CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD

A Heterogeneity Aware Federated Learning Framework for Cross Country Iris Verification

by

Owais Ali Khan

A thesis submitted in partial fulfillment for the
degree of Master of Science

in the

Faculty of Engineering
Department of Electrical Engineering

2026

CERTIFICATE OF APPROVAL

A Heterogeneity Aware Federated Learning Framework for Cross Country Iris Verification

by

Owais Ali Khan

(MEE243002)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Abdul Ghafoor	MCS-NUST, Rwp
(b)	Internal Examiner	Dr. Nadeem Anjum	CUST, Islamabad

Dr. Imtiaz Ahmad Taj
Thesis Supervisor
May, 2026

Dr. Noor Muhammad Khan
Head
Dept. of Electrical Engineering
May, 2026

Dr. Imtiaz Ahmad Taj
Dean
Faculty of Engineering
May, 2026

Author's Declaration

I, **Owais Ali Khan** hereby state that my MS thesis titled "**A Heterogeneity-Aware Federated Learning Framework for Cross Country Iris Verification**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Owais Ali Khan)

Registration No: MEE243002

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**A Heterogeneity Aware Federated Learning Framework for Cross-Country Iris Verification**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Owais Ali Khan)

Registration No: MEE243002

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Dr. Imtiaz Ahmad Taj, for his guidance, critical feedback, and continuous encouragement throughout the course of this research. His insistence on methodological rigor, originality, and clarity of thought played a decisive role in shaping this work and in developing my approach to research.

I would also like to acknowledge my parents and my family for their unconditional support, motivation and encouragement throughout my whole academic journey. Their belief in my abilities enabled me to complete this journey successfully.

Owais Ali Khan

Abstract

The centralization of biometric data for training deep learning models is frequently found to be unfeasible due to privacy laws, institutional standards, and the sensitive nature of raw biometric information. To overcome these operational challenges, this thesis proposes a federated learning framework for cross country iris verification, specifically designed to handle significant client heterogeneity. The proposed framework employs a SwinV2 Tiny Siamese network, which is trained using hybrid metric learning objectives, namely batch hard triplet loss and supervised contrastive loss.

To overcome the shortcomings of conventional aggregation techniques when confronted with heterogeneous data distributions, a novel aggregation strategy, termed FedHAT (Federated Heterogeneity Aware Training Framework), is presented. This method incorporates a two stage approach, a heterogeneity aware stabilization (warm-up) phase designed to collect supervision signals, succeeded by a learned aggregation phase. During this phase, client contributions are adaptively determined through a regression based decision function. Furthermore, client aggregation weights are modeled as a polynomial function of validation performance indicators, including Equal Error Rate (EER) and True Accept Rates (TAR) at low False Accept Rates, alongside dataset characteristics such as scale and identity richness.

The proposed framework is evaluated using seven iris datasets from different geographic regions. This includes the CUST-Iris dataset, which is part of the federated benchmark. The experimental results, based on identity disjoint test splits, show consistent improvements compared to standard federated learning methods like FedAvg, FedProx, and FedYogi. Superior performance is specifically observed at security critical operating points for underrepresented and heterogeneous clients, providing a scalable foundation for privacy preserving, multi institutional biometric collaboration.

Contents

List of Figures

List of Tables

Abbreviations

AUC	Area Under the Curve
BN	Batch Normalization
CNN	Convolutional Neural Network
EAS	Eye-Aware Identity Split
EER	Equal Error Rate
EM	Expectation–Maximization
FAR	False Accept Rate
FedAvg	Federated Averaging
FedHAT	Federated Heterogeneity-Aware Training
FedProx	Federated Proximal Optimization
FedYogi	Federated Yogi (Adaptive Server-Side Optimizer)
FL	Federated Learning
GDPR	General Data Protection Regulation
GELU	Gaussian Error Linear Unit
GPU	Graphics Processing Unit
IID	Independent and Identically Distributed
non-IID	non-Independent and Identically Distributed
LR	Learning Rate
ML	Machine Learning
NIR	Near Infrared
ROC	Receiver Operating Characteristic
SOTA	State of the Art
Swin	Shifted Window Transformer

SwinV2	Shifted Window Transformer Version 2
TAR	True Accept Rate
U-Net	U-shaped Convolutional Neural Network
ViT	Vision Transformer

Symbols

θ	Model parameters
D_k	Local dataset of client k
N_k	Number of local training samples
q_k	Number of unique iris identities
ϕ_k	Validation reliability score
m	Triplet loss margin
τ	Temperature parameter
λ	Hybrid loss balancing weight
W_k^{warm}	Warm-up aggregation weight
$\alpha, \beta, \gamma, \delta$	Heuristic control coefficients
$\hat{y}_k$	Predicted expected client utility
s_k	Softplus activated utility score
α_k	Final normalized aggregation weight

Chapter 1

Introduction

1.1 Background and Motivation

Iris identification has become a core biometric modality in a broad spectrum of high security systems such as border control, national identity management and in large scale access control systems. Because of its long term reliability, complex textural characteristics and innate resistance to spoofing attempts, the iris is often regarded as a highly reliable physiological property to be used in biometric authentication. These features have attracted significant studies on how to come up with more efficient and accurate iris recognition systems particularly in real life situations where security and reliability are the most important factors.

1.1.1 Iris Recognition and High Security Applications

In practical deployments, iris recognition is commonly preferred in scenarios where strong identity assurance is required and operational costs of false accepts are high. Border control and national identity programs rely on accurate verification at scale, while enterprise and institutional access management systems require consistent performance under diverse capture environments. The stability of the iris pattern over time, combined with its discriminative texture structure, makes iris based verification a dependable option for these use cases [1].

1.1.2 Centralized Deep Learning for Iris Recognition

Despite the maturity of research on iris recognition, most of the recent deep learning based pipelines are based on centralized training assumptions. In such setups, biometric data of different institutions are pooled into a common repository to support model development and refinement. Even though this methodology presents the benefit of larger datasets and possibly greater diversity to the model, it has a significant practical drawback, the centralized collection of biometric data is often not practical in real world applications. The centralization of biometric data has serious limitations because of privacy laws, including the GDPR, restrictions on cross border data regulation, institutional policies, and the sensitivity of biometric identifiers. This forbids the transmission, pooling, or storage of raw biometric samples outside their institution of origin, even when such collaborative learning would otherwise be beneficial.

1.1.3 Transformer Based Architectures for Fine Grained Iris Texture Modeling

Recent progress in deep learning has been strongly influenced by Transformer based architectures, most notably the Swin Transformer family [2]. These models have demonstrated strong performance for fine grained texture analysis in iris recognition settings [3–5]. The attention based modeling and hierarchical feature representation mechanisms of these architectures allow detailed texture patterns to be captured more effectively than many earlier backbone designs.

At the same time, such models are widely known to benefit substantially from large scale and diverse training corpora [6]. In many realistic deployments, however, access to sufficiently varied biometric data is constrained. This is especially true in cross institutional or cross country scenarios, where the exchange of biometric information is restricted, and where datasets differ considerably in sensor type, capture protocol, and population demographics.

1.1.4 Federated Learning as a Privacy Preserving Alternative

Federated Learning has become a desirable method of preserving privacy when training collaborative models [7, 8]. The cooperative learning mechanism in the Federated Learning model is done through sharing model parameters, including weight updates, instead of sensitive biometric information. Consequently, the institutions can collaboratively enhance a common global model and at the same time make sure that raw biometric data are localised and securely stored [9–11].

1.2 Federated Iris Verification Setting

This work explores federated iris verification across geographically dispersed institutions based on a SwinV2 Tiny Siamese architecture and trained with the metric learning objective. In particular, Batch hard triplet loss and supervised contrastive loss are deployed to generate iris embeddings in an unusual form as of verification tasks [12–14].

The datasets, which are a representation of countries, serve as individual federated clients. This arrangement ensures that iris information of a given identity is not shared among clients. Fig. 1.1 depicts an overview of federated framework.

1.3 Client Heterogeneity in Cross Institutional Biometric Federated Learning

Although Federated Learning offers a platform of privacy preserving collaboration, traditional aggregation schemes such as FedAvg make implicit assumptions about the contribution of participating clients that they contribute on equal footing and originate from similar data distributions [15]. This is usually not true in cross institutional biometric scenarios. In particular, there is a high level of heterogeneity of clients in various dimensions:

i. Dataset Scale: The volume of training samples vary significantly across institutions, potentially disadvantaging less dominant clients when sample weighted aggregation is employed.

ii. Identity Diversity: Clients differ in the number of unique identities and the extent of within identity variation, thereby influencing the quality and generalizability of the learned embeddings.

iii. Sensing Conditions and acquisition protocols: Cross country deployments frequently involve heterogeneous sensors and capture environments, which can alter feature distributions and modify embedding behavior.

iv. Image Quality: Variations in focus, illumination, occlusion, and noise can affect the reliability and utility of client updates.

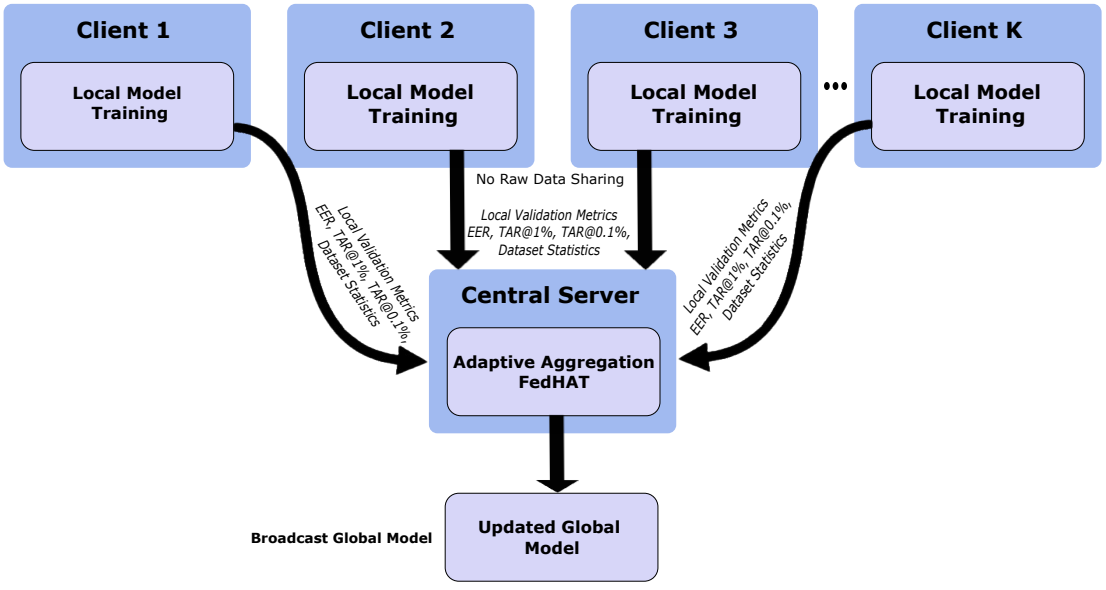

FIGURE 1.1: Overview of the federated iris verification framework.

In the case of naive aggregation on non-IID and imbalanced data, the global model may be biased with respect to the clients whose dataset is larger or of higher quality.

Consequently, the validity of the model in terms of verifying the data of under-represented or poorer quality may be reduced. This restricts the feasibility of the conventional Federated Learning procedures in actual biometric solutions.

1.4 Motivation for Heterogeneity Aware Aggregation

This research presents a heterogeneity conscious federated aggregation algorithm, which aims at overcome the shortcomings of the fixed or heuristic aggregation strategies. The aggregation process is designed to transition from an initial rule based weighting system to a learned aggregation process. In particular, the correlation between the client-specific traits, the metrics of the validation performance, and the global utility are modeled with the help of the polynomial regression. This causes aggregation weights to be acquired in an adaptive manner, enabling the empirical utility of every client to dictate its input to the global objective. Hence, client contribution is not viewed as consistently beneficial, rather, contributions are reweighted based on observed behavior from the past training rounds.

1.5 Summary of Empirical Findings

Substantial experimental testing on seven heterogeneous iris datasets shows consistent gains over standard FedAvg and other heuristic aggregation methods. The findings affirm that it is feasible to attain accurate and robust iris verification while maintaining strict privacy considerations. In addition, the proposed framework enables scalable and realistic cross-country biometric collaboration by explicitly considering client heterogeneity.

1.6 Problem Statement

A federated learning setup with K clients is considered. Each client i holds a local iris dataset $\mathcal{D}_i$ with identities that are not found in any other client. Because of privacy laws, institutional policies, and the fact that biometric data is very confidential, raw iris images and identity labels can not be shared between clients or dispatched to a central server. So, training a collaborative model must be done without direct access to centralized biometric data. The primary challenge

arises from severe client heterogeneity inherent to cross country iris verification. Participating clients differ substantially in dataset size, number of identities, image quality, sensor technology, and acquisition conditions. These elements result in non IID data distributions and make client updates inconsistent and unevenly contribute during federated optimization. In these cases, traditional aggregation methods that rely on fixed rules or the size of the dataset fail to work well.

This can lead to biased optimization and less favorable outcomes for clients that are not well represented or ones with lower quality data. The main problem addressed in this thesis is the design of a privacy preserving federated learning system for identity disjoint iris verification. While existing solutions rely on equal weight or sample size based aggregation, this framework maintains to be robust in the extreme case of highly non-IID data. This is achieved by dynamically weighting the contribution of clients based on their performance.

1.7 Research Objectives

The objectives of this thesis are summarized as follows:

i. Develop a privacy preserving federated iris verification framework using a SwinV2 Tiny Siamese backbone and metric learning objectives.

ii. Establish a realistic cross country, identity disjoint federated benchmark in which raw biometric data cannot be exchanged.

iii. Design an aggregation strategy that remains robust under extreme non IID distributions, dataset imbalance, and heterogeneous sensing conditions.

iv. Evaluate performance using verification metrics and security critical operating points to assess deployment relevant behavior.

1.8 Research Questions

This thesis is guided by the following research questions:

i. How does client heterogeneity (size, identity diversity, sensing conditions, and quality) affect federated iris verification under identity disjoint constraints?

ii. To what extent do standard aggregation strategies (e.g., FedAvg) degrade performance for under represented or low quality clients in cross country biometric FL?

1.9 Contributions

The main contributions of this work are summarized as follows:

i. Introduction of the CUST-Iris Pakistani benchmark: To make the federated benchmark more diverse in terms of geography and sensors, a new dataset of Pakistani iris imagery termed CUST-Iris is included. This dataset will be released to the public to help with other research.

ii. SwinV2 Tiny Siamese framework with hybrid metric learning: A privacy preserving pipeline is developed using a SwinV2 Tiny backbone integrated with specialized metric learning objectives. Stable collaborative optimization is attained via the integrated use of batch hard triplet and supervised contrastive losses, eliminating the need to share raw biometric data.

iii. FedHAT: learned heterogeneity aware aggregation: A novel aggregation strategy, which transitions from initial heuristic weighting to a learned polynomial regression method. This approach explicitly models client utility based on validation performance, dataset size, and identity richness.

Chapter 2

Literature Review

Recent developments in iris recognition and privacy preserving learning are due to the increased use of biometric systems in large, cross-institutional, and security sensitive applications. Recent studies have progressed in three key directions: the use of deep learning to extract iris features in a variety of acquisition settings, architectural and representational improvements to enhance robustness and security, and federated learning frameworks developed to allow decentralized training of biometric models as well as protecting data privacy. All these studies underscore the increasing need to have biometric verification systems.

The chapter is a review of the recent research published in 2022-2025. Advancements in cross domain iris recognition and privacy aware learning is focused on. In the sections that follow, the architectural decisions, training strategies and constraints of the proposed federated model, which takes into account the heterogeneity, will be discussed.

2.1 Deep Learning Architectures for Iris Recognition

The necessity to have reliable iris verification in different imaging conditions has triggered the purposeful exploration of deep learning models. Early works had focused largely on convolutional neural networks, but as research has continued more

recently, Transformer-based neural networks have increasingly gained popularity. Most of these earlier methodologies aimed at increasing resilience to segmentation errors and noise that were added when capturing the image.

Wei et al. [16] suggested a two stage model. It was a combination of parallelized Hough transform based iris localization and a lightweight CNN to extract features. Their findings revealed that they performed better verification in non-cooperative settings. This worked stressed that the correct localization of iris is a crucial need to achieve reliable recognition.

The strength of segmentation was also investigated by Jalal and Ghanim [17]. They combined deep convolutional learning of feature with SegNet based enhancement. It diminished the effects of imperfect isolations of the iris. Their research showed that the artifact of segmentation continues to be significant towards performance degradation, when the data is gathered with the help of different sensors and under different conditions.

Recent studies have focused on improving spatial stability and minimizing response to pupil enlargement, scale variations and off axis capture. Anti-aliased CNN architecture was proposed by Zambrano et al. [18]. It constrains the impacts of sampling distortions. Their solution had superior generalization to various acquisition geometries.

There has also been a growth of interest in security oriented feature representations. Lin and Chen [19] suggested multi scale dominant point descriptors and cryptosystem-based template protection. They found that privacy preserving representations are possible without a large compromised discriminative capability.

Other architectural orientations have also been ventured into. They consist of two stage pipelines, which explicitly isolate segmentation and recognition [20]. Linear discriminant analysis, coupled with deep neural networks, which are instant learning frameworks [21].

Style transfer based iris representations designed to improve privacy and domain adaptation [22] was also explored. Li et al. [23] investigated near infrared image

sequence of iris, and high throughput recognition. It highlighted the importance of temporal consistency in large scale biometric systems.

Despite such progress, the majority of deep learning based iris recognition approaches are centralized and assume that they have access to shared data. This supposition makes them less viable in privacy constrained and multi-institutional settings, where transfer of raw biometric data is not always allowed.

2.2 Transfer Learning and Pretrained Backbones

Labeled iris data is often limited. Training deep models from scratch is also expensive. For these reasons, transfer learning has become common in iris recognition research.

