to existing techniques.

The experimental validation on diverse datasets underscores the effectiveness of


Private Learner
Adversary
$\mathcal{D} \in \mathcal{Z}^{n-1}$, $\mathbf{z}_0, \mathbf{z}_1 \in \mathcal{Z}$
$b \sim \text{Bernoulli}(1/2)$
$\mathcal{D}_{\text{train}} \leftarrow \mathcal{D} \cup \{\mathbf{z}_b\}$
$h \leftarrow \mathcal{A}(\mathcal{D}_{\text{train}})$
$h, \mathcal{D}, \mathbf{z}_0, \mathbf{z}_1$
$\hat{b} \leftarrow \texttt{Att}(h, \mathcal{D}, \mathbf{z}_0, \mathbf{z}_1)$
$\texttt{Adv} = \mathbb{P}(\hat{b} = 0 \mid b = 0) - \mathbb{P}(\hat{b} = 1 \mid b = 0)$
Private Learner
Adversary
$\mathcal{D} \in \mathcal{Z}^{n-1}$, $\mathbf{z} \in \mathcal{Z}$
$\mathcal{D}_{\text{train}} \leftarrow \mathcal{D} \cup \{\mathbf{z}\}$
$h \leftarrow \mathcal{A}(\mathcal{D}_{\text{train}})$
$h, \mathcal{D}$
$\hat{\mathbf{z}} \leftarrow \texttt{Att}(h, \mathcal{D})$
$\texttt{MSE} = \mathbb{E}[\|\hat{\mathbf{z}} - \mathbf{z}\|_2^2/d]$

(a) Membership inference attack game

(b) Data reconstruction attack game

Figure 3.6: compares membership inference attacks (MIAs) and data reconstruction attacks (DRAs) by framing them as games between private learners and adversaries, with MIA measured by advantage (higher is better) and DRA by mean squared error (lower is better).. used by Guo C et al, (2022).

these approaches. One compelling figure, Figure 1 in the paper, presents the privacy loss in differentially private logistic regression on the Adult dataset, using Fisher information leakage. The results affirm the superiority of this method in establishing stringent privacy lower bounds in comparison to conventional alternatives.

This paper offers a comprehensive framework for assessing data leakage in machine learning models, highlighting the importance of the Fisher information matrix. By providing practical tools to enhance privacy analysis, the paper advances the state of the art in securing sensitive information in machine learning applications.

3.2.2.6 Gong H et al, (2022)

This paper discusses the potential privacy risks associated with exchanging gradients and weights of models in Federated Learning (FL) systems. The authors provide an overview of the FL framework and highlight the importance of privacy-preserving techniques in FL. They then discuss the threat model of FL and the potential privacy risks associated with exchanging gradients and weights. The authors also review several existing attacks on FL systems, including Single-Sample Reconstruction Attack System (SSRAS), Deep Leakage from Gradients (DLG), and Recursive Gradient Attack on Privacy (R-GAP), figure3.7 illustrates the framework of the Improved R-GAP Algorithm.

To address these privacy risks, the authors propose several defense mechanisms, including gradient perturbation, differential privacy, and secure aggregation. They also discuss the limitations of these defense mechanisms and suggest future research directions.

This paper [32] provides a comprehensive review of the privacy risks and defense mechanisms in FL systems. The authors highlight the importance of privacy-preserving techniques in FL and propose several defense mechanisms to address

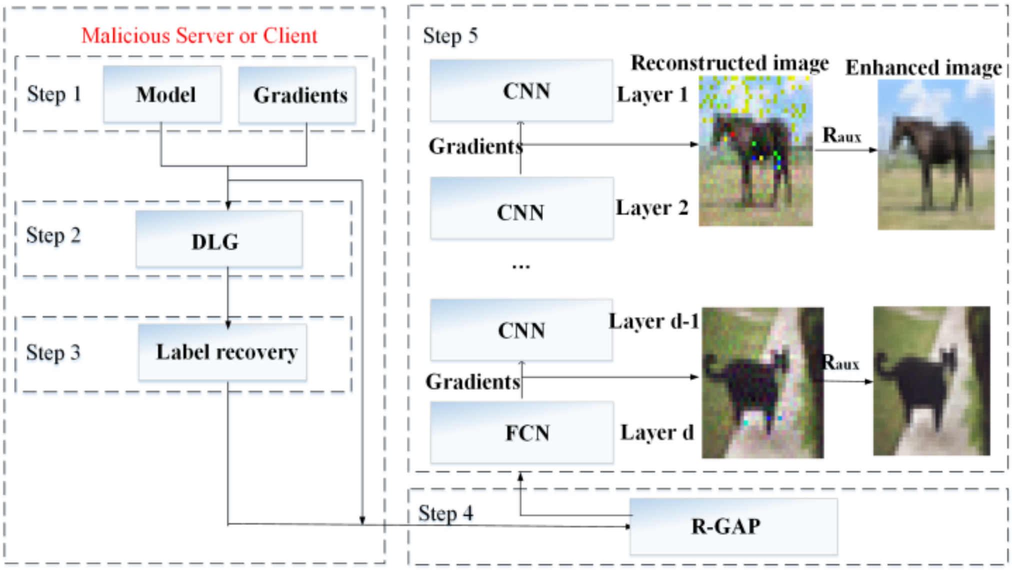


Figure 3.7: The framework of Improved R-GAP Alogrithm. used by Gong H et al, (2022).

the privacy risks associated with exchanging gradients and weights. The paper includes Figure 4, which illustrates the comparison of gradient inversion attacks on two datasets, and Figure 5, which presents ground truth and reconstructed images by the attack system and DLG.

The paper introduces the Single-Sample Reconstruction Attack System (SSRAS) and conducts experiments to assess its performance in the context of privacy attacks in federated learning. The study focuses on a classification task with CNN6 architecture, utilizing various activation functions and optimization techniques. Key metrics used for evaluation include the Rank Analysis Index (RA-I), Attack Success Rate (ASR), Attack iteration (Ai), Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity (SSIM).

In conclusion, this paper emphasizes the need for a deeper understanding of gradient leakage attacks and privacy in federated learning. The proposed SSRAS system and Improved R-GAP Algorithm provide valuable tools for privacy analysis and open the door to more secure and privacy-preserving intelligent systems. These findings offer important insights for researchers and practitioners in the field of federated learning and privacy preservation.

3.2.2.7 Wang Q et al,(2022)

The paper[33] presents an ensemble inversion technique that significantly improves model inversion tasks, with a focus on face classification. Utilizing multiple Models Under Attack (MUAs), the method reconstructs training images by guiding a generator with a combination of MUAs' predictions and a discriminator that en-

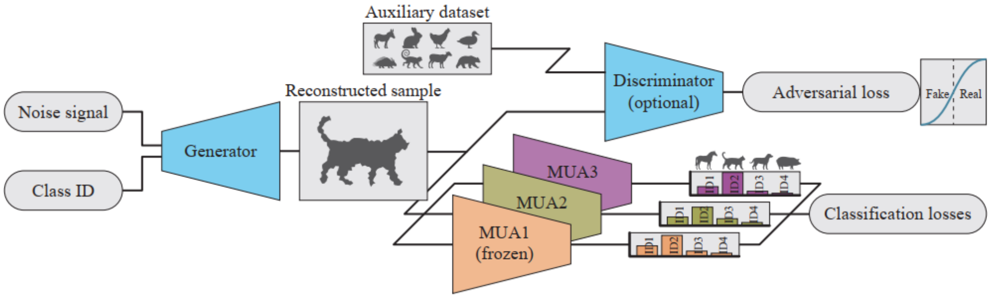


Figure 3.8: illustrates a model inversion method where a generator, guided by an ensemble of Models Under Attack, reconstructs original training images while being constrained by class prediction agreement and the realism of images compared to an auxiliary or, if unavailable, a data-free dataset. used by Wang Q et al,(2022).

sures realistic image generation, akin to a GAN, as depicted in Figure 3.8. This process benefits from diverse MUAs trained in different identity groups, enhancing the extraction of unique features from shared identities.

