ISSN 1688-2806

Universidad de la República
Facultad de Ingeniería

Optimization of Data Collection in Facial Recognition Models through Subsampling Strategies

Tesis presentada a la Facultad de Ingeniería de la Universidad de la República por

Silvana Tayler

en cumplimiento parcial de los requerimientos para la obtención del título de Magister en Ingeniería Matemática.

Directores de Tesis

Dr. Javier Preciozzi ..... Universidad de la República
Dr. Marcelo Fiori ..... Universidad de la República

Tribunal

Dr. Guillermo Carbajal ..... Universidad de la República
Dra. Marina Gardella ..... Universidad de la República
Dra. Lara Raad ..... Universidad de la República

Director Académico

Dr. Marcelo Fiori ..... Universidad de la República

Montevideo
Monday 18th May, 2026

Optimization of Data Collection in Facial Recognition Models through Subsampling Strategies, Silvana Tayler.

ISSN 1688-2806

Esta tesis fue preparada en LaTeX usando la clase iietesis (v1.1).
Contiene un total de 71 páginas.
Compilada el Monday 18th May, 2026.
http://fing.edu.uy/

Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.

George E. P. Box

This page has been intentionally left blank.

Acknowledgments

To my thesis advisors, Marcelo and Javier, for their guidance, their technical support, and, above all, the patience they showed me throughout this process. To my husband, Aldo, for his unconditional support and for standing by me during the long journey that this master's degree required. To my parents, for always providing me with the tools and encouragement to value education as a fundamental pillar in my life. And finally to my entire family and friends, for their constant encouragement and for reminding me at every stage that effort is always worthwhile.

This page has been intentionally left blank.

To my father, who saw the beginning of this work but could not be here for its completion.

This page has been intentionally left blank.

Abstract

Facial recognition systems have achieved remarkable performance in recent years; however, their accuracy remains highly dependent on the quality, diversity, and volume of training data. The widespread use of large-scale datasets, often collected without consent, raises significant ethical and legal concerns, while the storage and computational demands associated with such data present ongoing challenges. This thesis explores subsampling techniques to evaluate whether strategies can be identified that guide data collection, independently of the training process, with the goal of reducing data needs and improving computational efficiency.

ArcFace, a state-of-the-art facial recognition model, was selected as the baseline architecture due to its strong feature discrimination and generalization capabilities. Using the MS1M-ArcFace dataset for training and LFW, AgeDB-30, and CFP-FP benchmarks for evaluation, 53 experiments were conducted. Multiple sampling approaches were compared, at image and identity level, including uniform random selection, stratified sampling, k-means clustering and greedy Maximin selection. Both image and identity level subsampling were explored, with experiments designed to evaluate the effect of sample representativeness, intra- and inter-class variability, and the proportion of identities in the training set.

Results indicate that k-means clustering applied to ArcFace embeddings at the image level achieved the highest overall performance across all benchmark datasets, demonstrating its effectiveness in reducing redundancy while preserving intra-class and inter-class diversity. Alternatively, random sampling at the identity level yields competitive performance compared to more complex strategies, particularly when high intra-class variability is desired. This finding suggests that identity-level random sampling is a valid and cost-effective approach for training data selection, significantly reducing the costs of data collection, storage, and processing. Additionally, k-means clustering may serve as a more suitable alternative in scenarios with a limited number of identities and where greater intra-class variability is not required. These insights are especially relevant in ethically constrained environments, where biometric data collection is restricted by consent. In all cases, clustering can further guide the final image selection process once consent is obtained, enhancing both the efficiency and representativeness of the dataset.

This page has been intentionally left blank.

Contents

Chapter 1

Introduction

Facial recognition has rapidly become one of the most prominent applications of deep learning, demonstrating high performance and growing adoption across a variety of domains. However, the widespread deployment of facial recognition systems has also raised substantial ethical and legal concerns, particularly related to the use of large datasets compiled without the consent of data subjects.

Recent studies and legal cases [48] have highlighted significant issues associated with the collection of data for training facial recognition models, particularly regarding the legality and ethics of such data collection. Many datasets used to train these algorithms are collected without individuals' explicit consent, violating privacy rights, a practice considered illegal in several countries.