Lalitha et al. [24] evaluated several pretrained CNN backbones for iris feature extraction. They observed faster convergence during training. Generalization also improved. These gains were most noticeable under non cooperative imaging conditions. Their results indicate that features learned from large visual datasets provide a strong initialization for iris recognition.

Despite these benefits, most transfer learning studies rely on centralized training setups. Pretrained models are typically fine tuned using pooled data. This approach leaves some important questions unanswered. Data privacy continues to be a concern. Cross domain generalization is still limited. Institutional data governance is often ignored. These challenges become more severe in cross country biometric collaborations. In such settings, legal and ethical constraints prevent the sharing of raw biometric data.

2.3 Federated Learning for Biometrics

Federated learning has emerged as a useful way to support collaborative biometric model training. This is achieved while preserving data locality and privacy. Gupta

et al. [25] showed the feasibility of federated iris recognition. They trained CNN backbones across multiple institutions without sharing raw biometric data. This study established an important foundation for decentralized iris verification.

PM and Mahalakshmi [26] investigated privacy enhanced federated biometric strategies. They incorporated differential privacy mechanisms to reduce the risk of biometric information leakage. Their approach was evaluated in multimodal scenarios, where privacy concerns are especially critical. While improving privacy guarantees, such methods often introduce additional noise. It can reduce recognition accuracy.

Alternative communication strategies have also been proposed to reduce bandwidth usage and privacy exposure. Luo et al. [27] introduced FedIris, which exchanges iris templates instead of full model updates. While this method enhances communication efficiency, it relies on assumptions. They assume template security and limits adaptability across diverse client environments.

Sharma et al. [28] proposed a federated architecture for iris based authentication. They demonstrated that decentralized verification is feasible even under strict privacy constraints. Despite these developments, many federated biometric systems continue to rely on aggregation methods dominated by sample size. As a result, these methods are sensitive to extreme client heterogeneity, uneven identity distributions, and variations in data quality caused by sensor differences. The course of history of federated learning in biometrics is changing. The researchers have shifted out of the realm of simple feasibility to privacy preserving mechanisms. These developments enhance the locality of data and security of communication to a great extent. But, they also disclose profound architectural and methodological issues.

These limitations highlight the need for adaptive, performance-aware aggregation mechanisms that consider client utility beyond dataset size alone. A structured overview of recent advances in iris recognition and federated learning is presented in Table 2.1. It places the proposed framework within the widerresearch context and highlights its contributions.

Table 2.1: Comparison of related iris recognition and federated learning studies (2022–2025).

Study	Method Type	Architecture	Privacy or FL
Traditional and Deep Learning based Iris Recognition			
Wei et al. (2022)	DL iris verification	CNN with Hough transform	None
Jalal and Ghanim (2022)	Iris enhancement	SegNet and AlexNet	None
Zambrano et al. (2024)	Anti aliased CNN	ResNet AA	None
Lin and Chen (2024)	Template security	Multi scale features	Cryptosystem based
Abdulhasan et al. (2024)	LDA and DNN	DNN with LDA	None
Hsiao et al. (2024)	Two stage deep learning	CNN	None
Li et al. (2025)	High throughput iris recognition	CNN or Transformer	None
Transfer Learning Approaches			
Lalitha et al. (2024)	Transfer learning	Pretrained CNN backbones	None
Federated Learning for Biometrics			
Gupta et al. (2023)	Federated iris recognition	CNN	Federated learning
PM and Mahalakshmi (2024)	Privacy secure federated learning	CNN with differential privacy	Federated learning
Luo et al. (2022)	FedIris templates	CNN with template FL	Template sharing
Sharma et al. (2025)	FL for iris authentication	CNN	Federated learning
Style Transfer and Privacy Related Advances			
Wang et al. (2025)	Iris style transfer	Style Transformer	Privacy masking
This Work			
FedHAT (2026)	Federated metric learning with learned aggregation	SwinV2 Tiny Siamese	Federated Learning

2.4 Gap Analysis

Based on the reviewed literature, the following research gaps are identified:

i. Centralized Training Assumptions: Most iris recognition methods using deep learning are trained on large, centralized biometric datasets. While architectural improvements help systems handle noise, illumination changes, and segmentation errors, these methods struggle in real world situations. This is because different organizations can not share raw biometric data due to privacy, legal, and governance issues.

ii. Limited Exploration of Identity-Disjoint Settings: The current literature on biometric and federated learning research rarely uses per eyeball identity disjointed splits in clients. As a result, there are numerous performance increases reported without considering the issues which arise when being deployed in environments where no common identities exist. This gives overly optimistic evaluations that fail to consider the impact of non-IID data distributions.

iii. Inadequate Handling of Client Heterogeneity: Existing biometric federated learning models are largely based on either static or sample count aggregation strategy. These techniques implicitly assume that bigger datasets are more accurate in giving updates and fail to take into consideration the changes in sensing conditions, image quality and richness of identity that has a serious influence on verification performance.

iv. Bias Toward Dominant Clients: Clients with large and homogeneous data are generally prioritized by sample weighted aggregation schemes. This results in the under representation of smaller or poor quality datasets in the system. This results in an unfair treatment of such clients and leads to their under representation within the federated network.

v. Insufficient Modeling of Client Utility: Adaptive server side optimizers have been proposed to enhance convergence in the presence of heterogeneous updates; however, they fail to explicitly incorporate client utility regarding verification performance or dataset attributes. As a result, they are unable to distinguish between clients that are making a real difference in the global objective as opposed to those whose updates might not be as useful.

vi. Limited Focus on Security Critical Operating Points: Most of the current research focuses on average accuracy metrics and does not investigate performance at low FAR. This is important for biometric applications which require high confidentiality. This limits past research for actual-world scenarios.

vii. Lack of Performance Aware Aggregation in Biometric FL: There is a significant lack of federated learning methodologies that dynamically adjust client contributions according to empirical verification performance while adhering to strict privacy and identity separation guidelines, especially in multinational biometric environments.

These gaps directly motivate the problem formulation presented in Section 1.6 and inform the research objectives and questions defined in the subsequent sections.

Chapter 3

Identity Modeling, Datasets and Framework Selection

The effectiveness of any federated learning model is essentially dependent on the diversity, quality and organization of the underlying data. These variables in biometric systems are even more complex as they vary in sensor technology, locations where the data was gathered, and the regulations that the institutions adhere to. This chapter describes the wide-ranging experimental methodology that is utilized in this thesis, such as the datasets used, the definition of identification strategy that is used, as well as the preprocessing pipeline that is developed to normalize different iris samples before the federated training.

The overall aim of this framework is to design a realistic and challenging cross-country federated learning setting that represents the constraints that accompany real biometric deployments. Specific focus is laid upon strict identity segregation, sensor diversity and protection of privacy, thus ensuring that all experimental outcomes are a true reflection of the real world multi-institutional collaboration contexts.

This chapter outlines the approach to design a strong federated environment. It begins with the standardized method of collecting and processing biometric data in the next section.

3.1 Identity Modeling

A major problem with using iris recognition across different institutions is the lack of a standard way to label identities in different datasets. Subject-level labels often show inconsistencies and, in some cases, might not be directly comparable. This is because data collection is done independently. To address this issue and ensure clarity, this study used a standardized definition of identity. Instead of treating subjects as single entities, each eyeball, whether left or right, was considered a separate and unique identity. This definition is consistent with established biometric verification methods, allowing the model to learn features specific to each iris texture.

3.1.1 Rationale for Eyeball Level Identity Modeling

The definition of identity is crucial for evaluating performance in biometric verification. While subject level identity labels are often used in classification tasks, they can cause confusion in verification, especially when multiple biometric traits belong to one person. This thesis treats each eyeball, left and right, as a separate identity. This approach clearly separates biometric patterns, preventing information from one eye from affecting the other, which could artificially inflate verification accuracy. Moreover, from a security perspective, this definition aligns with how biometric systems work, where each enrolled biometric instance must be verified separately. Furthermore, eyeball level identity modeling enforces a more stringent evaluation protocol in federated learning. Since identities are disjoint across clients and splits, the model is required to generalize across unseen biometric patterns under heterogeneous sensing conditions, closely reflecting real world deployment scenarios.

3.2 Preprocessing Pipeline

The heterogeneity of datasets is a major problem with decentralized iris verification. As a result, standard preprocessing pipeline was created. This pipeline

converts raw iris images into a standard form that can be used in deep metric learning. It starts with accurate localization of the limbic and pupillary boundaries. The next step is normalization which maps the circular iris area to a fixed size rectangular template. Lastly, stabilization of the feature distribution is done through contrast enhancement and noise reduction. The following sections detail each stage of this workflow.

3.2.1 Segmentation and Masking

The initial preprocessing step necessitates pixel-wise iris segmentation. All raw images underwent processing via a U-Net based iris segmentation model, which had been trained on a varied iris dataset. This model generates a binary mask, thereby delineating the iris region from adjacent structures, including the pupil, sclera, eyelids, and eyelashes. Figure 3.1 presents U-Net model diagram.

Subsequent to segmentation, light morphological post-processing operations, encompassing noise removal and boundary smoothing, were implemented on the predicted masks. These operations serve to mitigate segmentation artifacts that could potentially introduce spurious features during the ensuing normalization and feature extraction phases.

3.2.2 Iris Normalization

Following segmentation, geometric normalization is performed to compensate for variations in pupil dilation, gaze direction, and capture distance. This work adopts a Daugman style rubber sheet normalization procedure, in which the circular iris region is unwrapped from Cartesian coordinates (x, y) into normalized polar coordinates (r, θ).

The resulting normalized iris representation is resized to a fixed resolution of 64×512. This representation ensures spatial alignment of iris texture patterns across all samples, regardless of the original image resolution or sensor characteristics, enabling consistent input to the SwinV2 Tiny backbone.

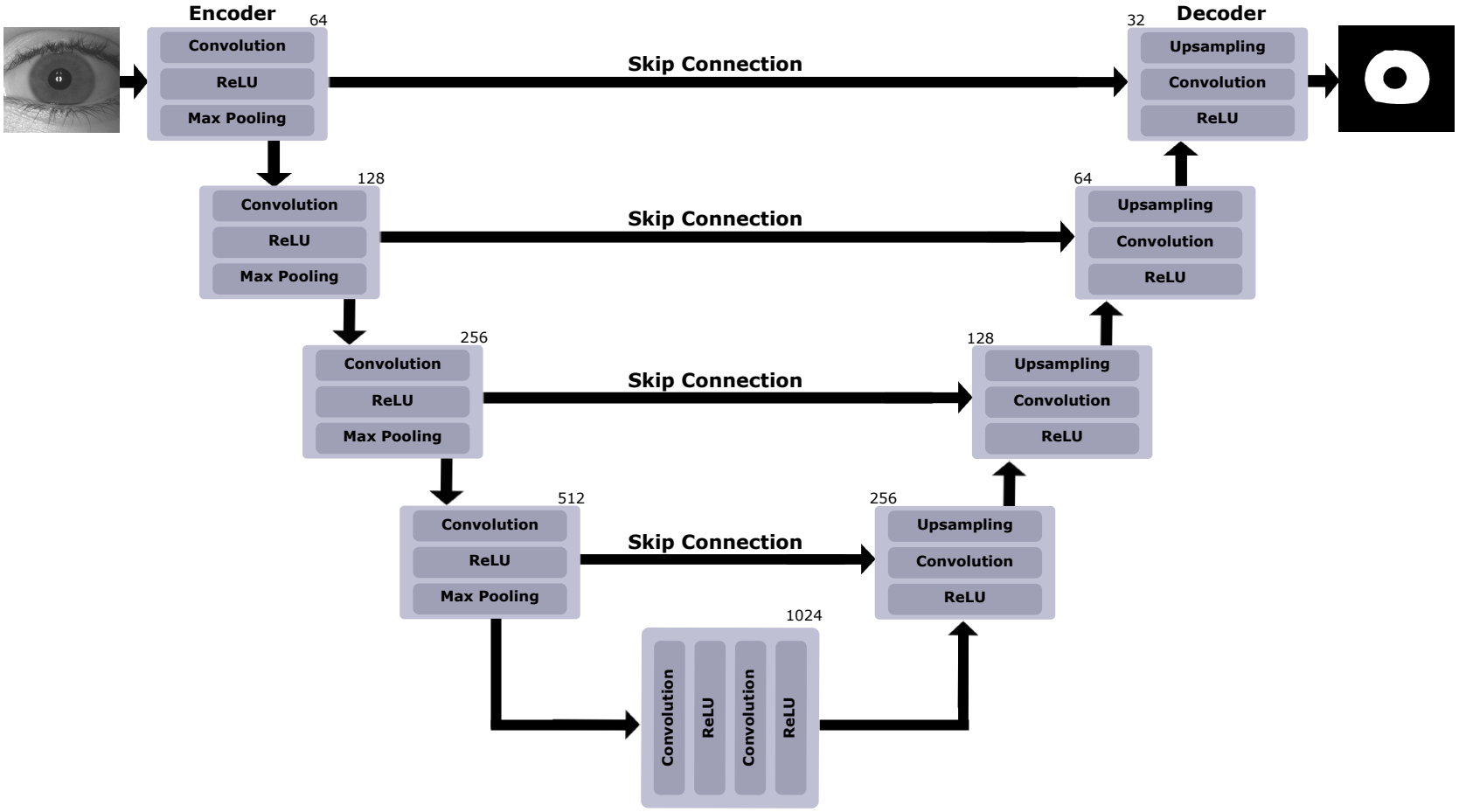

FIGURE 3.1: U-Net architecture used for iris segmentation in the preprocessing pipeline.

3.2.3 Image Enhancement

To mitigate inter dataset variability, mild, globally applied photometric enhancements were implemented. These enhancements were intended to improve visual consistency while preserving the fundamental biometric signature. The specific operations applied were Contrast Adjustment performed to normalize intensity distributions across datasets acquired in near-infrared and visible light, Brightness Correction applied to account for illumination variations stemming from differing acquisition environments, Sharpening utilized to enhance fine-grained textural details, especially in lower-resolution datasets like MMU.

Figures 3.2 and 3.3 show typical examples of the preprocessing pipeline that was used on all federated clients. A raw iris image, the corresponding binary segmentation mask, and the resulting mask overlay are shown for each dataset. The segmentation process consistently isolates the iris region while suppressing non-informative structures such as eyelids and sclera, even though the type of sensor, spatial resolution, lighting conditions, and imaging modality (near infrared versus visible light) can all be very different. This visual examination shows that the chosen segmentation method is strong and should be used as a standard preprocessing step before iris normalization and feature extraction.

3.2.4 Cross Dataset Normalized Iris Texture Comparison

Figures 3.4 and 3.5 present representative examples of normalized iris textures across all datasets used in this study, with three samples shown per dataset. Despite substantial variations in sensing modality, resolution, illumination conditions, and acquisition environments, the normalization process produces a consistent polar representation of iris texture across clients. Residual differences in texture sharpness, occlusion patterns, and noise characteristics reflect inherent sensor and population heterogeneity rather than preprocessing artifacts. This visual comparison highlights both the effectiveness of the normalization pipeline and the degree of cross-dataset variability that motivates the use of heterogeneity-aware aggregation in the proposed federated learning framework.

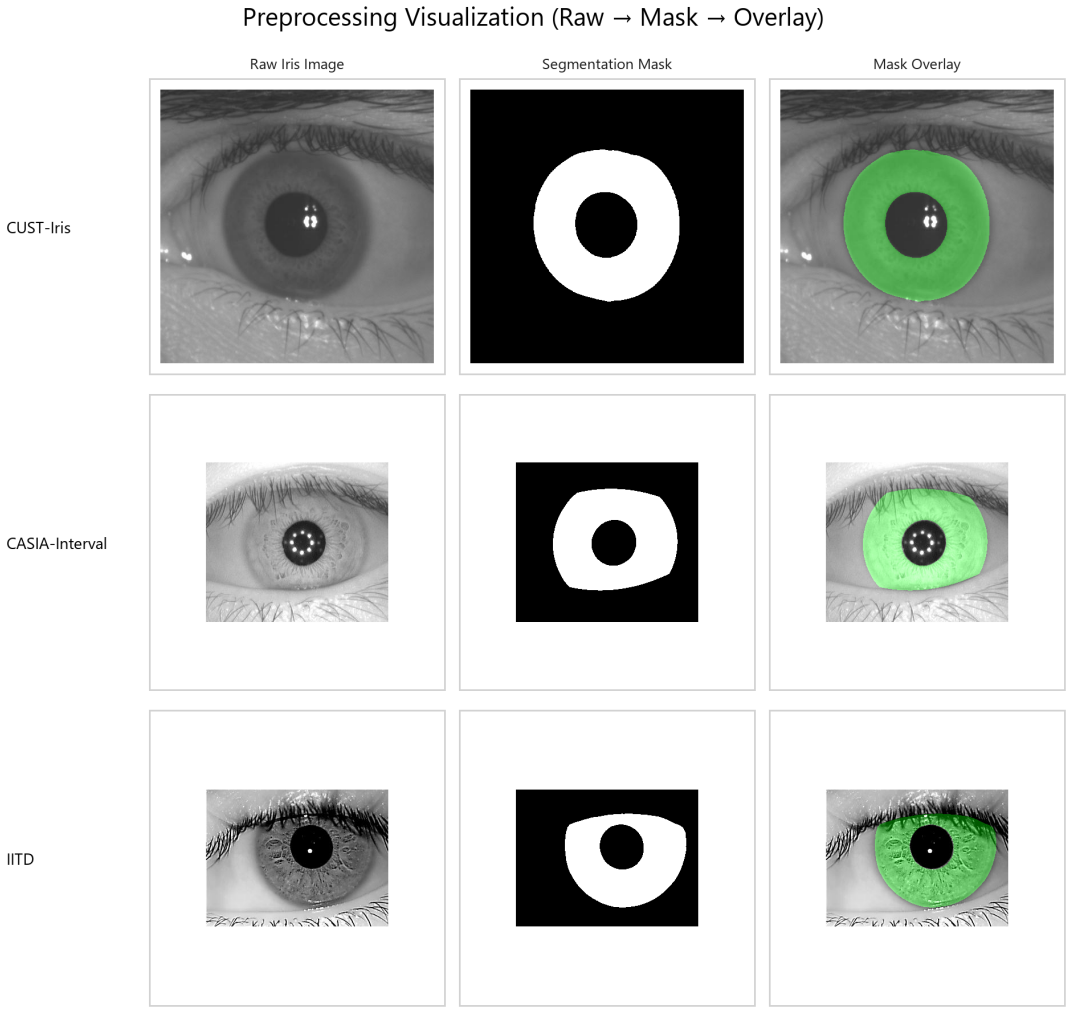

FIGURE 3.2: Visualization of the iris preprocessing pipeline for three representative datasets (CUST-Iris, CASIA-Interval, and IITD). For each dataset, the raw iris image, corresponding segmentation mask, and mask overlay are shown.