The researchers conducted experiments with ResNet-34-based face classifiers on the VGGFace2 dataset. Their findings illustrate that ensemble inversion not only refines the visual details of reconstructed images but also boosts attack accuracy compared to traditional single-model inversion methods. The accuracy of MNIST digit reconstruction rose by 70.9% in data-free scenarios and by 17.9% with auxiliary data, while face reconstruction accuracy increased by 21.1% over baseline methods.

Intriguingly, the paper[33] reveals that different face classifiers, even when trained with the same data, focus on various facial features, suggesting a combination of classifiers could lead to a more robust model inversion. With the application of tailored losses—such as one-hot loss and maximum output activation loss—further enhancements in sample quality were observed.

By highlighting the effectiveness of ensemble inversion in capturing distinct identity features without overlapping training and auxiliary data sets, the paper also underscores the need for developing defensive strategies against such sophisticated inversion techniques. The ultimate aim is a systematic exploration of model inversion's potential impact, setting the stage for future work in defense mechanisms.

3.2.3 Comparison table

Table 3.1 summarizes the key attributes of the reviewed articles on reconstruction attacks, providing details on the study year, the dataset used, the employed method, and the corresponding reconstruction performance metrics.

Table 3.1: Study characteristics of the articles included in the review for reconstruction attacks.

Study, yr	Dataset	Method	Reconstruction Performance Metrics.
Loo N et al, (2023)	MNIST and CIFAR-10	Neural Tangent Kernel-based attack; leverages MSE loss and network initialization; robust reconstruction on MNIST and CIFAR-10; connects reconstruction with dataset distillation for potential defense.	Reconstruction Quality: high. Dataset Distillation: high.
Haim N et al, (2022)	MNIST and CIFAR-10	Dataset Reconstruction via Gradient Flow Bias; employs KKT conditions in homogeneous networks; optimized using stationarity, dual feasibility, and dataset priors; demonstrated results on MNIST and CIFAR-10 with potential to identify training samples on margin.	SSIM metric. Train errors are zero. Test accuracy is 88.0% for MNIST. Test accuracy is 77.6% for CIFAR-10.
Balle B et al, (2022)	MNIST and CIFAR-10	Reconstructor network (RecoNN) method that predicts omitted training data using the weights of the targeted machine learning model. This approach effectively reconstructs image classifiers' training data from models trained on MNIST and CIFAR-10 datasets.	Mean Squared Error (MSE) is 0.0089 for MNIST and 0.0049 for CIFAR-10.Signal-to-Noise Ratio (PSNR) values for MNIST range from 30 dB to 20 dB.
Guan J et al, (2022)	multiple datasets such as CIFAR10, Tiny-ImageNet, CIFAR10-C, CIFAR100, and ImageNet.	SAC-w, which uses misclassified normal samples, and SAC-m, which employs augmented samples via CutMix and image flipping to detect model theft.	AUC-ROC curve. AUC value.
Guo C et al, (2022)	Adult dataset, MNIST dataset, CIFAR-10 dataset, CelebA dataset	Rényi Differential Privacy (RDP): Offers a unified framework for evaluating the tradeoff between privacy and utility within differential privacy mechanisms. Fisher Information Leakage (FIL): Quantifies data leakage potential through machine learning model parameters using the Fisher information matrix.	Mean Squared Error (MSE). Membership Inference Attack Advantage (Adv).
Gong H et al, (2022)	/	Single-Sample Reconstruction Attack System (SSRAS) and Improved R-GAP Algorithm: SSRAS is a privacy attack system capable of image reconstruction in federated learning without requiring label availability. It extends the attack to convolutional neural networks. The Improved R-GAP Algorithm enhances privacy attacks by utilizing the DLG algorithm to derive ground truth.	Rank Analysis Index (RA-I) is 95.5%. Attack Success Rate (ASR) is 0.87.
Wang Q et al,(2022)	MNIST and VGGFace2	Ensemble Inversion Technique; utilizes diversity among Models Under Attack (MUAs) with one-hot and maximum activation losses; enhances reconstruction fidelity on VGGFace2 using ResNet-34 classifiers; increases face reconstruction accuracy by 21.1%; underscores importance of developing defenses against model inversion attacks.	Attack Accuracy. Mean Squared Error (MSE).

Chapter 4

Materials and Methods

4.1 Dataset

The dataset which was used in this research is part of CASIA V1.0. This dataset comprises 756 iris images from 108 eyes, with each eye contributing seven images captured in two sessions using the CASIA close-up camera. The self-developed camera features eight 850mm NIR illuminators for uniform and ample illumination. To protect intellectual property rights (IPR), pupil regions in CASIA IrisV1 are automatically replaced with a constant intensity circular region, eliminating specular reflections from NIR illuminators. While this editing simplifies boundary detection, it has minimal or no impact on other components of an iris recognition system, such as feature extraction and classifier design.

The images in the database are stored in BMP format with a resolution of 320x280, providing a valuable resource for iris recognition system research. For within-class variability assessment, it is recommended to compare samples from the same eye captured in different sessions, facilitating effective training and testing scenarios. The arrangement of samples from different sessions is illustrated in Figure 4.1, where three samples are collected in the first session (Figure 4.1(a)) and four in the second session (Figure4.1 (b)). The data is available online [1]

4.2 Recognition

The realm of iris recognition, pivotal in biometric systems, has evolved with machine learning, promising heightened accuracy. In this context, we'll be able to explain the models which used in our experiment. the complexities of iris patterns necessitate refined computational approaches by leveraging deep learning.

This research aims to automate and optimize the recognition process and contribute to enhancing iris recognition systems, advancing their precision and reliability through the strategic reconstruction of trained data.

[1] http://biometrics.idealtest.org/findTotalDbByMode.do?mode=Iris%23/datasetDetail/1

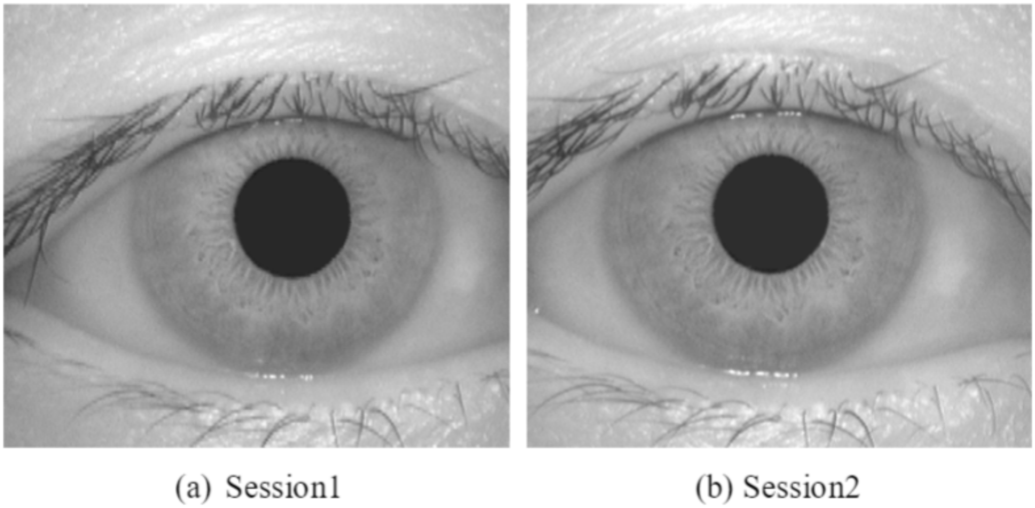


Figure 4.1: Example iris images in CASIA V1.0.