Additionally, despite advances in facial recognition, models still struggle under uncontrolled conditions. Variations in pose, lighting, and expression, the diversity of populations, and imbalanced datasets can impact model performance and generalization.

Furthermore, managing large volumes of data and their associated computational costs remains a critical challenge. Consequently, approaches that optimize model training through data reduction techniques have emerged [60], prioritizing the selection of the most relevant and representative information.

Given today's restrictions, as well as the significant costs associated with collecting, storing, and training on large-scale datasets, a fundamental question arises: *Is there a more efficient way to acquire data that still enables a model to achieve high effectiveness?*. To address this issue, the present work explores different subsampling strategies applied to large facial recognition datasets, with the aim of identifying potential patterns that may guide optimal data collection practices.

For illustration, consider a dataset containing a million images from 20,000 individuals, but suppose we aim to train a face recognition model using only a subset of 100,000 images. Several strategies are conceivable:

- selecting the most representative images from the entire dataset,

- selecting the most representative identities,
- prioritizing samples that are maximally distinct from one another,
- choosing a random subset,
- focusing on a small number of individuals with many images each, or
- covering a larger number of individuals with fewer images each.

Key questions naturally follow: How many samples are sufficient for an effective model? How should representativeness be defined? And ultimately, how much does this decision impact final model performance?

The main goal of this work lies in the identification of techniques or patterns that can be applied to data selection and collection, rather than to the training process itself, in order to improve data and computational efficiency, while ensuring compliance with ethical and regulatory standards. The main contribution is determining whether optimizing data collection strategies results in measurable improvements in model accuracy. The specific objectives are: (i) to investigate the selection of data subgroups for model training to identify patterns for future data collection, and (ii) to train a facial recognition model using various data subsets to test the hypothesis that there exists a subsampling method capable of consistently producing better results across all evaluations.

This document is structured as follows: Chapter 2 reviews related work on facial recognition and sample selection, Chapter 3 details the proposed methodology, Chapter 4 presents experimental results, and Chapter 5 concludes with a discussion of findings and directions for future research.

Some of the results discussed in this work were previously published in Optimization of Data Collection in Facial Recognition Models through Subsampling Strategies [56] and presented at the International Joint Conference on Biometrics (IJCB 2025), while additional experiments and analyses are included here for completeness.

Chapter 2

Theoretical Framework

Since the goal is to evaluate how different sampling techniques affect model performance, it is first necessary to define the underlying facial recognition model that will remain constant throughout the experiments. The theoretical framework of this work is structured around two key components that align with the objective of identifying effective strategies for sampling the data space while keeping all other experimental variables fixed. The first section, Neural Networks and Facial Recognition, provides an overview of the theoretical foundations of deep learning architectures and facial recognition, its challenges, the state of the art, and the rationale for selecting the ArcFace model as the experimental baseline. The second section, Data Sampling, reviews the main sampling methodologies relevant to this study, focusing on those that may theoretically inform or guide future data collection strategies prior to training.

2.1 Neural Networks and Facial Recognition

2.1.1 Neural Networks and Deep Learning

Artificial neural networks (ANNs) have become a central technique in modern machine learning. Inspired by the organization of biological neurons, they are designed as computational models that learn to approximate complex functions by adjusting the weights of interconnected nodes across multiple layers. Early neural network research established the theoretical foundations for pattern recognition, yet practical adoption was limited for decades due to computational constraints and insufficient training data. The resurgence of interest in the 2000s was fueled by advances in hardware, particularly graphics processing units (GPUs), as well as the availability of large datasets, which enabled neural networks to scale in depth and complexity [32].

Deep learning, a branch of machine learning based specifically on multiple levels of layers in a neural network, has since become the dominant approach in computer vision and related domains. In this framework, lower layers of the network capture simple patterns such as edges or textures, while deeper layers progressively encode

higher-level semantic concepts. Its key advantage lies in its ability to automatically learn hierarchical feature representations from raw data, where higher-level features are constructed from simpler ones, which has proven especially powerful in image analysis tasks, where variability in the images pose substantial challenges for traditional methods [20].