3.2.5 Design Considerations in Preprocessing

The preprocessing pipeline was designed with the objective of reducing inter dataset variability while preserving identity discriminative information. Several trade offs were considered during pipeline development.

First, aggressive normalization or heavy augmentation was deliberately avoided. While such techniques may improve robustness in classification tasks, they risk distorting fine grained iris texture patterns critical for verification. Second, all photometric enhancements were applied globally rather than adaptively to avoid introducing dataset specific artifacts that could bias federated training.

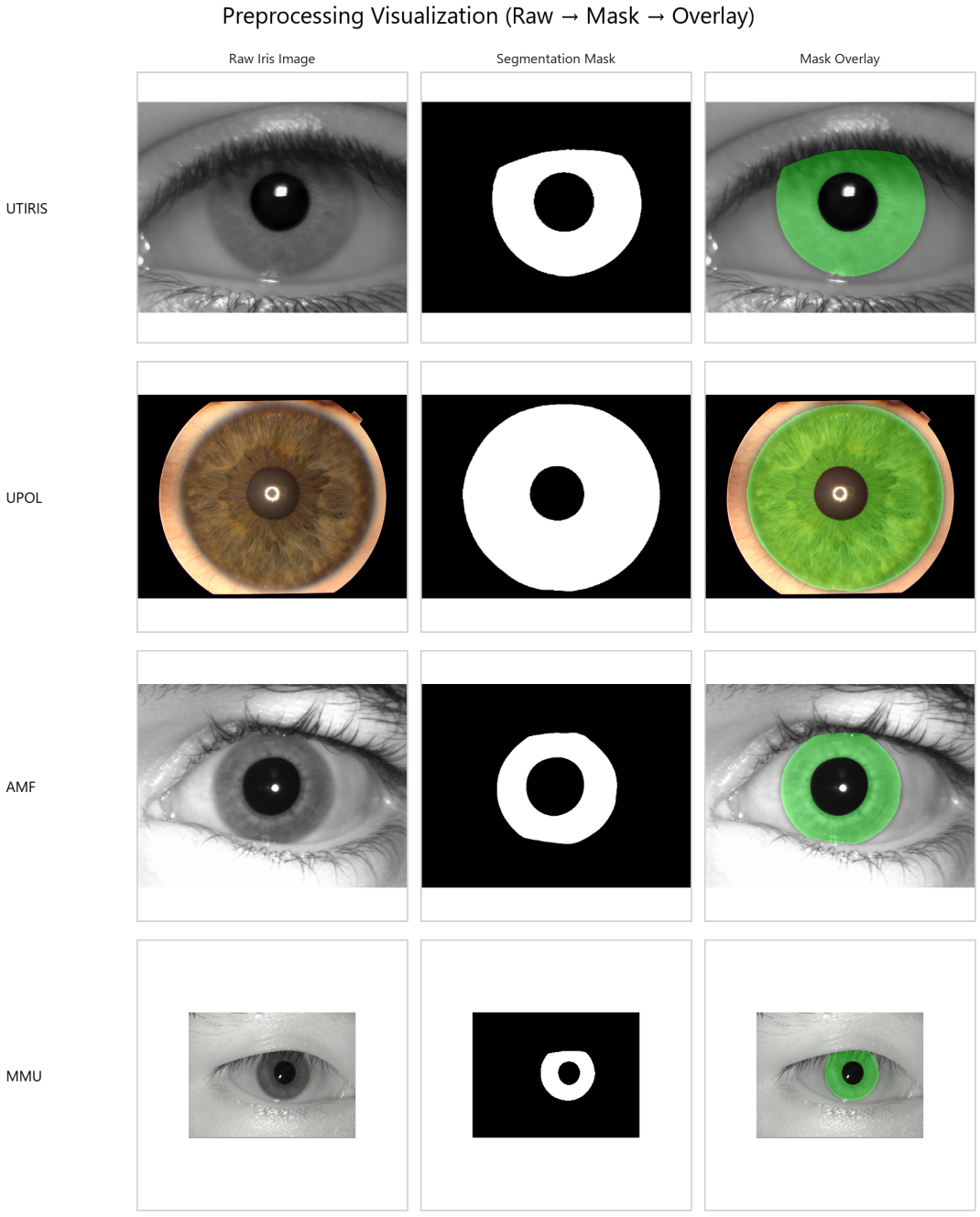

FIGURE 3.3: Visualization of the iris preprocessing pipeline for the remaining datasets (UTIRIS, UPOL, AMF, and MMU), illustrating robustness across visible light and near infrared acquisition conditions.

The choice of a fixed normalized resolution (64×512) balances spatial detail and computational efficiency. This resolution is sufficient to capture high frequency iris textures while remaining compatible with lightweight Transformer backbones such as SwinV2 Tiny. It facilitates faster convergence during the federated training

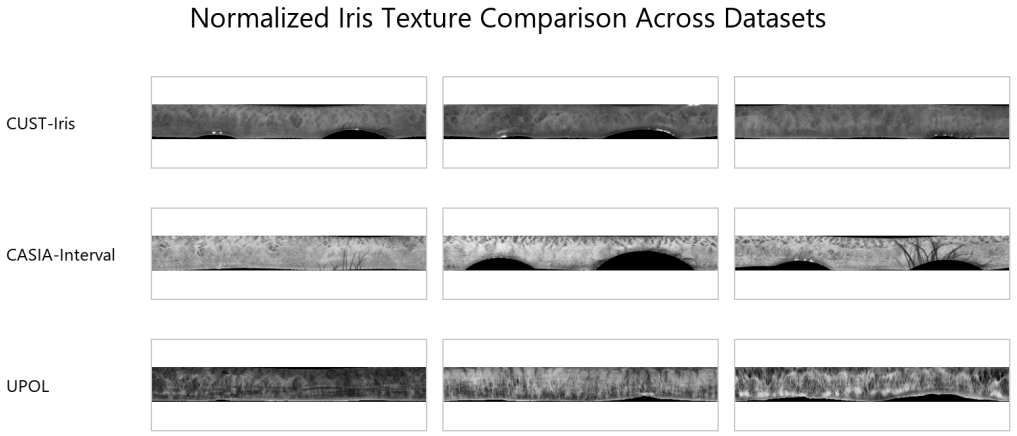

FIGURE 3.4: Normalized iris texture samples from CUST-Iris, CASIA-Interval, and UPOL datasets. Each row corresponds to one dataset, with three randomly selected normalized iris strips shown per dataset, illustrating cross-sensor texture characteristics after normalization.

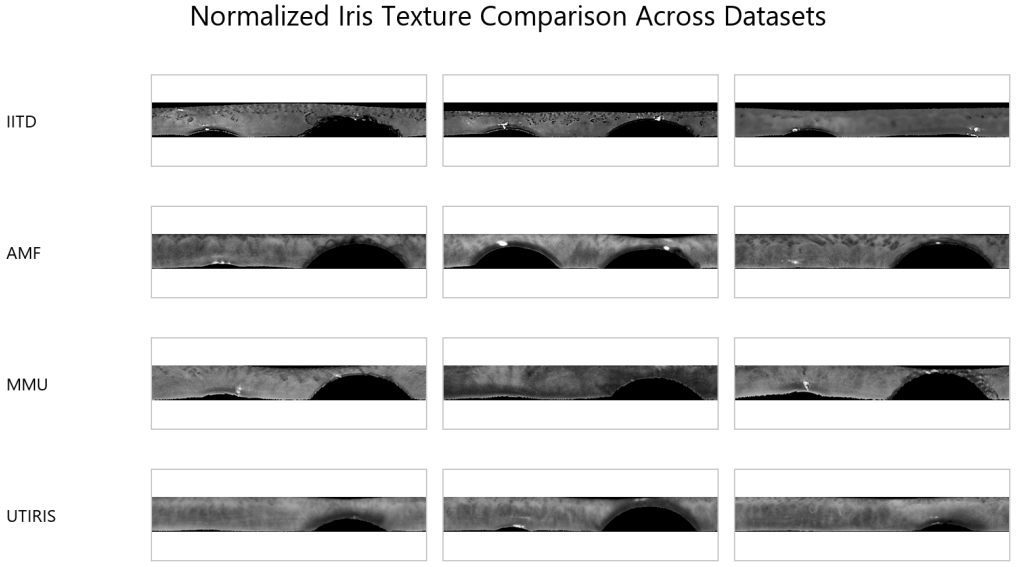

FIGURE 3.5: Normalized iris texture samples from IITD, AMF, MMU, and UTIRIS datasets. Despite differences in acquisition conditions and sensing modalities, the normalized representations exhibit consistent spatial alignment suitable for federated learning.

process by minimizing the input dimensionality for the SwinV2-Tiny Siamese network. It also ensures that the shifted window attention mechanism can efficiently handle fine grained iris images without a memory limit. In addition, this aspect ratio remains constant, thus stabilizing the embedding space over a wide range of datasets. This standardization becomes essential in attaining convergence in the

federated warm-up of the FedHAT framework.

3.2.6 Error Propagation and Robustness Considerations

In federated learning, the way preprocessing errors spread depends on the quality of the data and the characteristics of the sensors used. For example, segmentation errors might have a bigger impact on datasets with low resolution or those using visible light, which can lead to noisier embeddings and less useful contributions.

By standardizing preprocessing steps and avoiding overly aggressive transformations, the pipeline reduces the amplification of client specific errors. This design choice ensures that the performance differences seen during federated training mainly reflect the differences in the datasets, rather than being caused by problems with preprocessing.

3.3 Datasets

To establish a realistic cross country biometric collaboration, this study uses seven publicly available and proprietary iris datasets. These datasets were collected from different places, using different sensing technologies, and under different conditions. The datasets come from Pakistan (CUST-Iris), China (CASIA-Interval) [29], the Czech Republic (UPOL) [30–32], India (IITD) [33], Iraq (AMF), Iran (UTIRIS) [34], and Malaysia (MMU).

Each dataset exhibits unique characteristics in terms of illumination modality (near infrared versus visible light), image resolution, sensor quality, and subject cooperation. These variations introduce substantial statistical heterogeneity, making the experimental setting particularly challenging for federated learning algorithms that assume homogeneous client distributions.

In this thesis, each country level dataset is treated as an independent federated client. This design choice creates a setting that is clearly non-IID. It does not

assume anything about shared subject groups, sensor calibration, or how data is collected. This setup closely resembles real world biometric collaborations, where institutions have separate governance and can not share raw biometric data. The differences in these datasets allow for a thorough evaluation of the proposed Fed-HAT framework, considering various levels of client imbalance, data quality, and sensing differences.

3.3.1 Dataset Summary and Client Imbalance

A summary of identity counts, image counts, and train/validation/test splits for each dataset is provided in Table 3.1. A visual illustration of dataset scale and distribution across federated clients is shown in Fig. 3.6.

TABLE 3.1: Identity aware dataset statistics and train/validation/test splits. Left and right eyes are treated as separate identities.

Dataset	**IDs**	**Images**	**Train**	**Val / Test**	**Sensor / Acquisition**
CUST-Iris (PK)	720	2880	576	72 / 72	Crossmatch I-Scan 2, NIR, 480×480
CASIA-Interval (CN)	395	2639	317	39 / 39	CASIA V4, NIR 850 nm, Sony CCD, 320×280
UPOL (CZ)	128	384	104	12 / 12	Visible light, Nikon E5700, 2560×1920
IITD (IN)	425	2180	341	42 / 42	NIR 850 nm, JAI CV-M4+CL, 320×240
AMF (IQ)	108	540	88	10 / 10	NIR CCD, 640×480
MMU V1 (MY)	90	450	72	9 / 9	Visible light, Logitech web-cam, 320×240
UTIRIS (IR)	158	792	126	16 / 16	ISG Lightwise LW, 1000×776

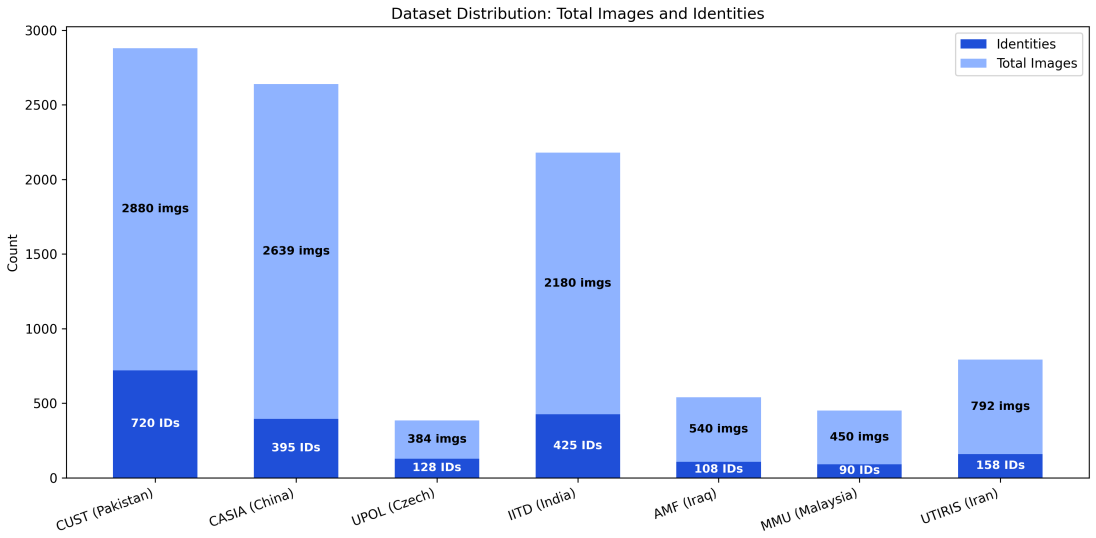

FIGURE 3.6: Distribution of total images and unique identities across all datasets.

3.3.2 Dataset Acquisition Characteristics and Challenges

Beyond dataset size and identity count, the conditions under which biometric samples are acquired play a critical role in determining the difficulty of federated learning. In real world deployments, institutions operate under independent acquisition protocols, hardware constraints, and subject compliance levels. These factors introduce structural heterogeneity that cannot be fully eliminated through preprocessing alone.

The datasets used in this study reflect a wide range of acquisition scenarios. Some datasets, such as CASIA Interval and IITD, were collected in controlled laboratory environments using near infrared sensors with fixed subject to camera distance and cooperative subjects. In contrast, datasets such as UPOL and MMU were captured under visible light conditions, introducing sensitivity to ambient illumination, reflections, and motion blur. The AMF and UTIRIS datasets further exhibit variability in image resolution and sensor quality, reflecting practical constraints often encountered in resource limited environments. These differences result in distinct data distributions across clients, affecting texture contrast, noise characteristics, and geometric consistency. As a result, models trained under centralized assumptions or uniform aggregation strategies are prone to bias toward dominant

acquisition conditions. This motivates the need for federated learning strategies that explicitly account for heterogeneity at both the dataset and acquisition levels.

3.3.3 Dataset Specific Characteristics

All the datasets used in this thesis present their own challenges that together make the federated learning environment highly heterogeneous. Such differences are based on the sensing hardware, acquisition protocols, and environmental differences between various institutions. An example would be the use of high-resolution near-infrared sensors by some clients and mobile or less-controlled capture systems by others.

Moreover, the richness of identity, as well as the sample distribution, is very different in each dataset. These inconsistencies result in significant changes in features which may cause instability of standard world models. With such a wide set of data sources, the study guarantees that the proposed framework is validated under activities of cross country deployment. To create a really robust and scalable iris verification system, it is necessary to address these multi-faceted imbalances.

3.3.3.1 CUST-Iris (Pakistan)

The CUST Iris dataset represents a near infrared acquisition environment using a dedicated commercial iris scanner. While the dataset benefits from consistent sensor characteristics and controlled capture conditions, it exhibits a relatively large number of identities compared to several other clients. This scale difference introduces imbalance during federated aggregation, making it a dominant contributor under sample count based weighting schemes.

3.3.3.2 CASIA-Interval (China)

CASIA-Interval is a widely used benchmark dataset collected under controlled laboratory conditions. Although the data quality is high, the dataset represents a

single acquisition modality and a specific demographic distribution. Its inclusion in this study provides a stable reference client against which the robustness of federated aggregation strategies can be evaluated.

3.3.3.3 UPOL (Czech Republic)

The UPOL dataset is captured under visible light conditions using a consumer grade camera. Compared to near infrared datasets, UPOL exhibits greater sensitivity to illumination changes, reflections, and occlusions. The relatively small dataset size and distinct sensing modality make UPOL a challenging client in federated learning, particularly under aggregation strategies that favor larger datasets.

3.3.3.4 IITD (India)

The IITD dataset was collected using a near infrared sensor but includes moderate variation in gaze direction and capture distance. This dataset represents an intermediate level of difficulty, combining controlled sensing with realistic variability in subject presentation.