4.2.1 Preprocessing

Before the images are used for the training of the model, a pre-processing step is carried out. The main steps of iris image preprocessing are (I) Threshoulding (II) Morphological Operations (III) Find the contours (IV)GaussianBlur for iris mask (V) Detect circles (VI) Iris mask and cropping.

4.2.1.1 Threshoulding

Thresholding, the simplest method, involves deriving a binary image from an original grayscale image [34]. The outcome of these thresholding strategies is illustrated in Figure 4.2.

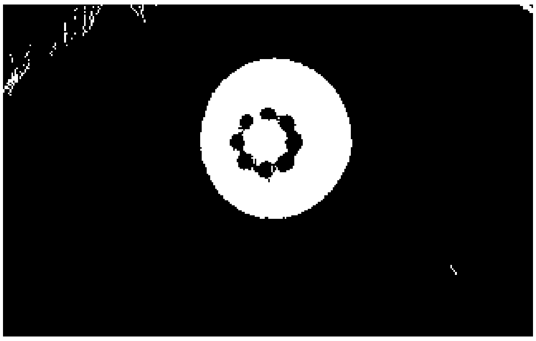


Figure 4.2: Outcome of the Thresholding Process.

4.2.1.2 Morphological Operations

I performed two morphological operations in this step: opening and a subsequent dilation process. The opening process involves a sequence of erosion and dilation operations for image enhancement which is defined by the equation [35]:

$$Opening = IM \ominus SE \oplus SE \tag{4.1}$$

Dilation involves systematically scanning an image with a structuring element, occasionally extending beyond its borders, enhancing pixel value alignment. This process contributes to the expansion of regions in the image [35]. The synergistic interplay of open and dilation, as depicted in Figure 4.3, within the opening process contributes to the enhancement of image clarity.

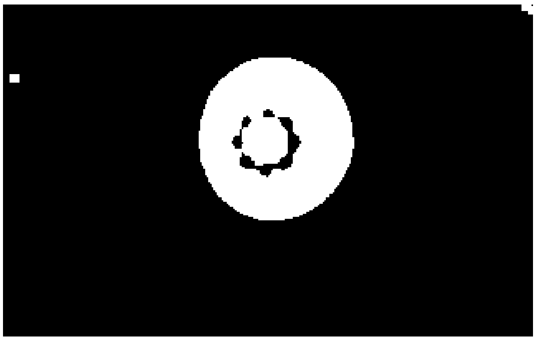


Figure 4.3: Result of the Opening and Dilation Morphological Processes.

4.2.1.3 Find the contours

The contours in images represent the boundaries of distinct objects or regions with similar pixel intensity. Utilizing the active contour model for segmentation to localize the boundary of the pupil with other unuseful details [36]. The figure 4.4 showed the result of this step.

4.2.1.4 GaussianBlur for iris mask

Gaussian blur is a versatile image processing technique for noise suppression, softening, and local averaging [37]. Gaussian Blur for an iris mask refers to the application of a Gaussian filter to smooth and reduce noise in the mask representing the iris region.

4.2.1.5 Detect circules

The foundation of many circle detection algorithms lies in the utilization of the Hough transform [38] which includes parameters representing ' dp ' adjusts accumulator resolution, 'minDist' sets minimum circle center distance, 'param1' influences edge detection, 'param2' is the accumulator threshold, and 'minRadius' and 'maxRadius' define accepted circle sizes and the result in figure 4.5.

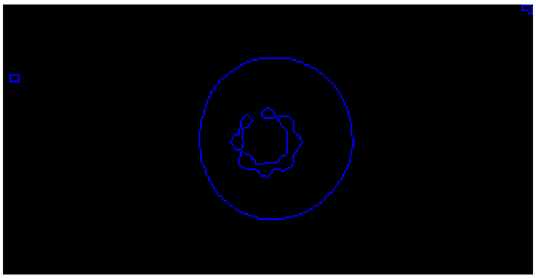


Figure 4.4: Outcome of Contour Detection Process.

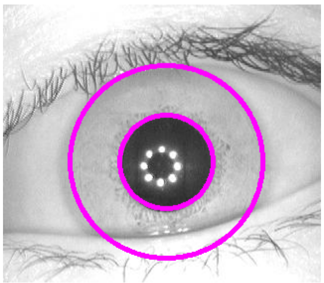


Figure 4.5: displays the outcome of the Hough Circles algorithm.

4.2.1.6 Iris mask and cropping

Iris mask creation involves accurately delineating the iris boundaries, followed by cropping the original image based on the extracted region. Figure 4.6 illustrates the outcome of cropping images using the iris mask, showcasing images focused solely on the iris region[39].

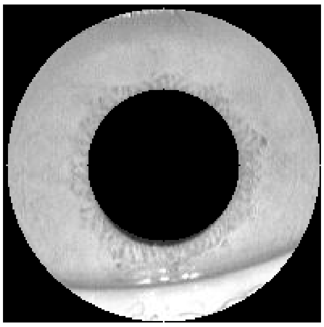


Figure 4.6: displays the outcome of the Hough Circles algorithm.

4.2.2 Model and Training

In this section, we present the architecture of the model and outline the chosen hyperparameters for training.

4.2.2.1 Prepare the data for one class classification

In implementing the recognition model, the utilization of One-Class Classification (OCC) stands out as a distinctive approach within the broader context of multi-class classification. OCC is tailored for scenarios where the training data exclusively comprises a single positive class, aiming to construct a representation and/or classifier that enhances the recognition of positively labeled queries during inference [40].

Upon partitioning the dataset into outcome target and non-target categories, as depicted in Figure 4.7, the division involves 70% for training data and 30% for validation data. This strategic division ensures a balanced and comprehensive assessment of the model's performance across various phases [40].

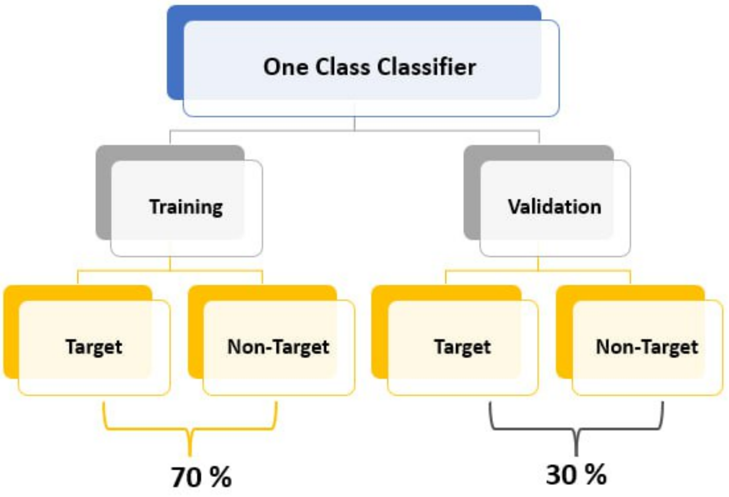


Figure 4.7: Dataset Division Using One-Class Classification.

4.2.2.2 Model Description

This model is inspired from this article [17].