Among deep learning models, convolutional neural networks (CNNs) have played a pivotal role in advancing computer vision. These networks are built as a series of stages, where each stage transforms the input into feature maps that highlight specific patterns in the data. For images, feature maps are like filtered versions of the original picture that capture things such as edges or textures. Each stage usually has three parts: filters that detect patterns, a non-linear function that makes the network flexible, and pooling that reduces the size while keeping important information [33]. By leveraging convolutional filters and local receptive fields, CNNs effectively exploit the spatial structure of images, allowing them to detect patterns invariant to translation and deformation. Architectures such as AlexNet [31], VGG [52], and ResNet [23] have demonstrated the scalability of CNNs to very deep structures, achieving breakthroughs in large-scale image recognition challenges. These successes have positioned CNNs as the backbone of modern facial recognition systems, where robust and discriminative feature extraction is critical.

Deep learning thus provides a flexible and powerful framework for representation learning across domains, particularly in tasks that involve high-dimensional and complex data such as images, audio, and text. Its ability to discover hierarchical and transferable representations has made it the foundation of modern artificial intelligence research. In the following sections, we examine how these principles extend to the specific case of facial recognition, where convolutional architectures and specialized methods have achieved remarkable performance.

2.1.2 Facial Recognition

Facial recognition aims to automatically detect and analyze human faces, extract discriminative features, and compare them against stored representations to perform identification or verification of individuals from images or video sequences.

This field has undergone significant advancements in recent decades, establishing itself as one of the most relevant applications in computer vision. This technology has become a fundamental tool across various domains, from security and surveillance to entertainment and e-commerce. Its widespread adoption is largely due to the ease of acquiring biometric information in uncontrolled environments and its strong identity discrimination capabilities.

The techniques used have evolved significantly in the last years. Specifically, deep convolutional neural networks (DCNNs) have transformed facial recognition, making models more robust, accurate, and adaptable to diverse conditions. These advancements have enabled performance levels comparable to human recognition in controlled environments, driven by advanced architectures, specialized loss functions, and increased computational capacity. However, model performance remains heavily dependent on the quantity, quality, and diversity of the training data.

Technical Challenges and Bias in Facial Recognition

Despite progress in facial recognition models, several challenges persist in uncontrolled conditions. Variations in pose, illumination, and facial expression (Pose, Illumination, Expression - PIE) [57] remain among the primary obstacles to accurate identification, as a single individual can appear significantly different across images. Additionally, occlusions—such as glasses, masks, or hair—can alter facial patterns, reducing a model's ability to reliably identify individuals.

Another key challenge is model generalization, ensuring that systems perform consistently across diverse populations and conditions. Studies have demonstrated that redundancy or imbalanced data distributions can introduce biases that undermine model effectiveness [9, 45]. These biases may originate from multiple sources, including dataset composition (e.g., age [11], race [62], gender [2]), class geometric characteristics [37], or model architecture [15]. To mitigate these issues, various strategies have been explored, such as the intelligent selection of representative data subsets [18], oversampling [24], undersampling [34], and more sophisticated approaches like DebFaceID [19], aimed at balancing demographic distributions.

Furthermore, managing large volumes of data and their associated computational costs remains a critical challenge. Consequently, approaches that optimize model training through data reduction techniques have emerged, prioritizing the selection of the most relevant and representative information, as discussed in Section 2.2.

Ethical Considerations and Data Collection Regulations

Beyond technical challenges, data collection for training facial recognition models raises significant ethical and legal concerns [40, 48]. Many facial recognition datasets have been obtained without explicit user consent, violating privacy rights and, in many jurisdictions, being considered illegal [40]. A well-known case is Clearview AI [44], currently facing multiple lawsuits for using facial data from social media without user authorization. The increasing regulation of biometric data collection has led to the establishment of new standards and legal frameworks aimed at ensuring ethical practices in this field [8, 58].

These regulatory changes have significantly impacted the availability of large-scale facial recognition datasets, such as MS1M [21] and VGGFACE2 [6], affecting both the quantity and diversity of data used for model training and increasing the costs associated with data acquisition. These trends underscore the importance of collecting ethically sourced data that ensures both user consent and model reliability.

2.1.3 State of the Art in Facial Recognition Tasks

The main goal of this study is not to identify the best-performing facial recognition model, but rather to evaluate sampling methods that can optimize model performance. Hence, the choice of the base model remains crucial, as it provides the foundation for consistent comparisons. To provide context for this choice, an overview of the state-of-the-art approaches that have demonstrated strong performance on the evaluation datasets used in this study is included.