3.3.3.5 AMF (Iraq)

AMF is characterized by a smaller number of identities and lower image resolution relative to other datasets. Such constraints reflect deployment scenarios where specialized biometric hardware may be unavailable. The dataset poses challenges related to limited data diversity and increased susceptibility to noise.

3.3.3.6 MMU (Malaysia)

MMU is a visible light dataset acquired using a standard webcam. The combination of low resolution, uncontrolled illumination, and limited identity count makes MMU one of the most challenging clients in this study. Performance on

this dataset is highly sensitive to aggregation bias and serves as a strong indicator of fairness in federated learning.

3.3.3.7 UTIRIS (Iran)

UTIRIS contains iris images captured using mixed acquisition conditions and higher image resolutions. The dataset introduces variability in both sensing modality and image scale, requiring robust normalization and aggregation strategies to ensure consistent performance.

3.4 Backbone Architecture Selection Rationale

Before conducting federated learning experiments, an initial centralized evaluation was performed to determine a suitable backbone architecture for iris verification under identity aware splitting. The objective of this experiment was not to optimize absolute verification performance, but rather to select a model that provides strong and consistent discriminative capability across diverse datasets while remaining computationally efficient and stable for federated training.

Several commonly used architectures were evaluated under identical training and evaluation conditions, including MobileNetV3, ResNet18, ViT Tiny, and SwinV2 Tiny. All models were trained in a centralized setting using the same preprocessing pipeline, metric learning objectives, and eye aware identity split protocol described earlier in this chapter. Each model was evaluated independently on the corresponding dataset it was trained on, ensuring a fair and controlled comparison focused solely on architectural capacity.

3.4.1 MobileNetV3

Across all evaluated datasets, MobileNetV3 exhibited limited discriminative power, particularly for small and heterogeneous datasets. While acceptable performance

was observed on some larger datasets, verification accuracy degraded sharply under challenging sensing conditions, as reflected by high Equal Error Rates and low True Accept Rates at strict operating points. This behavior suggests that lightweight convolutional architectures may lack sufficient representational capacity for fine grained iris texture modeling under heterogeneous conditions.

3.4.2 ResNet18

ResNet18 demonstrated a substantial improvement over MobileNetV3 and achieved strong performance on several datasets, particularly those captured under near infrared conditions. However, performance consistency varied across clients, with noticeable degradation on visible light datasets and smaller identity pools. This sensitivity indicates that fixed receptive field convolutional architectures may struggle to generalize uniformly across sensing modalities.

3.4.3 ViT Tiny Transformer

Transformer based architectures exhibited superior robustness. ViT Tiny achieved high verification accuracy across most datasets, particularly on larger clients. Nevertheless, its performance showed increased variability on smaller and noisier datasets, suggesting sensitivity to limited data regimes and reduced local context modeling.

3.4.4 SwinV2 Tiny Transformer

Among all evaluated architectures, SwinV2 Tiny consistently achieved the most balanced performance across datasets. It yielded low Equal Error Rates and high True Accept Rates at both moderate and security critical operating points, while maintaining stable behavior across near infrared and visible light conditions. The hierarchical window based attention mechanism of SwinV2 Tiny enables effective

local texture modeling while preserving global contextual awareness, making it well suited for iris verification tasks.

Based on this centralized evaluation, SwinV2 Tiny was selected as the backbone architecture for all subsequent federated learning experiments. This choice ensures that observed performance differences in the federated setting can be attributed primarily to aggregation strategy and data heterogeneity, rather than architectural limitations.

3.5 Rationale for Siamese Metric Learning

The backbone is arranged within a Siamese framework to meet the core demands of cross country iris verification. It directly reflects how the problem is defined in real world deployments. The design is guided by the following key considerations:

3.5.1 Verification as a Comparison Task

Iris recognition in high security environment is basically a comparison problem, rather than a classification problem. The idea is to establish whether two samples are of the same identity or not. A Siamese architecture is well suited for this purpose. It learns a function that maps raw iris textures to an embedding space. Similarity is reflected in this space in terms of distance. Decisions are thus made on relative proximity and not on the set class names.

3.5.2 Handling Identity-Disjoint Constraints

The key need of this work is the generalization to the unknown and unseen identities. The test subjects are totally dissociated with those which were used in model training. This constraint excludes the use of identity memorization. The model should be able to represent universal iris features. It must learn what makes two

samples similar or different at a structural level. The Siamese setup naturally implements this behavior by focusing on pairwise relationships.

3.5.3 Generalization under Heterogeneity

The data is from various countries and acquisition arrangements. The differences in sensors, lighting and imaging conditions are inevitable. The Siamese structure involves the sharing of weights between its branches. This imposes a uniform process of extracting features. Consequently, the acquired representation is constant across domains. This uniformity is essential for federated learning. The global model should be able to work consistently irrespective of heterogeneous distributions of clients.

By following these principles, the framework does not follow the traditional classification. It creates a comparison structure. This makes the system more scalable. It is also consistent with the need for privacy preserving, multi institutional, biometric collaboration.

Chapter 4

Client Side Training Pipeline

This chapter describes the client side training pipeline. It is the elementary unit of learning of every federating member since each member is storing sensitive biometric information. client side is the local computing environment in each of the participating institutions where sensitive biometric data is stored. The framework guarantees that raw iris images and the associated normalized polar strips do not leave the safe environment of institution. This is required to comply with privacy and data management measures. This study employs a metric learning model based on a Siamese architecture. This design, as it was established in the previous chapter, allows the model to map iris textures to a discriminative embedding space. The identity is verified by calculating the distances between vectors instead of class assignments. As shown in Fig. 4.1, all clients have the same architecture and optimization protocols. It makes sure that the local updates produced by each such institution despite originating from heterogeneous sensors and populations are mathematically compatible. Such compatibility allows steady global convergence in cases where such updates are ultimately aggregated by the central server.

To connect these local updates to the broader federated system, the pipeline involves certain validation procedures to know the reliability of the client prior to any communication occurring. The local progress of each client is evaluated using a subset of identities that are not revealed to the training phase.

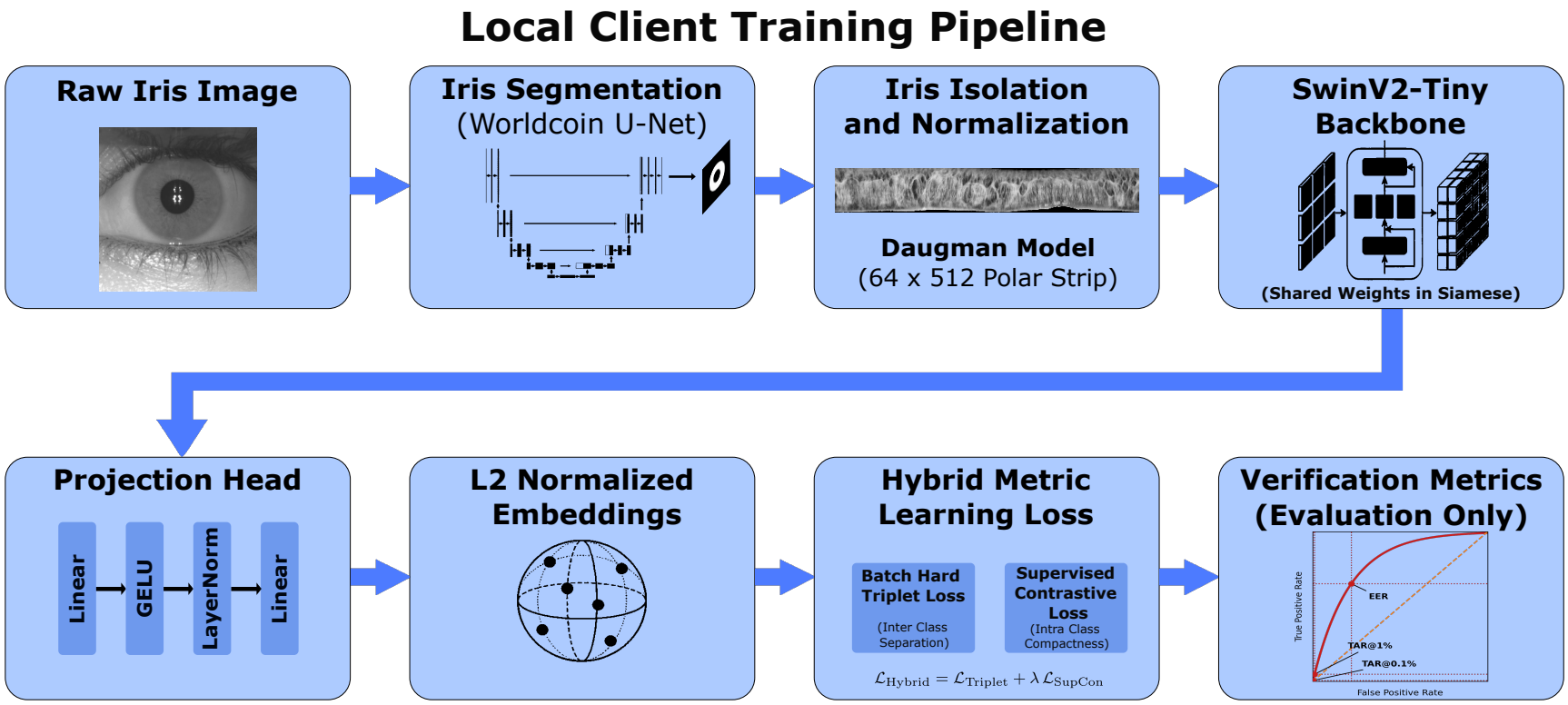

FIGURE 4.1: Local training and evaluation pipeline executed independently at each federated client using a SwinV2 Tiny Siamese network and metric learning objectives.

4.1 SwinV2 Tiny Siamese Backbone

The main feature extractor in this system is the SwinV2 Tiny backbone. Even though traditional CNNs have been extensively used in the extraction of biometric features, SwinV2 Transformer architecture is better at modeling long range spatial dependencies, a factor that is of importance in iris texture analysis. The backbone processes 64× 512 normalized iris strips. Due to the hierarchical nature of the Swin Transformer, the model can analyze iris texture at various scales.

One important element is the Shifted Window based Self Attention (S-WMSA) which limits the calculation of attention to local windows but can connect cross windows through cyclic shifting. This mechanism works well with the polar iris representation, in which the periodic character of the textural patterns along the horizontal (angular) axis are captured without sacrificing the local spatial resolution.

The last step of the backbone develops a high dimensional representation. This pictorial expression captures the finer aspects of the iris, including its crypts and furrows, and the overall structure of the iris.

4.2 Siamese Embedding Framework and Projection Head

The local model follows a Siamese formulation designed to learn an embedding function $f(\cdot;\theta)$. Given an input iris image x, which represents a 64×512 normalized polar strip, the backbone maps it to a d-dimensional feature vector:

$$z = f(x;\theta), \tag{4.1}$$

where θ denotes the learnable parameters of the network.

4.2.1 Projection Head Architectural Components

To refine the high dimensional features extracted by the SwinV2 Tiny backbone for biometric verification, the output is processed by a specialized Projection Head. This component acts as a bridge, connecting the general representation of the transformer with the discriminative embedding space needed for metric learning. The projection head consists of the following architectural components:

4.2.1.1 Initial Linear Transformation Layer

This layer performs a learned linear projection of the 768-dimensional (768-D) feature vector extracted from the final stage of the SwinV2 Tiny backbone. While the dimensionality remains consistent at 768-D in this stage, the main purpose is to transform the feature space into a representation optimized for the subsequent nonlinear operations.

4.2.1.2 Gaussian Error Linear Unit Activation

This component applies the GELU non linear activation function to each element of the projected feature vector. Unlike standard ReLU, GELU weights inputs by their percentile rather than a simple sign based gate, providing a smoother gradient flow that is particularly beneficial for transformer based architectures.

In the context of iris recognition, this non linearity allows the model to learn complex, non linear relationships between the iris patterns, which is essential for distinguishing between highly similar identities in a discriminative embedding space.

4.2.1.3 Layer Normalization

Layer Normalization is then applied to the features of each sample. This step is crucial for stabilizing the distribution of activations throughout the network. By keeping the mean and variance of the features consistent, it prevents internal

co-variate shift, which significantly speeds up training convergence. This stability is especially important in federated learning, as it helps to align updates from different clients, each with varying image qualities and sensor characteristics.

4.2.1.4 Final Linear Projection to 256-D Space

A final linear layer projects the 768-D normalized feature representation into a target 256-dimensional (256-D) embedding space. This layer performs the critical task of dimensionality reduction, acting as a bottleneck that compresses the rich transformer features into a compact biometric template.

The 256-dimensional space is chosen as an optimal balance between template compactness, which is necessary for efficient storage and communication, and enough capacity to represent the unique textural signature of the iris. This final vector is then L_2-normalized to reside on a unit hypersphere, ensuring that cosine similarity can be used as an effective and robust verification metric.

The resulting embedding is then L_2-normalized to project it onto a unit hypersphere:

$$\tilde{z} = \frac{W_2\text{LN}(\text{GELU}(W_1 f + b_1)) + b_2}{|W_2\text{LN}(\text{GELU}(W_1 f + b_1)) + b_2|_2} \tag{4.2}$$

The components of this equation are defined as follows:

i. f: The 768-dimensional feature vector produced by the final stage of the SwinV2-Tiny backbone.

ii. W_1, b_1: The weights and biases of the initial linear layer, where $W_1 \in \mathbb{R}^{768\times768}$, used to refine the feature space.

iii. GELU: The Gaussian Error Linear Unit activation function, which introduces non-linearity to capture complex iris patterns.

iv. LN: Layer Normalization, used to stabilize the distribution of activations across heterogeneous client updates.

v. W_2, b_2: The parameters of the final projection layer that map the refined feature space into the target 256-dimensional biometric template.

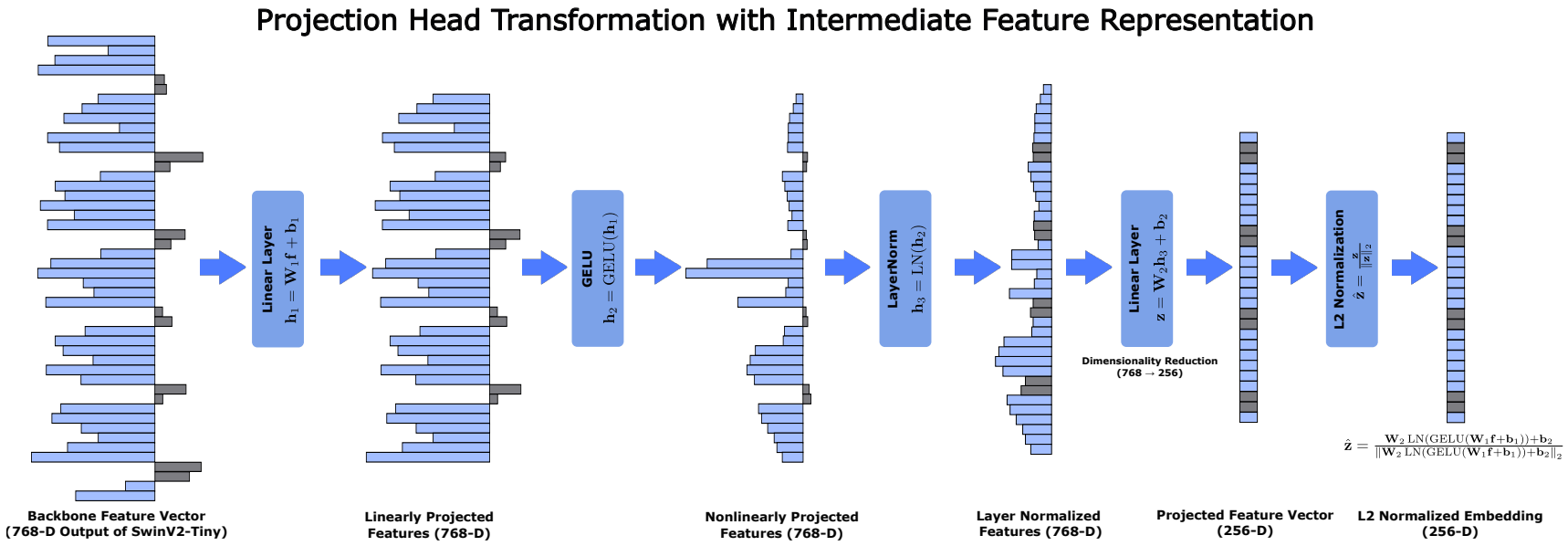

FIGURE 4.2: Structure of the projection head transforming the backbone output into the final embedding.

Normalization is critical because it ensures that the verification score depends solely on the angular relationship (cosine similarity) between vectors. The similarity s between two samples x_i and x_j is computed as:

$$s(x_i, x_j) = \tilde{z}_i^\top \tilde{z}_j \tag{4.3}$$

where $\tilde{z}_i$ and $\tilde{z}_j$ are the 256-D normalized templates.

4.3 Metric Learning Objectives

The model is trained to minimize a hybrid loss function that combines local distance constraints with global clustering objectives.

4.3.1 PK-Sampling and Batch Construction

To ensure the presence of informative gradients, each client utilizes a $P \times K$ sampling strategy. For every mini batch, P unique identities are randomly selected, and K images per identity are sampled. This ensures that every batch contains $\binom{K}{2} \times P$ positive pairs and $PK(PK - K)$ negative pairs, providing the necessary supervision for triplet and contrastive objectives.

4.3.2 Batch Hard Triplet Loss

The triplet loss enforces a margin between genuine and impostor distances. To accelerate convergence and avoid the "vanishing gradient" problem common in random triplet selection, this work employs a "Batch Hard" strategy. For each anchor i, the loss identifies the most difficult positive and negative samples within the batch:

$$d_i^+ = \max_{j:y_j=y_i} \|\tilde{z}_i - \tilde{z}_j\|_2^2, \quad d_i^- = \min_{j:y_j \neq y_i} \|\tilde{z}_i - \tilde{z}_j\|_2^2. \tag{4.4}$$

The loss is then defined as:

$$\mathcal{L}_{\text{triplet}} = \frac{1}{B}\sum_{i=1}^{B}\max(0, d_i^+ - d_i^- + m), \tag{4.5}$$

where the margin $m = 0.3$ prevents the model from collapsing all identities into a single point.