The architecture of the iris recognition system involves three layers of convolutional layer, each with a pooling layer of (2, 2) and dropout layer with a rate of 0.1. The first convolutional layer includes six filters of size (3, 3) with a ReLU activation function, and the strides parameter is set to (1, 1), meaning the filters move one pixel at a time. The second convolutional layer with 64 filters of size (3, 3) and a ReLU activation function. Third convolutional layer with 128 filters of size (3, 3) and a ReLU activation function. The network processes an input image with dimensions (280, 320, 3).

The subsequent flattened layer transforms the output of the convolutional layers into a 1D array, priming it for fully connected layers. Three Dense (Fully Connected) Layers follow, comprising 1028, 512, and 64 neurons with ReLU activation functions.

These layers are adept at discerning intricate patterns within the flattened feature set. The final Dense Layer has a singular neuron activated by the sigmoid function, suitable for binary classification tasks, where the model outputs a probability denoting the likelihood of the input belonging to a specific class.

We constructed a recognition model inspired by the research[29], which focuses on data reconstruction. The second model is designed to be simpler than the initial one.

The second model architecture consists of a Sequential model with a convolutional layer followed by max pooling and dropout layers. The initial Conv2D layer employs 32 filters of size (3, 3) with ReLU activation, processing images with dimensions (280, 320, 3). Subsequent max pooling reduces spatial dimensions, and a dropout layer with a dropout rate of 0.5 helps mitigate overfitting. The flattened output is connected to two dense layers with 64 and 1 neuron(s) respectively, activated by ReLU and sigmoid functions.

These two models are configured with the Adam optimizer utilizing an exponential decay learning rate schedule, commencing at an initial rate of 0.000001. Binary cross-entropy is the loss function for binary classification, and accuracy is the performance metric. Training unfolds over 40 epochs with a batch size of 16, dynamically adjusting the learning rate, and is validated using data generators during the training process. The summaries of the first and second recognition models are presented in Tables 4.1 and Table 4.2, respectively.

Layer (type)	**Output Shape**	**Param #**
conv2d_1 (Conv2D)	(None, 278, 318, 32)	896
max_pooling2d_1 (MaxPooling2D)	(None, 139, 159, 32)	0
dropout_1 (Dropout)	(None, 139, 159, 32)	0
conv2d_2 (Conv2D)	(None, 137, 157, 64)	18,496
max_pooling2d_2 (MaxPooling2D)	(None, 68, 78, 64)	0
dropout_3 (Dropout)	(None, 68, 78, 64)	0
conv2d_3 (Conv2D)	(None, 66, 76, 128)	73,856
max_pooling2d_3 (MaxPooling2D)	(None, 33, 38, 128)	0
dropout_43 (Dropout)	(None, 33, 38, 128)	0
flatten (Flatten)	(None, 160512)	0
dense_1 (Dense)	(None, 1028)	165,007,364
dense_2 (Dense)	(None, 512)	526,848
dense_3 (Dense)	(None, 64)	32,832
dense_4 (Dense)	(None, 1)	65
Total params: 165,660,357 (631.94 MB)		
Trainable params: 165,660,357 (631.94 MB)		
Non-trainable params: 0 (0.00 Byte)		

Table 4.1: Summary of the first Recognition Model, inspired by ZHUANG Y et al., 2020 [17].

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 278, 318, 32)	896
max_pooling2d (MaxPooling2D)	(None, 139, 159, 32)	0
dropout (Dropout)	(None, 139, 159, 32)	0
flatten (Flatten)	(None, 707232)	0
dense_1 (Dense)	(None, 64)	45,262,912
dense_2 (Dense)	(None, 1)	65
Total params: 45,263,873 (172.67 MB)		
Trainable params: 45,263,873 (172.67 MB)		
Non-trainable params: 0 (0.00 Byte)		

Table 4.2: Summary of the second Recognition Model, inspired by Balle B et al, 2022 [29]

4.2.3 Recognition Evaluation

Two visual representations depict the outcomes of the training and validation processes, illustrating the accuracy and loss. Figure 4.8 showcases the training and validation accuracy, while Figure 4.9 presents the training and validation loss.

The evaluation of our recognition model employed multiple metrics. Equation 4.2 defines the accuracy, Equation 4.3 quantifies Precision, Equation 4.4 specifies Recall, and Equation 4.5 delineates the F1_score. Additionally, Figure 4.10 visually portrays the Confusion Matrix, offering a comprehensive insight into the model's performance.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4.2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.4}$$

$$\text{F1\_score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.5}$$

4.3 Reconstructing

During the reconstruction phase, I applied a reconstruction algorithm inspired by the work of Balle B et al, (2022) as outlined in Section 3.2.2.3 of this research.

The algorithm proposed in this paper [29] is a reconstruction attack strategy based on training a reconstructor network. The reconstructor network is a neural network that is trained by the adversary to output a reconstruction of the target point when given as input the parameters of a released model.

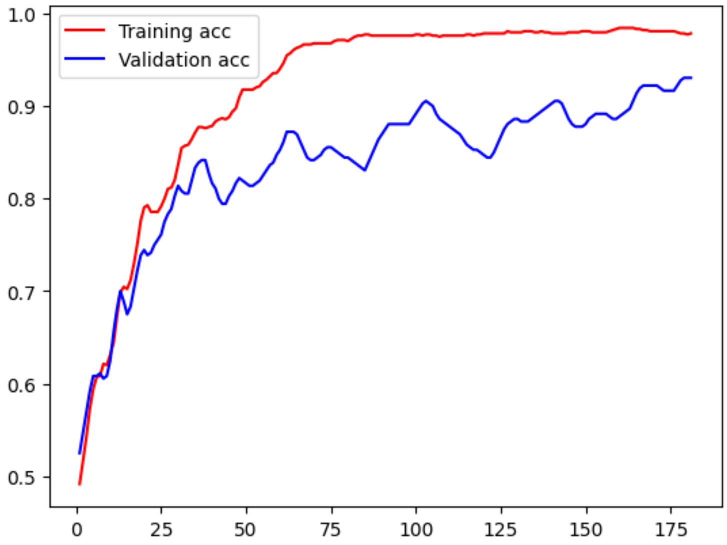


Figure 4.8: Training and validation accuracy

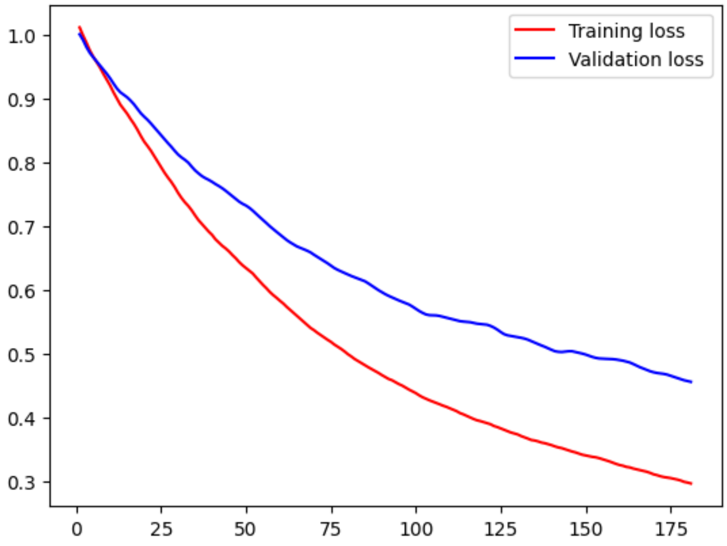
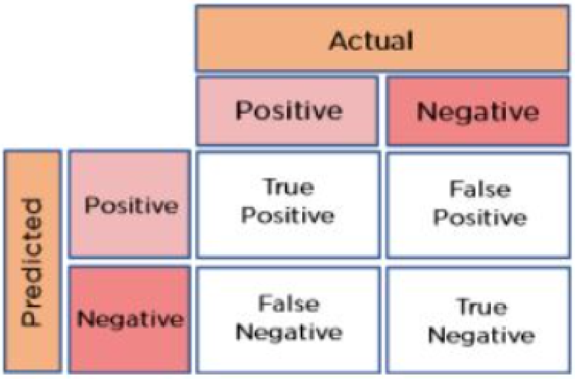


Figure 4.9: Training and validation loss



		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Figure 4.10: Confusion Matrix

The attack strategy involves the following steps:

1. The adversary obtains access to a released machine learning model and its parameters.

2. The adversary trains a reconstructor network using the model parameters and a subset of the training data.

3. The adversary uses the reconstructor network to reconstruct the remaining training data points.

4. The adversary can then use the reconstructed training data to train a new model that is similar to the original model.