Benchmark Datasets

Three benchmark datasets are widely used in the field, as they present diverse

challenges for biometric identification and serve as standard references for evaluating facial recognition methods. These datasets are:

Labeled Faces in the Wild (LFW) is one of the most commonly used datasets in facial recognition for evaluating models under unconstrained conditions [25]. It comprises 13,233 images of 5,749 individuals, collected from the web, with significant variations in illumination, expressions, and partial occlusions, which makes it an essential benchmark to assess the robustness of recognition models.

Cross-Pose Face Pairs - Frontal-Profile Matching (CFP-FP) dataset [53] is specifically designed to assess facial recognition under extreme pose variations. It consists of 500 individuals, each with 10 frontal images and 4 profile images, allowing models to be tested on significantly different views. This is one of the most challenging datasets, as the similarity between profile and frontal images is substantially lower than in other datasets, making it a reference for evaluating pose-invariant recognition models.

AgeDB-30 focuses on age variation in facial recognition, containing 16,488 images of 568 subjects. Image pairs were selected from the AgeDB database [39], ensuring a 30-year gap between them. Currently, it is one of the most demanding benchmarks for evaluating a model's robustness against natural aging effects, as it tests the ability of models to recognize individuals despite significant changes in facial structure over time.

Samples of the three datasets are shown in Figure 2.1.

Models

The following paragraphs present the models that have achieved the best-known performance as of the time of this study, based on the rankings provided by Papers With Code [12].

Figure 2.1: Examples of LFW, CFP-FP and AgeDB-30 benchmark datasets.

GhostFaceNets

GhostFaceNets introduces a novel lightweight facial recognition architecture designed for resource-constrained devices [1]. This model is based on GhostNetV1 and GhostNetV2, which were initially developed for general computer vision tasks.

GhostNetV1 [22] incorporates the Ghost Module, which reduces feature redundancy, improves efficiency, and generates additional feature maps without increasing

the number of parameters. This is achieved through low-cost linear transformations, producing Ghost Feature Maps that retain critical information without requiring additional convolutions.

GhostNetV2 [55] enhances this approach by incorporating Dynamic Fully-Connected Attention (DFC Attention), which improves global feature extraction and combines it with local features obtained from GhostNetV1. This way, the Ghost Feature Map generation is optimized, achieving greater expressiveness while maintaining low computational costs.

To adapt GhostFaceNets for facial recognition, several modifications were introduced. The Global Average Pooling (GAP) layer was replaced with a Global Depthwise Convolution layer (7×7 kernel + BN + 1×1×*embedding_size* convolution), which captures more relevant differences across feature units. The activation function was changed from ReLU to PReLU[1], allowing the model to better capture non-linear relationships in the data. Additionally, the Squeeze-and-Excitation (SE) modules were modified by replacing Fully Connected layers with 1×1×*channel_axis* convolutions, adjusting the weight of each channel according to importance, with minimal computational cost.

The loss function used in GhostFaceNets is ArcFace, detailed in Subsection 2.1.4, which minimizes intra-class variability and maximizes inter-class differentiation, demonstrating superior performance in facial recognition compared to Softmax and CosFace. Figure 2.2 shows the model architecture based on GhostNetV2.

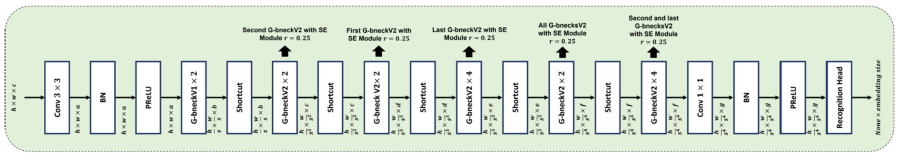


Figure 2.2: GhostFaceNetsV2 architecture, as presented by Alansari et al. (2023) [1].

During training, GhostFaceNets was trained on multiple versions of the MS1M [2] dataset and evaluated on multiple benchmarks, including LFW and CFP-FP. The results demonstrated state-of-the-art accuracy while improving computational efficiency. Specifically, the GhostFaceNetV2-1 (MS1MV3) model achieved an accuracy of 99.8667% on LFW and 99.33% on CFP-FP, making it one of the best-performing models for these tasks at the time of this study.