4.3.3 Supervised Contrastive Loss

While triplet loss focuses on the relative distance between three samples, the Supervised Contrastive (SupCon) loss considers all positive samples in a batch simultaneously, pulling them toward the anchor while pushing all negatives away. The scaled similarity logits are defined with a temperature $\tau = 0.07$:

$$\mathcal{L}_{\text{sup}} = -\frac{1}{B}\sum_{i=1}^{B}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(\tilde{z}_i^\top \tilde{z}_p/\tau)}{\sum_{j\neq i}\exp(\tilde{z}_i^\top \tilde{z}_j/\tau)}. \tag{4.6}$$

The final local training objective is defined as a weighted combination of the two loss terms, as illustrated in Fig. 4.3:

$$\mathcal{L} = \mathcal{L}_{\text{triplet}} + \lambda\,\mathcal{L}_{\text{sup}}, \tag{4.7}$$

where $\lambda = 0.5$ balances inter class separation and intra class compactness.

4.4 Optimization and Evaluation Protocol

The local optimization process utilizes the AdamW optimizer with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. To maintain training stability and prevent divergence amidst the non IID data distributions inherent in federated iris recognition, gradient clipping is strictly enforced at a threshold of 5.0. Each client independently executes this optimization procedure on its local dataset for a fixed number of epochs per communication round.

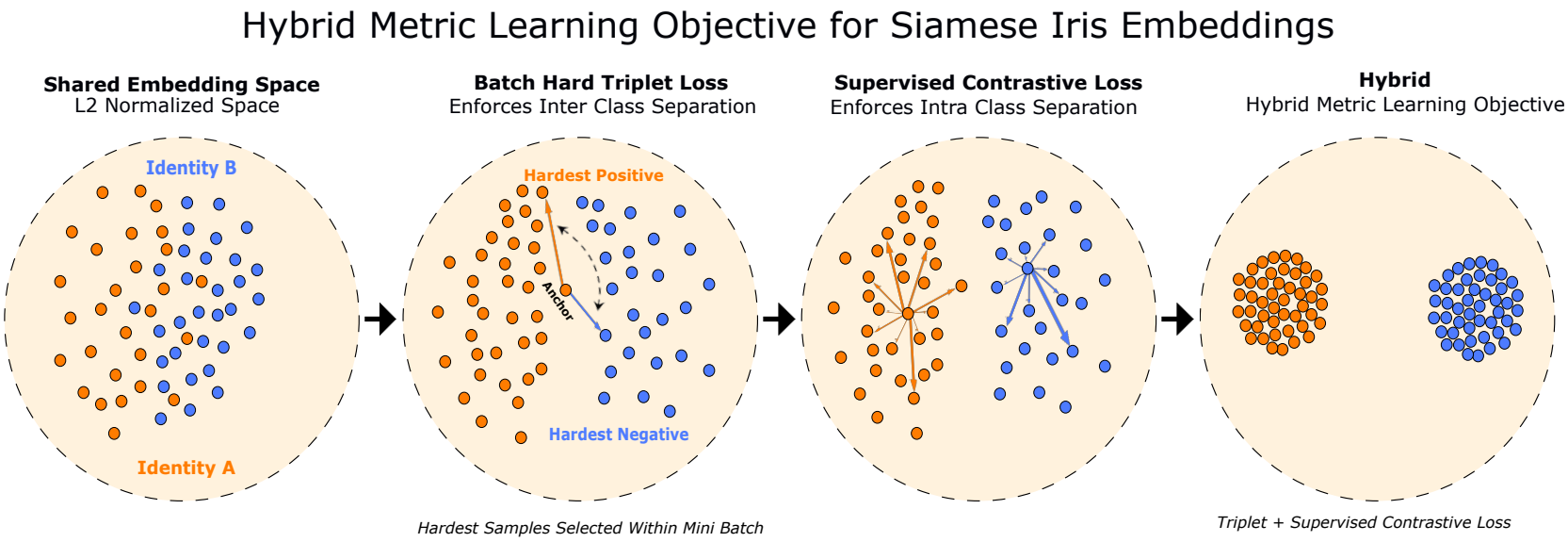

FIGURE 4.3: Hybrid metric learning objective used for local Siamese training.

Upon completion of the training rounds, the global model undergoes a rigorous evaluation protocol on the identity disjoint test split of each client. The verification performance is quantified using three primary biometric metrics. First, the ROC-AUC is computed to assess the overall discriminative capability and the degree of separation between authentic and impostor pairings across all potential operating thresholds. Subsequently, EER is determined as the specific operating point where FAR and FRR are equivalent; lower values signify enhanced verification accuracy. Ultimately, TAR at fixed False Accept Rates is presented at the 1% and 0.1% levels. This concluding metric is of particular importance, as it reflects the dependability and performance under the high security of the system. It demands operational conditions essential for real world biometric applications.

Chapter 5

Federated Learning Strategy

This chapter presents the federated learning formulation adopted in this thesis and details the proposed aggregation mechanism designed to operate under cross country biometric heterogeneity. The focus is on enabling collaborative training across institutions while preserving data privacy, identity separation, and robustness to dataset imbalance.

5.1 Federated Problem Formulation

Consider a federated learning environment consisting of K participating clients, where each client corresponds to a country level iris dataset. The k-th client, denoted by $\mathcal{C}_k$, maintains exclusive access to its local dataset

$$D_k = \{(x_i, y_i)\}_{i=1}^{N_k}, \tag{5.1}$$

where x_i represents a normalized iris image and y_i denotes its associated identity label.

Due to differences in data acquisition pipelines, sensing modalities, and collection protocols, the participating datasets vary significantly in both scale and composition. As detailed earlier in Table 3.1, the number of identities q_k, total samples N_k,

and image characteristics differ substantially across clients. Moreover, identities are strictly disjoint between clients, resulting in a federated learning problem that is simultaneously non IID, identity separated, and highly imbalanced.

A central requirement in biometric systems is the protection of sensitive identity bearing data. Direct exchange of raw iris images or identity labels is therefore prohibited. In the adopted federated setting, only model parameters are communicated between clients and the central server, enabling collaborative optimization while ensuring that all biometric data remain confined to their originating institutions.

5.2 Conventional Aggregation via Federated Averaging

Let θ_t denote the global model parameters at communication round t. At each round, the server distributes θ_t to all clients, which then perform local optimization for E epochs using their respective datasets. Each client returns an updated parameter set $\theta_{t+1}^{(k)}$.

In the classical Federated Averaging (FedAvg) approach, client contributions are weighted proportionally to dataset size, yielding the aggregation rule

$$\theta_{t+1}^{\text{global}} = \sum_{k=1}^{K} \left[\frac{N_k}{\sum_{j=1}^{K} N_j} \cdot \theta_{t+1}^{(k),\text{local}} \right] \tag{5.2}$$

The components of this equation are defined as follows:

i. $\theta_{t+1}^{\text{global}}$: The updated global model parameters for the subsequent communication round.

ii. K: The total number of participating federated clients (geographical institutions).

iii. N_k: The number of local training samples (iris images) held by the k-th institution.

iv. $\sum_{j=1}^{K} N_j$: The total dataset size across the entire federated network.

v. $\frac{N_k}{\sum N_j}$: The aggregation weight for client k, which is strictly proportional to its relative dataset size.

vi. $\theta_{t+1}^{(k),\text{local}}$: The parameters of the model trained locally on the k-th client's private biometric data.

While FedAvg is computationally efficient and widely adopted, its reliance on sample count based weighting introduces bias in identity disjoint biometric settings. Clients with large datasets exert disproportionate influence over the global model, whereas smaller datasets, often representing underrepresented populations, contribute marginally to the optimization process. This imbalance motivates the development of aggregation strategies that account for more than dataset size alone.

5.3 Heterogeneity Aware Aggregation Design

To mitigate the limitations of uniform or size based aggregation, this thesis introduces a heterogeneity aware federated training strategy. The key idea is to modulate client influence based on observed training behavior and validation performance rather than relying solely on static dataset statistics.

The proposed approach operates in two stages. An initial stabilization phase ensures reliable optimization during early training and gathers supervision signals. This is followed by an adaptive aggregation phase in which client weights are inferred using a learned regression model. An overview of the full training pipeline is illustrated in Figure 5.1.

5.3.1 Early Stage Stabilization

During the first T_{warm} communication rounds, aggregation weights are computed using a deterministic scoring function designed to balance dataset scale, identity diversity, and empirical validation quality.

For each client $\mathcal{C}_k$, the following quantities are considered:

i. N_k: number of local training samples,

ii. q_k: number of unique iris identities,

iii. ϕ_k: validation reliability score, defined as $\phi_k = 1 - \text{EER}_k$.

A composite warm up score is computed as

$$W_k^{\text{warm}} = \alpha \log(1 + N_k) + \beta \frac{q_k}{N_k} + \gamma \phi_k + \delta \frac{1}{\sqrt{q_k}}, \tag{5.3}$$

where α, β, γ, and δ regulate the influence of dataset scale, identity density, validation behavior, and rarity compensation, respectively.

These scores are normalized and used to aggregate client updates:

$$\theta_{t+1}^{\text{global}} = \sum_{k=1}^{K} \left[\left(\frac{W_k^{\text{warm}}}{\sum_{j=1}^{K} W_j^{\text{warm}}} \right) \cdot \theta_{t+1}^{(k),\text{local}} \right] \tag{5.4}$$

The components of this equation are defined as follows:

i. $\theta_{t+1}^{\text{global}}$: The updated global model parameters broadcast by the central server for the next communication round.

ii. K: The total number of participating federated clients.

iii. W_k^{warm}: The composite stabilization score for client k, which balances dataset scale, identity richness, and empirical validation quality.

iv. $\sum_{j=1}^{K} W_j^{\text{warm}}$: The normalization factor representing the sum of all client scores.

v. $\theta_{t+1}^{(k),\text{local}}$: The parameters of the model trained locally on the k-th client's private biometric data.

Normalization is mathematically necessary to maintain the stability of the global parameter space and prevents the model from being biased toward the absolute scale of any single heuristic metric. Beyond stabilizing early training, this phase serves an additional purpose: collecting behavioral data required to learn an adaptive aggregation function.The warm up aggregation weight defined in Eq. (5.3) is employed as an initial, rule based personalization mechanism. Its design is inspired by prior work such as FedFomo [35], which shows that effective aggregation should be informed by a client's observed impact on validation performance rather than dataset size alone. In heterogeneous learning environments, this perspective encourages weighting schemes that capture practical contribution to global optimization. In addition, the proposed formulation aligns with fairness aware aggregation principles in federated learning, which argue for non uniform client weighting to counteract the disproportionate influence of large or homogeneous clients under non IID data distributions [36]. By combining dataset statistics with validation reliability, the warm up strategy provides a stable and equitable initialization for subsequent learned aggregation.

5.3.2 Supervision Signal Construction

At each warm up round t and for each client k, validation performance is measured both before and after local training. Let $\text{EER}_{k,\text{global}}^{(t)}$ denote the validation error of the global model and $\text{EER}_{k,\text{local}}^{(t)}$ the validation error of the locally updated model.

A non negative utility target is defined as

$$y_k^{(t)} = \max\left(0,\ \text{EER}_{k,\text{global}}^{(t)} - \text{EER}_{k,\text{local}}^{(t)}\right), \tag{5.5}$$

capturing the extent to which a client update improves validation performance.

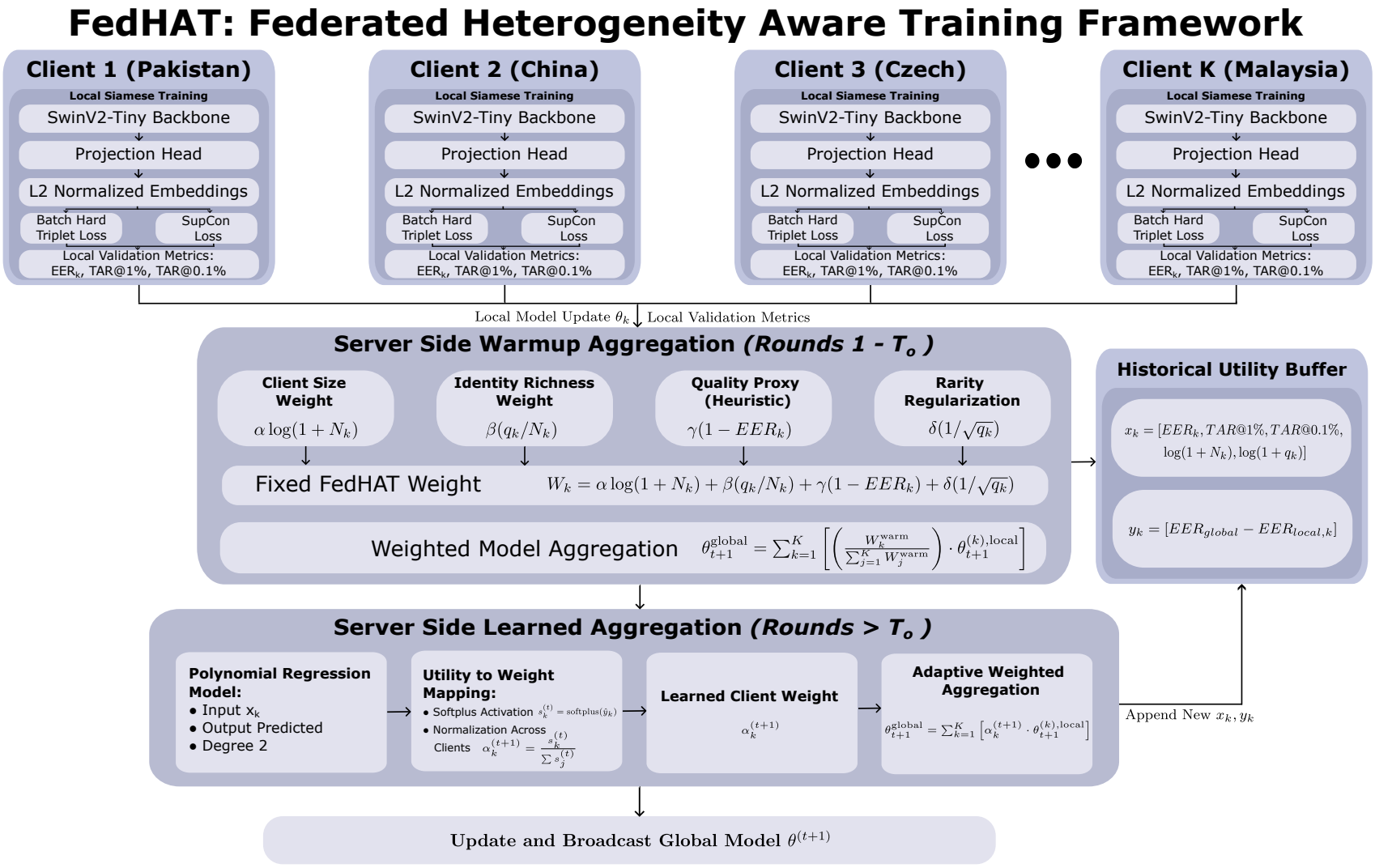

FIGURE 5.1: FedHAT: Federated heterogeneity-aware training framework.

Each target is paired with a feature descriptor

$$\mathbf{x}_k^{(t)} = \begin{bmatrix} \mathrm{EER}_{k,\mathrm{local}}^{(t)} \\ \mathrm{TAR}_{1\%,k}^{(t)} \\ \mathrm{TAR}_{0.1\%,k}^{(t)} \\ \log(1+N_k) \\ \log(1+q_k) \end{bmatrix}, \tag{5.6}$$

which jointly encodes validation behavior and dataset characteristics.

The primary objective of constructing these supervision pairs $\{(\mathbf{x}_k^{(t)}, y_k^{(t)})\}$ is to create a training dataset for the central server's adaptive aggregator. By pairing empirical dataset characteristics and performance metrics ($\mathbf{x}_k$) with the actual observed improvement in verification accuracy (y_k), the framework establishes a historical record of client utility. This data allows the central server to move beyond static, heuristic based weighting. It enables the server to learn a predictive mapping that identifies which clients are likely to contribute high quality updates in future rounds based on their current profile. The goal is to shift from human defined weighting rules to an automated, performance aware strategy that prioritizes updates with the highest expected empirical utility.

5.4 Adaptive Aggregation via Learned Regression

After completion of the stabilization phase, aggregation weights are no longer computed using fixed coefficients. Instead, a regression model is trained on the collected supervision pairs $\{(\mathbf{x}_k^{(t)}, y_k^{(t)})\}$ to estimate the expected utility of each client update as shown in Figure 5.2.

A second order polynomial regression model with ridge regularization is employed to capture nonlinear interactions between validation metrics and dataset attributes:

$$\hat{y}_k = f_{\mathrm{poly}}(\mathbf{x}_k). \tag{5.7}$$

To ensure numerical stability and positivity of aggregation weights, predicted utilities are transformed using the softplus function:

$$s_k = \log(1 + \exp(\hat{y}_k)) \,. \tag{5.8}$$

To ensure non negativity and numerical stability of the learned aggregation weights, the predicted utility values are transformed using a softplus function. This design choice is motivated by the expectation–maximization (EM) perspective on federated learning introduced by Louizos et al. [37]. In the EM formulation, client contributions can be interpreted as latent variables representing their expected responsibility in improving the global model. These latent responsibilities are inherently non negative and continuous. The softplus transformation provides a smooth, strictly positive mapping from unconstrained regression outputs to valid aggregation weights, aligning with this probabilistic interpretation. Unlike hard thresholding functions such as ReLU, softplus preserves gradient information for small or uncertain utility estimates, preventing abrupt client exclusion and enabling gradual reweighting as validation behavior evolves. This property is particularly important in federated settings with heterogeneous clients, where temporary performance degradation may occur due to domain shift or limited local data. Furthermore, softplus avoids the saturation effects associated with exponential or softmax based transformations when applied independently to client utilities.