This process is visually represented in Figure 4.11.

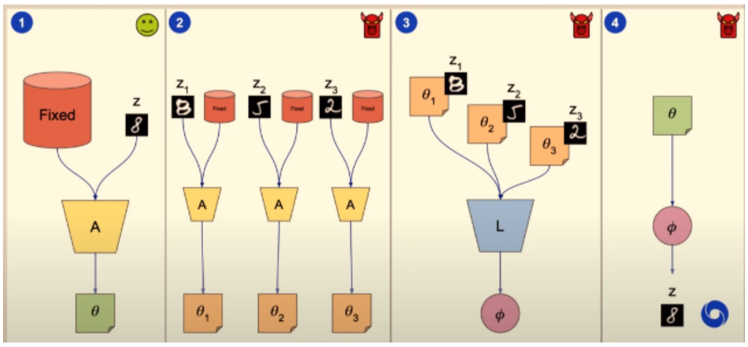


Figure 4.11: A general Reconstruction Attack

The paper provides a theoretical analysis of reconstruction attacks against simple machine learning models like linear, logistic, and ridge regression, as well as against standard neural network architectures for image classification.

Algorithm 1 formalizes the reconstruction attack, detailing the interaction between the model developer and the informed adversary. It utilizes the trained model θ on data D with an additional point $\{z\}$, employing attack algorithm R to produce a faithful reconstruction $\hat{z}$, evaluated based on the chosen error function ℓ, reflecting privacy concerns and context-specific considerations. Privacy expectations may not require perfect reconstruction, and the paper explores various metrics, including MSE and model output similarities, with the choice of ℓ and the threshold for success depending on the specific application and potential harm to individuals.

The paper draws a connection between membership inference attacks (MIA) showed algorithm 2 and the proposed attack strategy. It introduces an informed MIA adversary, more powerful than standard MIA, and emphasizes the significance of differential privacy (DP) as strong privacy protection against both informed MIA and accurate reconstruction.

This algorithm was firstly applied using the Github repository of the work [29]. During this step, MNIST dataset was utilized as input for training the MLP model. The results of the reconstruction model are visually depicted in Figure 4.12.

Algorithm 1 Reconstruction attack with an informed adversary. (Auxiliary side knowledge aux is optional).

```
procedure RECONSTRUCTION($A, R, D, z$; aux)
    $\theta \leftarrow A(D \cup \{z\})$
    $\hat{z} \leftarrow R(\theta, D_-, A; \text{aux})$
    return $\ell(z, \hat{z})$
end procedure
```

Algorithm 2 Informed Membership Inference Attack

```
procedure INFORMED-MIA($A, M, D, z_0, z_1$)
    $b \leftarrow \text{Unif}(\{0, 1\})$
    $\theta \leftarrow A(D \cup \{z_b\})$
    $\hat{b} \leftarrow M(\theta, D_-, A, z_0, z_1)$
    return $b = \hat{b}$
end procedure
```

4.4 Reconstruction Implmenetation

Upon completion of training for the Recognition Models outlined in Section 4.2.2.2, both the first and second models were saved using two distinct methods:

1. Save Model H5: The entirety of a deep learning model's architecture, weights, optimizer state, and training setup are stored in the Hierarchical Data Format (HDF5) file format when a model is saved as an H5 file.

2. Side Knowledge Model: Pre-trained models are employed as feature extractors or customized for particular tasks in side knowledge models, which are commonly employed in transfer learning scenarios. Unlike H5 files, which save the whole model architecture and weights, side knowledge models simply save particular layers or components that provide important information about broad patterns discovered during pre-training.

In order to test the two models (outlined in Section 4.2.2.2) with the two different methods (h5 and side knowledge), four scenarios were implemented. The four scenarios include matching the two recognition models and two saving methods together. The general scheme is illustrated in Fig 4.13, showing the overarching steps involved in the reconstruction process.

4.4.1 Evaluation

In order to evaluate the implemented reconstruction, two methods of evaluation were followed. The first method involved the usage of the same input model. The reconstructed data were used as input to try to gain non-legitimate authentication. This was tested using both h5 and side knowledge models. The second method includes measuring the similarity between the legitimate eye iris data and the

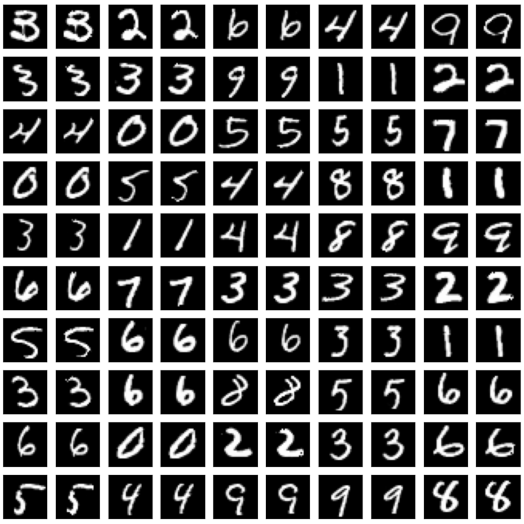


Figure 4.12: displays the output of the Reconstruction model, where odd columns represent the target, and even columns showcase the corresponding reconstructions.

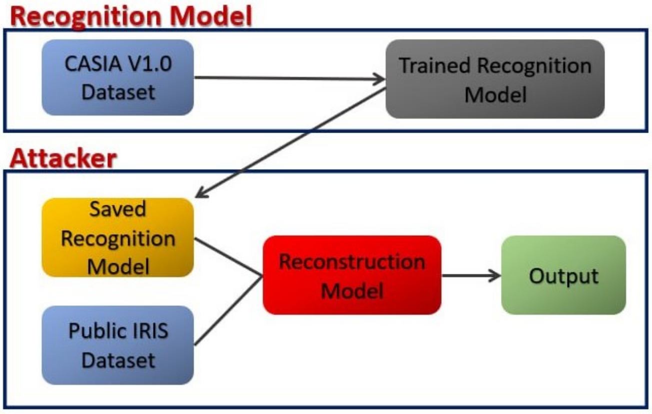


Figure 4.13: Reconstruction process scheme.

reconstructed ones. This evaluation aims to measure to what extent the reconstructed data have the ability to gain false authentication.

Chapter 5

Experimental Results

5.1 Recognition Model Accuracy

Table 5.1 provides a comparative analysis of the accuracy achieved by our Recognition model and those discussed in the literature review(Section 2.5), all of which utilized the same CASIA dataset.

Table 5.1: Recognition Models Accuracy

The model for Recognition	Dataset	Accuracy
Vijaykumar V et al, 2022	CASIA V1.0	99.85%
ZHUANG Y et al, 2020	CASIA-iris V4.0	99%
Thakkar S et al, 2020	CASIA-iris V4.0	CRR 99%
First Model (inspired by [17])	CASIA V1.0	96.43%
Second Model (inspired by [29])	CASIA V1.0	92.86%

5.2 Reconstrution Result

We utilized all four Recognition Model saving scenarios outlined in Section 4.4 as input for the reconstruction model. The results, depicted in Figure 5.1, consistently demonstrated similar outcomes across all scenarios.

5.3 Measuring the Similarity

5.3.1 Similarity Descriptive Statistics

We conducted a comprehensive assessment of similarity across three stages, employing Descriptive Statistics to gain insights into data distribution.

- In the first stage, we evaluated the similarity score between each reconstructed image and its corresponding actual image. Table 5.2 encompasses the Descriptive Statistics values, while Figure 5.2 visually represents the data distribution through a Histogram with mean and median, Violin plot, and Box plot.

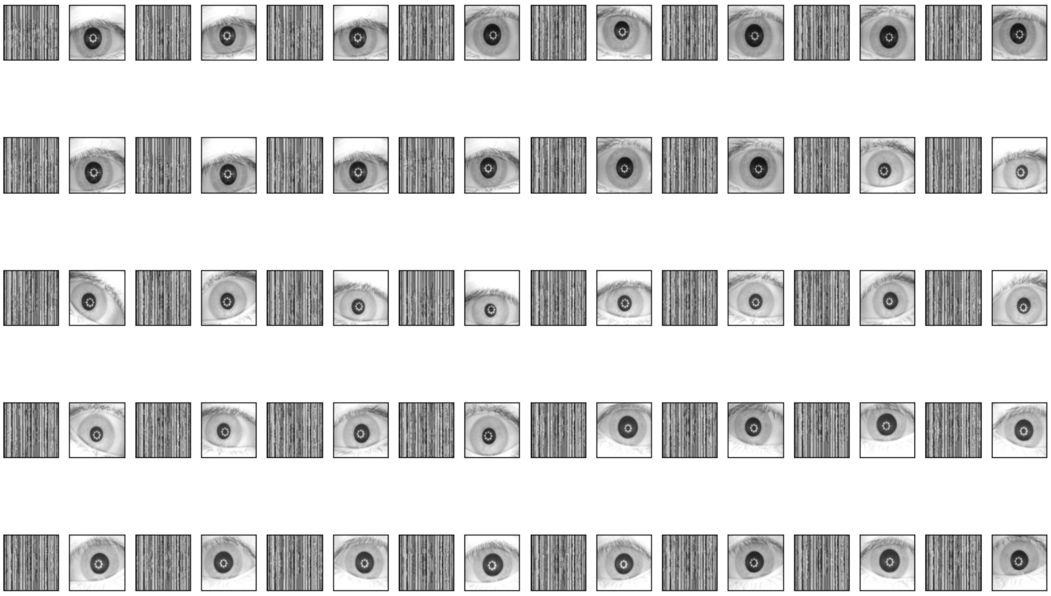


Figure 5.1: illustrates the outcomes of the reconstruction model, showcasing a side-by-side comparison of the reconstructed (odd columns) and actual (even columns) images.

Table 5.2: Descriptive Statistics Summary for all stages.

	Value (Stage 1)	Value (Stage 2)	Value (Stage 3)
Mean	0.0078	0.0078	0.7675
Median	0.008	0.008	0.7577
Minimum	0.0065	0.0065	0.7348
Maximum	0.0096	0.0096	0.818
Range	0.0031	0.0031	0.0832
Standard Deviation	0.001	0.001	0.0299
Variance	0.0	0.0	0.0009
Skewness	0.1093	0.1093	0.4477
Kurtosis	-1.0007	-1.0007	-1.1883
25th Percentile	0.0069	0.0069	0.7423
50th Percentile	0.008	0.008	0.7577
75th Percentile	0.0085	0.0085	0.7869
Interquartile Range	0.0016	0.0016	0.0446

- For the second stage, we considered the first image of actual images as a template. Afterwards, we have measured the similarity between each reconstructed image and the template. Descriptive Statistics for this stage are presented in Table 5.2, and Figure 5.3 provides a graphical representation of the data.
- In the third stage, we extended our analysis to measure the similarity between

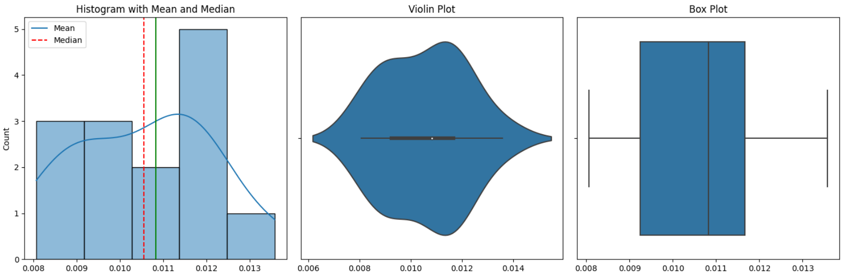


Figure 5.2: encompassing a Histogram with mean and median, a Violin plot, and a Box plot for the first stage.

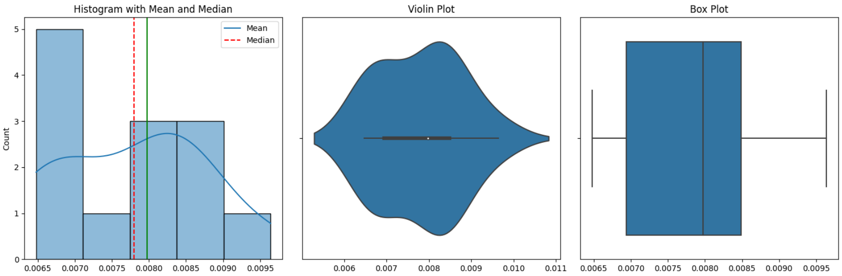


Figure 5.3: encompassing a Histogram with mean and median, a Violin plot, and a Box plot for the second stage.

each actual image and the template. The aim here is to make a comparison between similarities of remonstrated images and actual images with the template. Table 5.2 compiles the Descriptive Statistics values for this stage, while Figure 5.4 depicts the data distribution through various plots.

We utilized Histogram with Kernel Density Estimation as depicted in Figure 5.5 to visually represent the data distribution between the reconstructed images and the actual images. This shows both histograms of the second and the third stages.

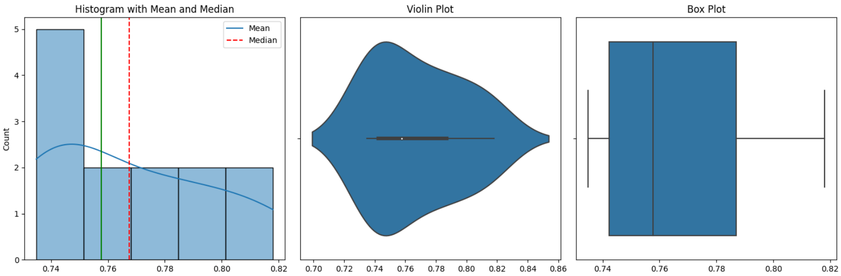


Figure 5.4: encompassing a Histogram with mean and median, a Violin plot, and a Box plot for the third stage.