5.4.1 Temporal Decoupling and Weight Application

To avoid same round information leakage, utilities predicted using statistics from round t are applied only in the subsequent aggregation step. The normalized aggregation coefficient for client k at round $t + 1$ is defined using the softplus-activated utility scores s from the previous round:

$$\alpha_k^{(t+1)} = \frac{s_k^{(t)}}{\sum_{j=1}^{K} s_j^{(t)}} \tag{5.9}$$

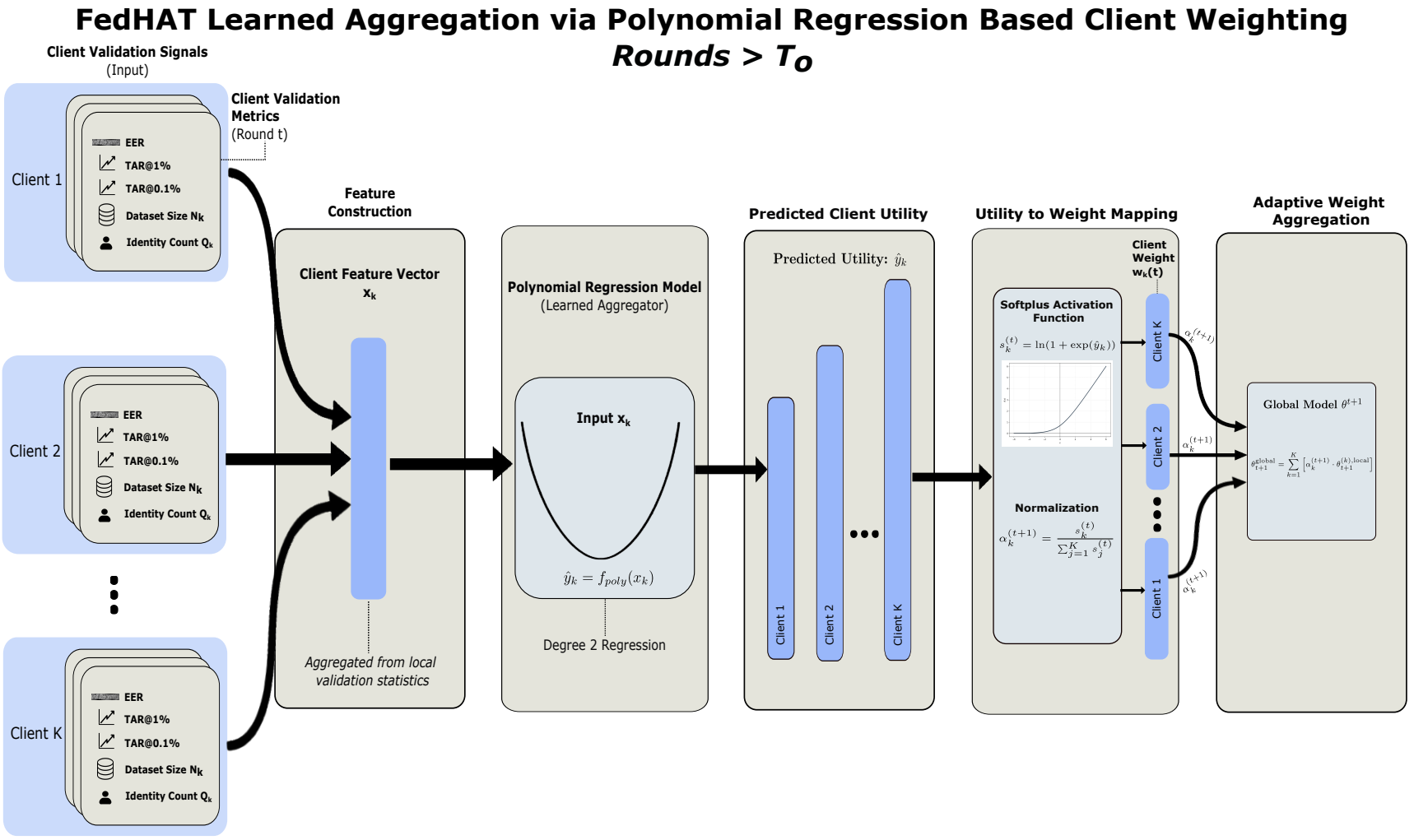

FIGURE 5.2: Learned aggregation phase of FedHAT.

The global model is updated by aggregating the local parameters returned by each client:

$$\theta_{t+1}^{\text{global}} = \sum_{k=1}^{K} \left[\alpha_k^{(t+1)} \cdot \theta_{t+1}^{(k),\text{local}} \right] \tag{5.10}$$

The components of these equations are defined as follows:

i. $\theta_{t+1}^{\text{global}}$: The updated global model parameters for the next communication round.

ii. K: The total number of participating federated clients.

iii. $\alpha_k^{(t+1)}$: The learned aggregation weight for client k, derived from its predicted utility in the preceding round.

iv. $s_k^{(t)}$: The softplus-activated utility score assigned to client k based on observed validation performance.

v. $\theta_{t+1}^{(k),\text{local}}$: The model parameters optimized locally by the k-th institution during the current round.

5.5 Handling of Batch Normalization Layers

Batch normalization statistics are known to be sensitive to dataset specific imaging characteristics such as illumination and sensor noise. To prevent distributional distortion, all batch normalization parameters are retained locally at each client and excluded from global aggregation. This design choice allows each client to preserve its own feature normalization behavior while still benefiting from shared representation learning. Following each communication round, the aggregated global model is evaluated on identity disjoint validation splits across all clients using the embedding level verification protocol described earlier. Training is terminated using early stopping when no further improvement is observed in the macro averaged Equal Error Rate across clients.

5.6 Overall Training Procedure

Algorithm 1 summarizes the complete federated optimization process, highlighting the transition from stabilization based aggregation to learned heterogeneity aware weighting.

Algorithm 1 FedHAT - Heterogeneity Aware Federated Training

Initialize global model θ_0^{global}
Initialize aggregation model $\mathcal{G}$
Initialize regression buffer $\mathcal{B} \leftarrow \emptyset$
for $t = 0$ to $T - 1$ **do**
  Broadcast θ_t^{global} to all clients
  for all clients $k = 1, \ldots, K$ **in parallel do**
    $\theta_{t+1}^{(k),\text{local}} \leftarrow \text{LocalTrain}(\theta_t^{\text{global}}, D_k)$
    Compute N_k, q_k and validation metric ϕ_k
  end for
  if $t < T_{\text{warm}}$ **then** ▷ Warm up stage
    for all clients k **do**
      $W_k^{\text{warm}} \leftarrow \alpha \log(1 + N_k) + \beta(q_k/N_k) + \gamma\phi_k + \delta/\sqrt{q_k}$
    end for
  else ▷ Learned aggregation stage
    for all clients k **do**
      $\mathbf{x}_k^{(t)} \leftarrow [\text{EER}_k, \text{TAR}_k^{1\%}, \text{TAR}_k^{0.1\%}, \log(1 + N_k), \log(1 + q_k)]$
      $W_k \leftarrow \text{Softplus}(\mathcal{G}(\mathbf{x}_k^{(t)}))$
    end for
  end if
  $\alpha_k^{(t+1)} \leftarrow W_k / \sum_j W_j$
  $\theta_{t+1}^{\text{global}} \leftarrow \sum_{k=1}^{K} \left[\alpha_k^{(t+1)} \cdot \theta_{t+1}^{(k),\text{local}}\right]$
  for all clients k **do**
    $y_k^{(t)} \leftarrow \max(0, \text{EER}_{\text{global},k} - \text{EER}_{\text{local},k})$
    Append $(\mathbf{x}_k^{(t)}, y_k^{(t)})$ to $\mathcal{B}$
  end for
  if $|\mathcal{B}| \geq N_{\min}$ **then**
    Update $\mathcal{G}$ via polynomial regression
  end if
end for
return θ_T^{global}

Chapter 6

Experimental Results

This section presents a comprehensive experimental evaluation of the proposed federated learning framework under identity disjoint, cross country iris recognition settings. Performance is reported for multiple federated aggregation strategies while enforcing strict data locality and privacy constraints.

All experiments follow the federated configuration described in Chapter 5, where each dataset functions as an independent client and raw biometric data are never exchanged. To ensure unbiased evaluation, all reported results are obtained exclusively on held out, identity disjoint test splits.

6.1 Computational Cost

The computational cost of the proposed FedHAT framework was evaluated under identical experimental settings. All experiments were conducted on a laptop equipped with an Intel Core Ultra 7 processor and NVIDIA RTX 5060 Laptop GPU. Standard federated baselines, including FedAvg, FedProx, and FedYogi, required approximately one hour to complete end to end training across all communication rounds. In contrast, FedHAT required approximately one hour and thirty minutes. The additional training time is mainly attributed to the validation driven aggregation stage, where client utility estimation and learned aggregation weights

are computed at the server. Despite this increase, the overall training time remains practical for offline federated biometric training and is justified by the consistent performance improvements observed under heterogeneous client conditions.

TABLE 6.1: Comparison of training time for federated learning models under identical experimental settings on GPU and CPU.

Model	**GPU Time**	**CPU Time**	**Relative Overhead**
FedAvg	$\sim$ 1 hour	$\sim$ 6 hours	Baseline
FedProx	$\sim$ 1 hour	$\sim$ 6 hours	Baseline
FedYogi	$\sim$ 1 hour	$\sim$ 6 hours	Baseline
FedHAT	$\sim$ 1 hour 30 min	$\sim$ 9 hours	+50%

6.2 Evaluation Metrics

Model performance is assessed at the embedding level using verification metrics commonly adopted in biometric systems. These metrics quantify the ability of the model to distinguish between genuine and impostor comparison pairs. For each client dataset, results are reported in terms of ROC–AUC, EER, and TAR at fixed False Accept Rates (FAR) of 1% and 0.1%. All metrics are computed using cosine similarity between ℓ_2-normalized embeddings.

6.2.1 Equal Error Rate

The EER is a standard measure used to summarize the accuracy of a biometric system. It is the specific operating point where the False Accept Rate (FAR) and the False Reject Rate (FRR) are equivalent.

i. FAR: The probability that the system incorrectly accepts an impostor sample as a genuine match.

ii. FRR: The probability that the system incorrectly rejects a genuine sample as an impostor.

A lower EER value signifies superior overall verification accuracy and better balance between security and user convenience.

6.2.2 Receiver Operating Characteristic and Area Under the Curve

The ROC curve visualizes the verification behavior across all potential operating thresholds by plotting the True Accept Rate (TAR) against the FAR. The Area Under the Curve (AUC) provides a single scalar value representing the overall discriminative capability of the model. An AUC of 1.0 indicates a perfect separation between genuine and impostor distributions.

6.2.3 True Accept Rate at Security-Critical Operating Points

In high security applications, the system must maintain a very low FAR to prevent unauthorized access. Consequently, performance is evaluated at fixed, strict security thresholds:

i. TAR @ 1% FAR: The percentage of genuine users correctly identified when the system is tuned to allow only one false acceptance in every 100 impostor attempts.

ii. TAR @ 0.1% FAR: A security critical metric reflecting performance under extremely tight constraints, allowing only one false acceptance in every 1,000 impostor attempts.

Reporting TAR at these levels is essential for assessing deployment relevant behavior in real world biometric environments.

6.3 Baselines and Compared Methods

The proposed federated learning framework is evaluated against widely adopted federated baselines to contextualize performance under severe client heterogeneity and identity disjoint biometric settings. All compared methods share identical network architectures, loss functions, local optimization hyperparameters, and evaluation protocols. The only point of variation lies in the server side aggregation strategy, ensuring a controlled and fair comparison.

6.3.1 FedAvg

Federated Averaging (FedAvg) serves as the primary baseline and represents the most commonly used aggregation strategy in federated learning. In this method, client updates are aggregated in proportion to the number of local training samples, implicitly assuming that larger datasets provide more reliable model updates. An overview of FedAvg pipeline is presented in Figure 6.1

While this assumption holds under relatively homogeneous data distributions, it is frequently violated in cross country biometric scenarios. In such settings, substantial variation exists across clients in terms of dataset size, identity diversity, and acquisition conditions, leading to aggregation bias in favor of dominant datasets.

6.3.2 FedProx

FedProx extends FedAvg by introducing a proximal regularization term during local optimization. This term constrains client updates to remain close to the global model parameters, with the objective of stabilizing training under non IID data distributions. An overview of FedProx is presented in Figure 6.2. Despite improved stability during local optimization, FedProx retains the same sample count based aggregation mechanism as FedAvg. As a result, explicit accounting for dataset quality, identity diversity, or contribution disparity across heterogeneous biometric clients is not incorporated.

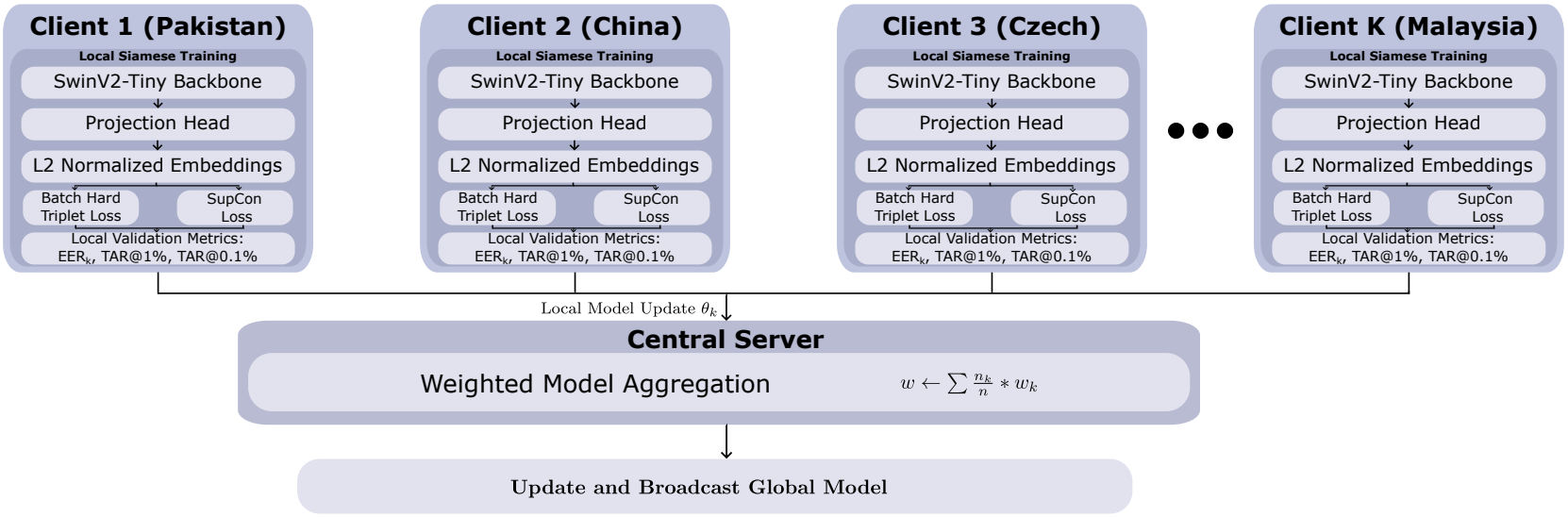

FIGURE 6.1: Federated Averaging aggregation mechanism.

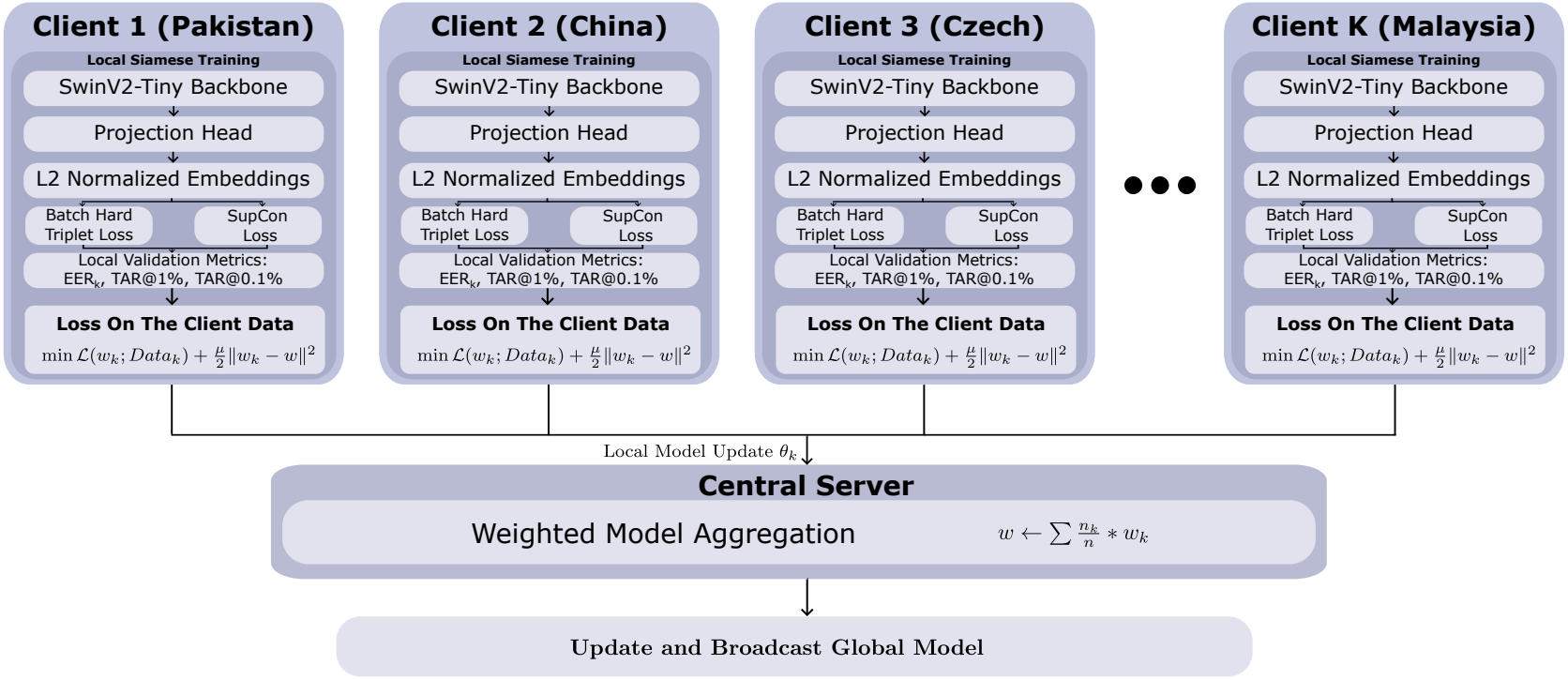

FIGURE 6.2: FedProx aggregation mechanism with proximal regularization.