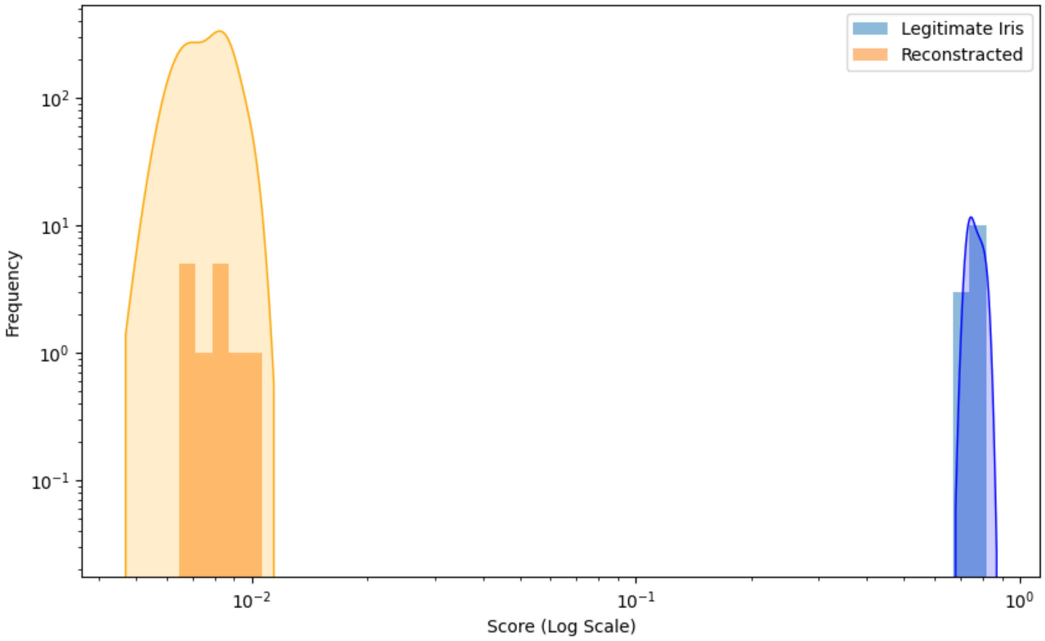


Figure 5.5: Histograms and KDE of two datasets.

5.3.2 Similarity Statistical Analysis

The aim of the statistical analysis is to find whether there is a statistically significant difference between the similarities of reconstructed and actual images with the template. We consider the the alpha level of 0.05 ($\alpha = 0.05$)

Firstly. Shapiro-Wilk test was conducted to assess the normality of data. This will lead us to choose a parametric or non-parametric test for the difference test.

Starting with Shapiro-Wilk test, we firstly present

1. **Null Hypothesis (H0):** The sample data follow a normal distribution. In other words, there is no significant departure from normality in the population distribution from which the sample was drawn.

2. **Alternative Hypothesis (H1):** The sample data do not follow a normal distribution. This implies that there is a significant departure from normality in the population distribution.

After Conducting Shapiro-Wilk test, the following results were obtained:

- **Actual images similarities:** the p-value is 0.1361. Since this value is greater than the alpha level of 0.05, we do fail to reject the null hypothesis. This means there is not enough evidence to say that the data is not normally distributed.

- **Reconstructed images similarities:** the p-value is 0.4304, which is also greater than 0.05. This leads to the same conclusion: there is not sufficient evidence to reject the null hypothesis of normal distribution.

Given that the two lists appear to be normally distributed, it's appropriate to use a parametric test for comparing the two lists. In this case, the t-test is a suitable choice. The paired t-test is a statistical test used to compare the means of two related groups to determine if there is a statistically significant difference between them.

Moving topaired t-test, we present present.

1. **Null Hypothesis (H0):** There is no significant difference between the means of the two related groups.

2. **Alternative Hypothesis (H1):** there is a significant difference between the means of the two related groups.

The output of the paired t-test provides three key statistics:

1. Statistic: The t-statistic is 93.41.

2. P-value: The p-value is approximately 1.52×10^{-18}.

Given that the p-value is statistically smaller than alpha value, we can reject the Null Hypothesis and say that the Alternative Hypothesis is true. We can say that there is a significant difference between the means of the two related groups. In other words, there is a significant difference between the biometric similarities and actual similarities.

Chapter 6

Discussion and Conclusion

6.1 Discussion

The work of this project started with reproducing an iris recognition model based on [17]. and Building another iris recognition model inspired from [29]. Looking at table 5.1, we can compare the performance of our models with the literature review (section 2.5). It showed that our system has a high authentication rate as a One-Class Classifier. It shows that our model has a slightly lower accuracy that could be neglected since the focus of our work is to test the reconstructing attacks.

The reconstruction attack was applied to the model given that the attacker knows the model and has a side knowledge about the context. The results of the reconstructing attacks are shown in Fig 5.1. For a general comparison of the naked eye, we can say that. The reconstruction attack failed to regenerate images with a similar look to the actual original ones. However, it might have the same embedded patterns as the actual ones.

In order to catch any hidden similar patterns between the reconstructed images and the actual ones, we have measured the similarity between each group of images (reconstructed and actual) with an iris template of the person we want to authenticate. The similarity score based on was calculated using SSIM (The Structural Similarity Index). The results of the similarity indices are presented in Fig 5.2 and Fig 5.3 and 5.4 and also in table 5.2. From the descriptive statistics of the similarity indices, we can observe that the similarity between the reconstructed images with templates has a mean value of 0.0078 ± 0.001 while the actual similarities mean equles to 0.7675 ± 0.0299. This showed that there is a significant difference between the similarity score of the two groups, that is visually proven in Fig 5.5.

Moving on to find whether there is a statistical difference between the similarity indices, we have conducted a statistical analysis. Based on the analysis in section 5.3.2, we can conclude with a high degree of confidence that there is a statistically significant difference between the similarity scores of legitimate, original biometric data and reconstructed biometric data, when each is compared to the same reference template. The exceedingly high t-statistic, combined with an exceptionally low p-value, strongly indicates that reconstructing biometric data has a measurable and significant impact on its similarity to the reference template.

The significance of these findings extends to the realm of biometric security and authentication systems, shedding light on potential challenges in maintaining the fidelity of reconstructed biometric data. The notable disparity in similarity scores raises concerns about the reliability and security of reconstructed data compared to the original, legitimate biometric data. This variance has the potential to impact the overall effectiveness and security of biometric verification and identification processes. While the current reconstructed images exhibit discernible content, their limited ability to accurately recognize fundamental features, such as the shapes of eyes, poses a current challenge to iris-based biometric authentication. Although existing reconstruction methods may not currently pose a significant threat, it's imperative to recognize the possibility of more sophisticated techniques emerging in the future, presenting a heightened risk to the security of iris-based biometric authentication.

6.2 Conclusion

In conclusion, the research presented in this project concentrated on the development and evaluation of iris recognition models, with an emphasis on their robustness against reconstruction attacks. Two models were created, inspired by previous research, and tested using the CASIA V1.0 dataset. Despite the models' high authentication rates, their accuracy was slightly lower, which is to be expected given the emphasis on testing reconstruction attacks.

The investigation of reconstruction attacks on iris recognition models is the study's main contribution. While such attacks could reproduce patterns similar to the actual images, they were unable to closely replicate the images' appearance. This finding was supported by extensive similarity testing with the Structural Similarity Index (SSIM), which revealed a significant difference in similarity scores between original and reconstructed biometric data. This significant difference was confirmed by statistical analysis, emphasising the impact of reconstruction on biometric data fidelity.

These findings are especially important in the context of biometric security and authentication systems. The significant difference in similarity scores between original and reconstructed data suggests that reconstructed biometric data may be untrustworthy and insecure. This has significant implications for the effectiveness and security of biometric systems, implying that reconstruction methods and their impact on biometric verification and identification processes must be carefully considered.

Overall, the study emphasises the difficulties in maintaining the integrity and reliability of biometric data in the face of advanced reconstruction attacks, emphasising the importance of continuous advancements in biometric security measures.