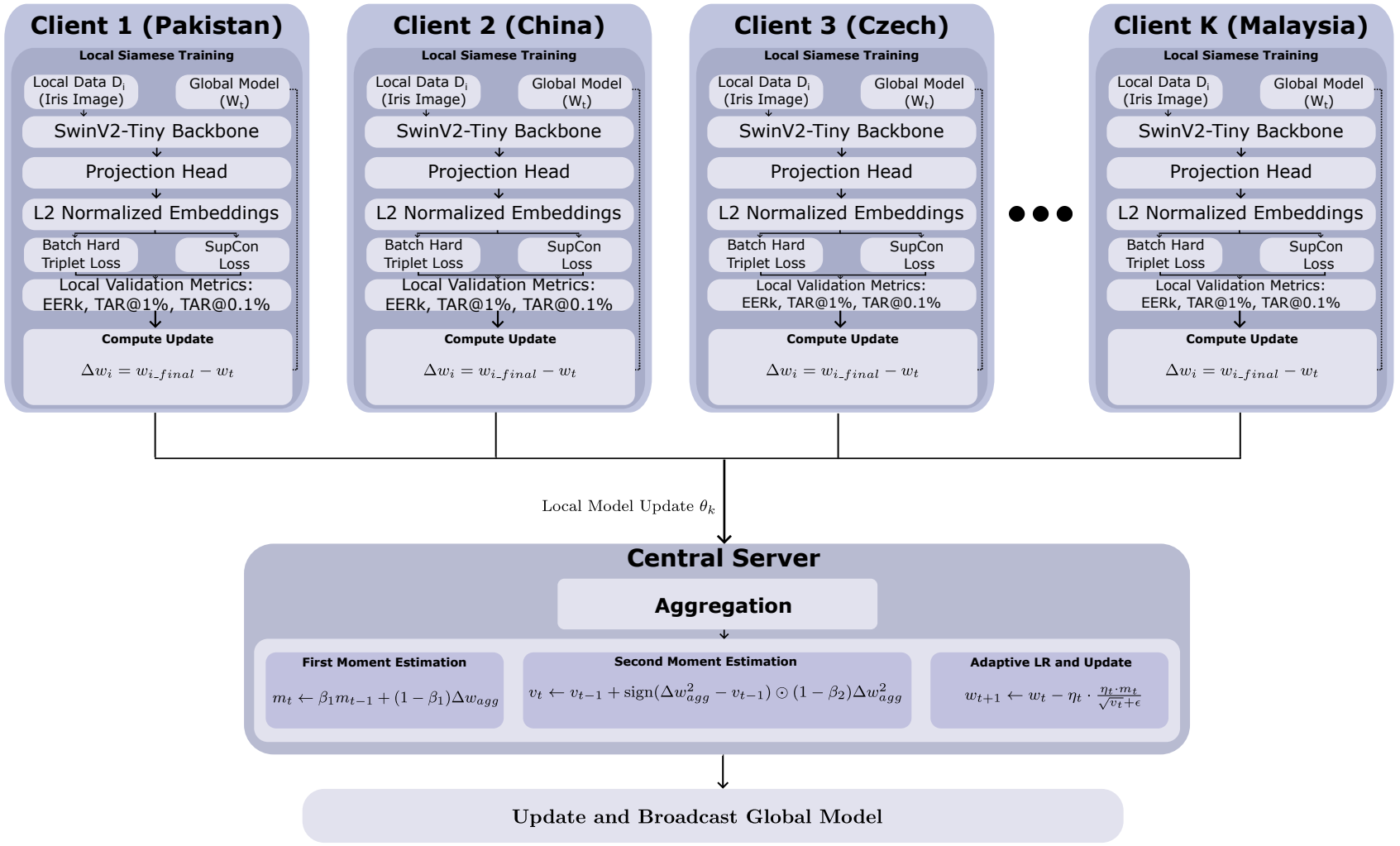

FIGURE 6.3: FedYogi aggregation mechanism using adaptive server-side optimization.

6.3.3 FedYogi

FedYogi represents an adaptive server side optimization approach that incorporates second order moment estimates to improve convergence under heterogeneous client updates. Unlike FedAvg and FedProx, FedYogi modifies the server update dynamics rather than the client weighting scheme. An overview of FedYogi is presented in Figure 6.3

While effective in certain distributed learning scenarios, the suitability of FedYogi for identity disjoint biometric verification under extreme heterogeneity remains unclear, as it does not explicitly model client contribution or dataset imbalance.

6.3.4 Proposed Method: FedHAT

The proposed method, FedHAT, introduces a heterogeneity aware aggregation strategy designed specifically for cross country biometric federated learning. FedHAT operates as a single federated algorithm with two sequential stages.

In the first stage, a stabilization phase is employed in which aggregation weights are computed using heuristic statistics derived from dataset size, identity richness, and validation performance. In the second stage, aggregation weights are learned adaptively using a polynomial regression model trained on observed validation utility. By explicitly modeling client utility as a function of both performance and dataset characteristics, FedHAT enables dynamic reweighting that mitigates dominance by large datasets while improving robustness for underrepresented clients.

6.3.5 Reporting Strategy

To facilitate detailed analysis and avoid metric coupling, verification performance is reported independently for each evaluation criterion. Per dataset comparisons of Equal Error Rate (EER), True Accept Rate (TAR) at 1% and 0.1% false accept rates, and ROC–AUC are presented separately in Tables 6.2–6.5. This separation

enables precise interpretation of system behavior under different operating conditions and highlights metric specific trends across federated aggregation strategies.

6.4 Federated Training Configuration

All federated experiments are conducted for a total of 250 communication rounds under identical communication schedules and local optimization settings.

For baseline methods (FedAvg, FedProx, and FedYogi), the corresponding aggregation strategy is applied consistently throughout all communication rounds.

For the proposed method (FedHAT), training is divided into two stages. During the initial 50 communication rounds, a warm up phase is employed in which aggregation weights are computed using heterogeneity aware heuristic statistics derived from dataset size, identity richness, and validation performance. This phase stabilizes early training and populates the regression buffer required for learning the aggregation function. During the remaining 200 rounds, learned aggregation is applied, with client contributions adaptively reweighted based on validation metrics and dataset characteristics from previous rounds.

Early stopping is applied uniformly across all methods based on the macro averaged validation Equal Error Rate (EER), with a patience of 150 rounds.

6.4.1 Per Dataset Verification Results

Verification performance is reported independently for each evaluation criterion to avoid metric coupling. Per dataset comparisons of Equal Error Rate (EER), True Accept Rate (TAR) at 1% and 0.1% false accept rates, and ROC-AUC are presented separately.

The subsequent subsections give a detailed analysis of the effectiveness of the proposed framework in different operational contexts. Detailed per dataset comparisons for the primary error measure are given in Table 6.2. Sensitivity and

security critical performance at the different operating points are reported in Table 6.3 and Table 6.4. Lastly, the total discriminative power of the model is recapped through the area under the curve in Table 6.5. Together, these statistics provide an overall picture of how the framework can respond to the inherent asymmetry of multi-institutional biometric data.

6.4.1.1 Equal Error Rate

TABLE 6.2: Per dataset comparison of Equal Error Rate (EER). Lower values indicate better performance. Best results are shown in **bold**, while second best results are highlighted in blue. Relative gains show FedHAT improvement over each baseline.

Dataset	FedAvg	FedProx	FedYogi	FedHAT	**Δ% vs Avg**	**Δ% vs Prox**	**Δ% vs Yogi**
CUST-Iris	0.0097	0.0127	0.0247	**0.0093**	+4.12	+26.77	+62.35
CASIA-Interval	0.0098	0.0109	0.0518	**0.0078**	+20.41	+28.44	+84.94
UPOL	**0.0556**	**0.0556**	0.1326	**0.0556**	+0.00	+0.00	+58.07
IITD	**0.0088**	0.0132	0.0132	**0.0088**	+0.00	+33.33	+33.33
AMF	**0.0091**	0.0200	0.0498	0.0100	-9.89	+50.00	+79.92
MMU	0.0446	0.0378	0.1682	**0.0357**	+19.96	+5.56	+78.78
UTIRIS	0.0164	0.0256	0.0787	**0.0150**	+8.54	+41.41	+80.94

6.4.1.2 True Accept Rate at 1% False Acceptance Rate

TABLE 6.3: Per dataset comparison of TAR at 1% FAR. Higher is better. Relative gains show FedHAT improvement over each baseline.

Dataset	FedAvg	FedProx	FedYogi	FedHAT	**Δ% vs Avg**	**Δ% vs Prox**	**Δ% vs Yogi**
CUST-Iris	0.9907	0.9861	0.9352	**0.9931**	+0.24	+0.71	+6.19
CASIA-Interval	0.9902	0.9883	0.6963	**0.9932**	+0.30	+0.50	+42.64
UPOL	0.8889	0.8889	0.5052	**0.9167**	+3.13	+3.13	+81.45
IITD	**0.9912**	0.9868	0.9824	**0.9912**	+0.00	+0.45	+0.90
AMF	**1.0000**	0.9600	0.7187	0.9900	-1.00	+3.13	+37.75
MMU	0.4625	0.4750	0.3113	**0.5000**	+8.11	+5.26	+60.62
UTIRIS	0.9487	0.9551	0.7002	**0.9808**	+3.38	+2.69	+40.07

6.4.1.3 True Accept Rate at 0.1% False Acceptance Rate

TABLE 6.4: Per dataset comparison of TAR at 0.1% FAR. Higher is better. Relative gains show FedHAT improvement over each baseline.

Dataset	FedAvg	FedProx	FedYogi	FedHAT	**Δ% vs Avg**	**Δ% vs Prox**	**Δ% vs Yogi**
CUST-Iris	0.9514	0.9236	0.6181	**0.9653**	+1.46	+4.52	+56.17
CASIA-Interval	0.8730	0.9268	0.3027	**0.9326**	+6.83	+0.63	+208.09
UPOL	0.8056	**0.8333**	0.3002	0.8056	+0.00	-3.32	+168.35
IITD	0.9670	0.9758	0.8462	**0.9824**	+1.59	+0.68	+16.10
AMF	**0.9900**	0.9600	0.1880	0.9800	-1.01	+2.08	+421.28
MMU	0.2250	0.3025	0.1104	**0.3125**	+38.89	+3.31	+183.06
UTIRIS	0.8718	**0.9103**	0.4150	0.8974	+2.94	-1.42	+116.24

6.4.1.4 Receiver Operating Characteristic–Area Under the Curve

TABLE 6.5: Per dataset comparison of ROC-AUC. Higher is better. Relative gains show FedHAT improvement over each baseline.

Dataset	FedAvg	FedProx	FedYogi	FedHAT	**Δ% vs Avg**	**Δ% vs Prox**	**Δ% vs Yogi**
CUST-Iris	**0.9996**	0.9995	0.9965	**0.9996**	+0.00	+0.01	+0.31
CASIA-Interval	0.9995	0.9995	0.9879	**0.9997**	+0.02	+0.02	+1.19
UPOL	0.9865	0.9829	0.9475	**0.9933**	+0.69	+1.06	+4.83
IITD	0.9979	0.9974	0.9974	**0.9995**	+0.16	+0.21	+0.21
AMF	**0.9999**	0.9992	0.9882	0.9998	-0.01	+0.06	+1.17
MMU	0.9785	**0.9897**	0.9085	0.9847	+0.63	-0.51	+8.39
UTIRIS	0.9984	0.9985	0.9687	**0.9993**	+0.09	+0.08	+3.16

6.4.2 Receiver Operating Characteristic Curve Evaluation

Receiver Operating Characteristic (ROC) curves are used to visualize verification behavior across different operating thresholds. These curves are generated by calculating the True Positive Rate (TPR) and the False Positive Rate (FPR) for every unique cosine similarity score produced by the test pairs.

The calculation process involves the following steps:

i. Pairwise Scoring: For a given client, cosine similarity scores are computed for all possible genuine pairs (same identity) and impostor pairs (different identities) in the test set.

ii. Threshold Sweeping: A threshold τ is moved across the entire range of similarity scores $[-1, 1]$. For each value of τ, the following rates are calculated:

   i) True Positive Rate (TPR): The proportion of genuine pairs whose similarity score is greater than or equal to τ. This is also known as the True Accept Rate (TAR).

   ii) False Positive Rate (FPR): The proportion of impostor pairs whose similarity score is greater than or equal to τ. This is also known as the False Accept Rate (FAR).

iii. Curve Generation: The ROC curve is formed by plotting the resulting (FPR, TPR) pairs on a 2D plane. Comparative ROC curves obtained using the FedAvg baseline and the proposed FedHAT global models across all identity disjoint test splits are shown in Figure 6.4. The curves illustrate how FedHAT maintains a higher TPR at lower FPR values compared to traditional aggregation methods.

6.4.3 Macro Averaged Performance

To summarize overall behavior across heterogeneous clients, macro averaged verification performance is computed as the unweighted mean across all datasets.

TABLE 6.6: Macro average verification performance across all federated clients.

Method	**EER**	**TAR@1%**	**TAR@0.1%**	**AUC**
FedAvg	0.0220	0.8960	0.8120	0.9943
FedProx	0.0251	0.8915	0.8331	0.9952
FedYogi	0.0741	0.6928	0.3972	0.9707
FedHAT (Proposed)	**0.0203**	**0.9093**	**0.8394**	**0.9967**

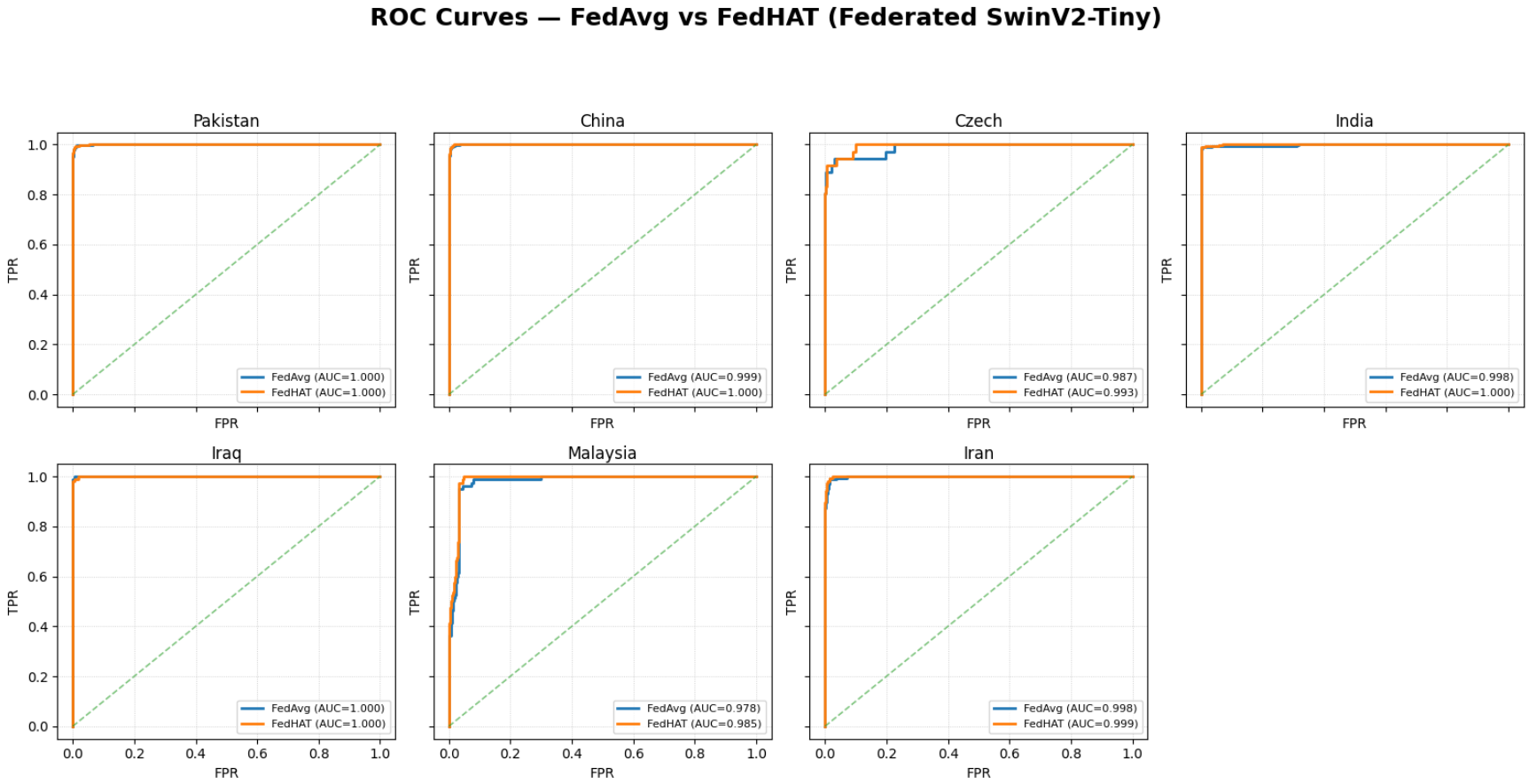

FIGURE 6.4: ROC curves for FedAvg and FedHAT evaluated on identity-disjoint test splits across all datasets. Each subplot corresponds to one client dataset.

6.5 Discussion

The experimental findings illustrate that the explicit acknowledgment of client heterogeneity is essential for federated learning in cross country iris verification. Conventional aggregation strategies do not take into account differences in dataset scale, sensing conditions, and identity distributions when strict identity disjoint and privacy preserving constraints are in place. The proposed FedHAT framework, on the other hand, improves and balances performance across all measured metrics while keeping data locality.

6.5.1 Effect of Heterogeneity Aware Aggregation

As it has been shown in the results of the per dataset reported in Tables 6.2–6.5, sample size based approaches to aggregation, including FedAvg and FedProx, are biased toward. large clients in disproportion. This conduct makes the process of verification less accurate, particularly when dealing with smaller or more complicated data sets, specifically ones that possess exclusive sensing abilities or identity variety.

FedHAT addresses this disparity by separating client input and dataset. size in itself and rather incorporating updates with an empirically determined utility. With this mechanism, the accuracy of verification of underrepresented is enhanced. clients and does not compromise performance in dominating datasets. These findings ensure that quality of contribution and not quantity of samples are isolated, is a better predictor of aggregation in biometric federated. learning.

6.5.2 Learned Aggregation versus Heuristic Weighting

Heuristic aggregation methods typically employ static or minimally adaptive weighting strategies, which are either pre-defined or amenable to manual adjustment, and these weights are either constant or exhibit limited adaptability throughout training iterations. Therefore, these methods aren't ideally suited for accurately

representing how the importance of clients changes across different data contexts. In contrast, learned aggregation directly calculates client weights based on how well they perform during training. This allows the aggregation process to adjust to changes in contribution for each client. This key difference allows learned aggregation to provide more consistent and fair model updates compared to heuristic weighting, particularly when client influence varies.

6.5.2.1 Ablation Analysis

The benefit of using learned aggregation weights, rather than using fixed heuristic weights, is shown in the ablation results in Table 6.7. While heuristic warm up aggregation provides stable optimization and competitive performance during early training, its limitations become apparent under pronounced heterogeneity in sensing conditions and identity distributions.

Across most datasets, consistent reductions in EER and improvements in TAR at both 1% and 0.1% FAR are observed when transitioning from heuristic weighting to the learned FedHAT aggregation. These gains are particularly evident for datasets such as CASIA-Interval, AMF, CUST-Iris, and UTIRIS, where fixed coefficients are less effective at modeling non linear variations in client utility. When client contributions are adjusted according to observed validation behavior, the learned aggregation method enhances reliability at security-critical operational junctures. These findings suggest that using heuristic weighting is enough for initial stabilization. However, learned aggregation provides better adaptability, which helps maintain consistent performance despite changing biometric differences.

6.5.2.2 Heuristic Weight Optimization

The stabilization phase coefficients ($\alpha = 0.35$, $\beta = 0.25$, $\gamma = 0.25$, $\delta = 0.15$) were selected via grid search to balance dataset scale and identity richness. While these optimized values provide a stable initialization, superior performance is achieved after transitioning to the learned FedHAT aggregation strategy.

TABLE 6.7: Ablation comparison between heuristic weighting (warm up aggregation) and learned aggregation (FedHAT).

Dataset	**EER ↓**		**TAR@1% ↑**		**TAR@0.1% ↑**	
	Heuristic	FedHAT	Heuristic	FedHAT	Heuristic	FedHAT
CUST-Iris	0.0096	**0.0093**	0.9907	**0.9931**	0.9537	**0.9653**
CASIA-Interval	0.0079	**0.0078**	0.9902	**0.9932**	0.9258	**0.9326**
UPOL	**0.0471**	0.0556	0.8889	**0.9167**	**0.8333**	0.8056
IITD	**0.0088**	**0.0088**	**0.9912**	**0.9912**	0.9758	**0.9824**
AMF	0.0151	**0.0100**	0.9700	**0.9900**	0.9500	**0.9800**
MMU	0.0500	**0.0357**	**0.6250**	0.5000	**0.4000**	0.3125
UTIRIS	0.0171	**0.0150**	0.9679	**0.9808**	0.8910	**0.8974**

6.5.2.3 Exploratory Development and Model Lineage

FedHAT was developed after testing out a few different setups, such as ArcFace-based optimization, GuardRail-style update constraints, and FedProx combined with optimized heuristic weighting. Table 6.8 gives a summary of these exploratory configurations.

These methods did make things more stable or improve performance in some cases, but they failed to perform well when there was an excessive degree of variety. The switch to a two stage learned aggregation framework fixed these problems by making it possible to adapt to each round, which led to better results than manually adjusted and strictly regularized baselines.

TABLE 6.8: Developmental lineage of exploratory configurations and technical hypotheses.

Hypothesis/Configuration	**Outcome Summary**
FedProx + Optimized Heuristic Weights	Improved stability; lacked round adaptivity
Manual Grid Search for Heuristic Weights	Non-linear utility variations not captured
ArcFace Loss Optimization	Suboptimal convergence
Update Constraints (GuardRail)	Over-restricted updates for diverse sensors
Learned Polynomial Aggregation	**Best Macro Average Performance**

6.5.3 Performance at Security Critical Operating Points

In biometric verification systems, low false accept rates are the most essential. FedHAT consistently improves TAR at FAR = 0.1% across most datasets. This implies that it does a better job of separating genuine and impostor distributions. These changes make not only the average verification accuracy better, but also the tail behavior of similarity score distributions, which is very important for high security deployment scenarios.

6.5.4 Robustness across Sensors and Acquisition Conditions

FedHAT shows strong performance across datasets collected using near-infrared, visible light, and mixed imaging methods. By retaining batch normalization layers locally at each client, sensor specific feature statistics are preserved. This design choice avoids the performance issues often seen when normalization parameters are shared across the entire model.

Chapter 7

Conclusion

7.1 Conclusion

This thesis examined federated learning for cross country iris verification within strict identity disjoint and privacy preserving parameters. The main objective was to establish how severe client heterogeneity affects federated biometric verification and whether adaptive aggregation strategies can help improve performance with variations in dataset size, identity diversity, sensing conditions, and data quality.

The experimental outcomes directly respond to the research questions presented in this study. First, it is demonstrated that client heterogeneity significantly influences federated iris verification, especially under circumstances surrounding identity disjointness and non-IID data distributions. FedAvg and FedProx favor large, similar clients. This makes them less effective for smaller or more difficult datasets. This confirms that using sample size to aggregate is not enough for cross country biometric federated learning.

Second, the results show that optimizer-level adaptations alone, like one used in FedYogi, can not fully make up for differences in client utility caused by variations in sensor variations and data quality. These methods could make convergence more stable, but they do not address how much each client should affect the global model.

Lastly, the proposed FedHAT framework answers the third research question by demonstrating that a dual stage aggregation strategy, merging heuristic stabilization with learned performance aware aggregation can greatly enhance reliability and impartiality among diverse clients. FedHAT achieves steady gains across seven datasets that are geographically and sensor rich by weighting client contributions based on empirical validation utility instead of just dataset size. Improved verification accuracy is particularly crucial for practical biometric deployments because it takes place at security critical operating points. The results show that learned aggregation that takes into account heterogeneity is a useful and effective way for federated biometric learning in environments where privacy is important and numerous institutions involved.

7.2 Limitations

Despite the positive results of this study, there are some limitations that need to be considered. The aggregation method used in FedHAT is intentionally limited to low-degree polynomial regression. This choice was made to improve stability, generalization, and interpretability. While this design choice reduces the likelihood of overfitting and unstable optimization, it could also hurt the ability of aggregation model to accurately represent complex, non linear client behaviors. Furthermore, the experimental assessment is confined to iris verification, representing a singular biometric application. Although the datasets incorporate diverse sensors and acquisition scenarios, the performance of the framework has not been evaluated within multimodal or open set biometric environments, where additional sources of variability could potentially arise.

7.3 Future Work

Based on these limitations, several directions for future research emerge. The aggregation mechanism can be extended to more expressive yet controlled models. Examples include shallow neural aggregators or lightweight attention based

weighting. Any such extension must preserve stability and interpretability. These models may better handle highly dynamic clients. This includes clients that participate intermittently.

Future work may also explore the use of FedHAT in multimodal biometric systems. Possible combinations include iris–face and iris–periocular verification. These settings introduce additional heterogeneity. This arises from sensing conditions and modality specific feature distributions. Another important direction is evaluation in open set environments. Dynamic biometric scenarios should also be considered.

Together, these directions support the development of more general federated learning systems. Such systems must remain adaptive and reliable. This is especially important in complex, real world biometric applications.

Bibliography

[1] John Daugman. "How Iris Recognition Works". In Al Bovik, editor, *The Essential Guide to Image Processing*, pages 715–739. Academic Press, 2009. doi: 10.1016/B978-0-12-374457-9.00025-1.

[2] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. "Swin Transformer V2: Scaling Up Capacity and Resolution". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019, 2022.

[3] Runqing Gao and Thirimachos Bourlai. "On Designing a SwinIris Transformer Based Iris Recognition System". *IEEE Access*, 12:30723–30737, 2024. doi: 10.1109/ACCESS.2024.3369035.

[4] Xianyun Sun, Caiyong Wang, Yunlong Wang, Jianze Wei, and Zhenan Sun. "IrisFormer: A Dedicated Transformer Framework for Iris Recognition". *IEEE Signal Processing Letters*, 32:431–435, 2025. doi: 10.1109/LSP.2024.3522856.

[5] Lin Jia, Bo Zhang, and Peng Li. "Stochastic Stylization Transformer with Self-Supervision for Iris Recognition". *Multimedia Systems*, 31(18), 2025. doi: 10.1007/s00530-024-01619-y.

[6] Lubos Omelina, Jozef Goga, Jarmila Pavlovicova, Milos Oravec, and Bart Jansen. "A survey of iris datasets". *Image and Vision Computing*, 108: 104109, 2021. doi: 10.1016/j.imavis.2021.104109.

[7] Kristtopher K. Coelho, Eduardo T. Tristão, Michele Nogueira, Alex B. Vieira, and José A. M. Nacif. "Multimodal biometric authentication method by

federated learning". *Biomedical Signal Processing and Control*, 85:105022, 2023. doi: 10.1016/j.bspc.2023.105022.

[8] Jing Guo, Haoran Mu, Xiaoming Liu, et al. "Federated learning for biometric recognition: a survey". *Artificial Intelligence Review*, 57(208), 2024. doi: 10.1007/s10462-024-10847-7.

[9] Fan Bai, Jiaxiang Wu, Pengcheng Shen, Shaoxin Li, and Shuigeng Zhou. "Federated Face Recognition". *arXiv preprint arXiv:2105.02501*, 2021.

[10] Guangyi Sun, Matías Mendieta, Abhishek Dutta, Xiaoming Li, and Chen Chen. "Towards Multi-modal Transformers in Federated Learning". In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gülşe Varol, editors, *Computer Vision – ECCV 2024*, volume 15073 of *Lecture Notes in Computer Science*. Springer, 2025. doi: 10.1007/978-3-031-72633-0_13.

[11] Guang Chen, Dacan Luo, Fengzhao Lian, Feng Tian, Xu Yang, and Wenxiong Kang. "A Multimodal Biometric Recognition Method Based on Federated Learning". *IET Biometrics*, 2024(1):5873909, 2024. doi: 10.1049/2024/5873909.

[12] Kevin Musgrave, Serge Belongie, and Ser Nam Lim. "A Metric Learning Reality Check". In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12370 of *Lecture Notes in Computer Science*, pages 681–699. Springer, 2020. doi: 10.1007/978-3-030-58595-2_41.

[13] Weiqing Deng, Liang Zheng, Yifan Sun, and Jian Jiao. "Rethinking Triplet Loss for Domain Adaptation". *IEEE Transactions on Circuits and Systems for Video Technology*, 31(1):29–37, 2021. doi: 10.1109/TCSVT.2020.2968484.

[14] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. "Supervised Contrastive Learning". In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020.

[15] H. Brendan McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. "Communication-Efficient Learning of Deep Networks from Decentralized Data". In *AISTATS*, 2017.

[16] Yinyin Wei, Xiangyang Zhang, Aijun Zeng, and Huijie Huang. "Iris Recognition Method Based on Parallel Iris Localization Algorithm and Deep Learning Iris Verification". *Sensors*, 22(20):7723, 2022. doi: 10.3390/s22207723.

[17] R. W. Jalal and M. F. Ghanim. "Enhancement of Iris Recognition System Using Deep Learning". In *Proceedings of the 2022 IEEE Symposium on Industrial Electronics and Applications (ISIEA)*, pages 1–7, 2022. doi: 10.1109/ISIEA54517.2022.9873666.

[18] J. E. Zambrano, J. I. Pilataxi, C. A. Perez, and K. W. Bowyer. "Iris Recognition Using an Enhanced Pre-Trained Backbone Based on Anti-Aliased CNNs". *IEEE Access*, 12:94570–94583, 2024. doi: 10.1109/ACCESS.2024.3425648.

[19] Kuan-Cheng Lin and Yu-Min Chen. "A High-Security-Level Iris Recognition System Based on Multi-Scale Dominating Feature Points". *IEEE Signal Processing Letters*, 31:1600–1604, 2024. doi: 10.1109/LSP.2024.3411513.

[20] Cheng-Shun Hsiao, Chia-An Chang, and Chih-Peng Fan. "Two-Stage Deep Learning Technology Based Iris Recognition Methodology for Biometric Authorization". *Journal of Advanced Information Technology*, 15(2):212–218, 2024.

[21] R. A. Abdulhasan, S. T. Abd Al-latief, and S. M. Kadhim. "Instant Learning Based on Deep Neural Network with Linear Discriminant Analysis Features Extraction for Accurate Iris Recognition System". *Multimedia Tools and Applications*, 83:32099–32122, 2024. doi: 10.1007/s11042-023-16751-6.

[22] Mengdi Wang, Efe Bozkir, and Enkelejda Kasneci. "Iris Style Transfer: Enhancing Iris Recognition with Style Features and Privacy Preservation through Neural Style Transfer". *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 8(2):21, 2025. doi: 10.1145/3729413.

[23] M. Li, Y. Wang, K. Zhang, Z. He, and Z. Sun. "Exploring Near-Infrared Iris Image Sequences for High Throughput Iris Recognition". *IEEE Transactions on Information Forensics and Security*, 20:5718–5731, 2025. doi: 10.1109/TIFS.2025.3573663.

[24] S. Lalitha, B. Padmavathy, C. S, R. Gomalavalli, P. V. Rajlakshmi, and V. Dhanakodi. "Improving Iris Recognition Systems with Transfer Learning and Pretrained CNN Models". In *Proceedings of the 2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET)*, pages 1–6, 2024. doi: 10.1109/ACET61898.2024.10730317.

[25] Himanshu Gupta, T. K. Rajput, Rahul Vyas, O. P. Vyas, and Antonio Puliafito. "Biometric Iris Identifier Recognition with Privacy Preserving Phenomenon: A Federated Learning Approach". In *Neural Information Processing (ICONIP 2022)*, volume 1794 of *Communications in Computer and Information Science*, pages 505–517. Springer, 2023. doi: 10.1007/978-981-99-1648-1_41.

[26] P. M. and K. Mahalakshmi. "Privacy-Secure and Decentralized Biometric Authentication Models Using Federated Learning Frameworks". In *2024 International Conference on System, Computation, Automation and Networking (ICSCAN)*, pages 1–7, 2024. doi: 10.1109/ICSCAN62807.2024.10894284.

[27] Zhengquan Luo, Yunlong Wang, Zilei Wang, Zhenan Sun, and Tieniu Tan. "FedIris: Towards More Accurate and Privacy-Preserving Iris Recognition via Federated Template Communication". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4016–4025, 2022.

[28] Harsh Sharma, Kush Pandey, and Divyashikha Sethia. "Preserving Privacy in Iris Recognition: A Federated Learning Approach". In *AIP Conference Proceedings*, volume 3297, page 060017, 2025. doi: 10.1063/5.0286884.

[29] Chinese Academy of Sciences' Institute of Automation. "CASIA Iris Image Database Version 4.0", 2010.

[30] Michal Dobeš and Libor Machala. "UPOL Iris Database". Accessed: 2025-02-13.

[31] M. Dobeš, L. Machala, P. Tichavský, and J. Pospíšil. "Human eye iris recognition using the mutual information". *Optik*, 115(9):399–405, 2004.

[32] M. Dobeš, J. Martinek, D. Skoupil, Z. Dobešová, and J. Pospíšil. "Human eye localization using the modified Hough transform". *Optik*, 117(10):468–473, 2006.

[33] Ajay Kumar and Arun Passi. "Comparison and combination of iris matchers for reliable personal identification". *Pattern Recognition*, 43(3):1016–1026, 2010. doi: 10.1016/j.patcog.2009.08.001.

[34] M. S. Hosseini, B. N. Araabi, and H. Soltanian-Zadeh. "Pigment Melanin: Pattern for Iris Recognition". *IEEE Transactions on Instrumentation and Measurement*, 59(4):792–804, 2010. doi: 10.1109/TIM.2009.2037996.

[35] Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M. Alvarez. "Personalized Federated Learning with First Order Model Optimization", 2021. arXiv:2012.08565.

[36] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. "Fair Resource Allocation in Federated Learning", 2020. arXiv:1905.10497.

[37] Christos Louizos, Matthias Reisser, Joseph Soriaga, and Max Welling. "An Expectation-Maximization Perspective on Federated Learning", 2021. arXiv:2111.10192.