Bibliography

[1] V. Panwar *et al.*, "A review on iris recognition system using machine and deep learning," in *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pp. 857–866, IEEE, 2022.

[2] K. Nguyen, H. Proença, and F. Alonso-Fernandez, "Deep learning for iris recognition: A survey," *arXiv preprint arXiv:2210.05866*, 2022.

[3] S. Minaee, A. Abdolrashidi, H. Su, M. Bennamoun, and D. Zhang, "Biometrics recognition using deep learning: A survey," *Artificial Intelligence Review*, pp. 1–49, 2023.

[4] N. A. Alzahab, M. Baldi, and L. Scalise, "Efficient feature selection for electroencephalogram-based authentication," in *2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 1–6, IEEE, 2021.

[5] N. A. Alzahab, A. Di Iorio, M. Baldi, and L. Scalise, "Effect of auditory stimuli on electroencephalography-based authentication," in *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*, pp. 388–392, IEEE, 2022.

[6] D. Bhattacharyya, R. Ranjan, F. Alisherov, M. Choi, *et al.*, "Biometric authentication: A review," *International Journal of u-and e-Service, Science and Technology*, vol. 2, no. 3, pp. 13–28, 2009.

[7] S. R. Kodituwakku, "Biometric authentication: A review," *Int. J. Trend Res. Dev*, vol. 2, pp. 113–123, 2015.

[8] I. Natgunanathan, A. Mehmood, Y. Xiang, G. Beliakov, and J. Yearwood, "Protection of privacy in biometric data," *IEEE access*, vol. 4, pp. 880–892, 2016.

[9] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on circuits and systems for video technology*, vol. 14, no. 1, pp. 4–20, 2004.

[10] Y. Sutcu, E. Tabassi, H. T. Sencar, and N. Memon, "What is biometric information and how to measure it?," in *2013 IEEE international conference on technologies for homeland security (HST)*, pp. 67–72, IEEE, 2013.

[11] J. Wayman, A. Jain, D. Maltoni, and D. Maio, “An introduction to biometric authentication systems,” in *Biometric systems: Technology, design and performance evaluation*, pp. 1–20, Springer, 2005.

[12] K. W. Bowyer, K. Hollingsworth, and P. J. Flynn, “Image understanding for iris biometrics: A survey,” *Computer vision and image understanding*, vol. 110, no. 2, pp. 281–307, 2008.

[13] M. N. T. A. Chmielewski *et al.*, “An iris biometric system for public and personal use,” *IEEE catalog*, no. 0018-9162, 2000.

[14] J. Zuo, N. K. Ratha, and J. H. Connell, “Cancelable iris biometric,” in *2008 19th International conference on pattern recognition*, pp. 1–4, IEEE, 2008.

[15] J. J. Winston and D. J. Hemanth, “A comprehensive review on iris image-based biometric system,” *Soft Computing*, vol. 23, pp. 9361–9384, 2019.

[16] V. Vijaykumar and K. Selvam, “Kcir: A novel iris recognition system using deep cnn with kalman filtering,” in *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*, pp. 997–1003, IEEE, 2022.

[17] Y. Zhuang, J. H. Chuah, C. O. Chow, and M. G. Lim, “Iris recognition using convolutional neural network,” in *2020 IEEE 10th International Conference on System Engineering and Technology (ICSET)*, pp. 134–138, IEEE, 2020.

[18] S. Thakkar and C. Patel, “Iris recognition supported best gabor filters and deep learning cnn options,” in *2020 International Conference on Industry 4.0 Technology (I4Tech)*, pp. 167–170, IEEE, 2020.

[19] M. V. Vizoni and A. N. Marana, “Ocular recognition using deep features for identity authentication,” in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 155–160, Ieee, 2020.

[20] T. Sudhakar and M. Gavrilova, “Multi-instance cancelable biometric system using convolutional neural network,” in *2019 International Conference on Cyberworlds (CW)*, pp. 287–294, IEEE, 2019.

[21] A. Boyd, A. Czajka, and K. Bowyer, “Deep learning-based feature extraction in iris recognition: Use existing models, fine-tune or train from scratch?,” in *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–9, IEEE, 2019.

[22] Y. Li, L. Zhu, X. Jia, Y. Jiang, S.-T. Xia, and X. Cao, “Defending against model stealing via verifying embedded external features,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 1464–1472, 2022.

[23] S. Chang and C. Li, “Privacy in neural network learning: threats and countermeasures,” *IEEE Network*, vol. 32, no. 4, pp. 61–67, 2018.

[24] E. De Cristofaro, "A critical overview of privacy in machine learning," *IEEE Security & Privacy*, vol. 19, no. 4, pp. 19–27, 2021.

[25] V. Chandrasekaran, K. Chaudhuri, I. Giacomelli, S. Jha, and S. Yan, "Exploring connections between active learning and model extraction," in *29th USENIX Security Symposium (USENIX Security 20)*, pp. 1309–1326, 2020.

[26] G. Wang, J. C. Ye, K. Mueller, and J. A. Fessler, "Image reconstruction is a new frontier of machine learning," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1289–1296, 2018.

[27] N. Loo, R. Hasani, M. Lechner, and D. Rus, "Dataset distillation fixes dataset reconstruction attacks," *arXiv preprint arXiv:2302.01428*, 2023.

[28] N. Haim, G. Vardi, G. Yehudai, O. Shamir, and M. Irani, "Reconstructing training data from trained neural networks," *arXiv preprint arXiv:2206.07758*, 2022.

[29] B. Balle, G. Cherubin, and J. Hayes, "Reconstructing training data with informed adversaries," in *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1138–1156, IEEE, 2022.

[30] J. Guan, J. Liang, and R. He, "Are you stealing my model? sample correlation for fingerprinting deep neural networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36571–36584, 2022.

[31] C. Guo, B. Karrer, K. Chaudhuri, and L. van der Maaten, "Bounding training data reconstruction in private (deep) learning," in *International Conference on Machine Learning*, pp. 8056–8071, PMLR, 2022.

[32] H. Gong, L. Jiang, X. Liu, Y. Wang, L. Wang, and K. Zhang, "Recover user's private training image data by gradient in federated learning," *Sensors*, vol. 22, no. 19, p. 7157, 2022.

[33] Q. Wang and D. Kurz, "Reconstructing training data from diverse ml models by ensemble inversion," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2909–2917, 2022.

[34] M. Kamel and A. Zhao, "Extraction of binary character/graphics images from grayscale document images," *CVGIP: Graphical Models and Image Processing*, vol. 55, no. 3, pp. 203–217, 1993.

[35] K. A. M. Said, A. B. Jambek, and N. Sulaiman, "A study of image processing using morphological opening and closing processes," *International Journal of Control Theory and Applications*, vol. 9, no. 31, pp. 15–21, 2016.

[36] J. Koh, V. Govindaraju, and V. Chaudhary, “A robust iris localization method using an active contour model and hough transform,” in *2010 20th International Conference on Pattern Recognition*, pp. 2852–2856, IEEE, 2010.

[37] J. Flusser, S. Farokhi, C. Höschl, T. Suk, B. Zitova, and M. Pedone, “Recognition of images degraded by gaussian blur,” *IEEE transactions on Image Processing*, vol. 25, no. 2, pp. 790–806, 2015.

[38] Z. Yao and W. Yi, “Curvature aided hough transform for circle detection,” *Expert Systems with Applications*, vol. 51, pp. 26–33, 2016.

[39] M. Karakaya, “Deep learning frameworks for off-angle iris recognition,” in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–8, IEEE, 2018.

[40] P. Perera, P. Oza, and V. M. Patel, “One-class classification: A survey,” *arXiv preprint arXiv:2101.03064*, 2021.