ProdPoly

The Π-Nets architecture introduces a polynomial-based neural network designed to enhance both generative and discriminative tasks beyond traditional deep convolutional networks [10]. It has achieved state-of-the-art results in image generation, face verification, and 3D mesh representation learning.

Unlike conventional CNNs, this model outputs a high-order polynomial function based on input data, as can be shown in Figure 2.3, where unknown parameters are

[1]Parametric ReLU modifies ReLu by allowing a small slope α for negative values, where α is a learnable parameter, improving convergence and accuracy.

[2]MS1M is a large-scale facial recognition dataset containing millions of images of celebrities collected from the internet, which will be described in greater detail later in this document.

Method	CCP	NCP	NCP-Skip
Rec. expr. for N-th order	$x_n = \left(U^T_{[n]z}\right) * x_{n-1} + x_{n-1}$ with $x_1 = \left(U^T_{[1]z}\right)$ and $x = Cx_N + \beta$	$x_n = \left(A^T_{[n]z}\right) * \left(S^T_{[n]}x_{n-1} + B^T_{[n]}b_{[n]}\right)$ with $x_1 = \left(A^T_{[1]z}\right) * \left(B^T_{[1]}b_{[1]}\right)$ and $x = Cx_N + \beta$	$x_n = \left(A^T_{[n]z}\right) * \left(S^T_{[n]}x_{n-1} + B^T_{[n]}b_{[n]}\right) + V_{[n]}x_{n-1}$ with $x_1 = \left(A^T_{[1]z}\right) * \left(B^T_{[1]}b_{[1]}\right)$ and $x = Cx_N + \beta$
Learnable Parameters	$C \in \mathbb{R}^{O\times k}, U_{[n]} \in \mathbb{R}^{d\times k}$	$B_{[n]} \in \mathbb{R}^{\omega\times k}, b_{[n]} \in \mathbb{R}^{\omega}$	same as in NCP and $V_{[n]} \in \mathbb{R}^{k\times k}$

Table 2.1: Recursive expressions for the three polynomial decomposition methods proposed by the authors.

represented as high-order tensors, and with no need for non-linear activation functions. To prevent exponential growth in parameters, the authors propose a coupled tensor factorization approach, effectively reducing the number of polynomial parameters needed.

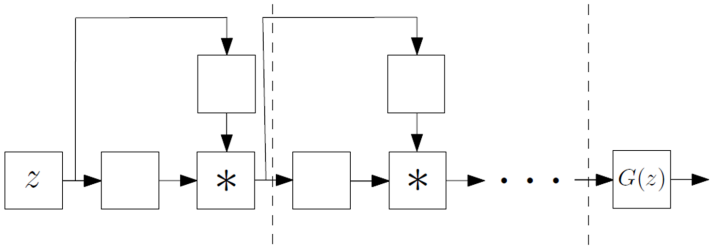


Figure 2.3: Π-Nets architecture, as presented by Chrysos et al. (2020) [10].

Starting from the objective of learning the function $G : \mathbb{R}^d \to \mathbb{R}^O$ of order $N \in \mathbb{N}$, such that:

$$x = G(z) = \sum_{n=1}^{N} \left(\mathcal{W}^{[n]} \prod_{j=2}^{n+1} \times_j z \right) + \beta, \tag{2.1}$$

where:

- $n \in \mathbb{N}$ and $N \in \mathbb{N}$ denote the polynomial term order and the total approximation order, respectively,
- $\times_j$ is the mode-j tensor–vector product
- $z_i \in \mathbb{R}^d$ is the input to the polynomial approximator,
- $\beta \in \mathbb{R}^O$ and $\{\mathcal{W}^{[n]} \in \mathbb{R}^{O\times\prod_{m=1}^{n}\times_m d}\}_{n=1}^{N}$ are the learnable parameters.

Since directly parameterizing a full N-th order tensor is infeasible, the authors propose three recursive formulations based on tensor decompositions: Coupled CP decomposition (CCP), Nested Coupled CP decomposition (NCP), and Nested Coupled CP decomposition with skip connections (NCP-Skip). The recursive expressions for the three variants are given in Table 2.1.

Instead of relying on a single polynomial, the function approximation is expressed as a product of polynomials, that allow using different decompositions, which is why the model is referred to as ProdPoly.

One of the primary advantages of ProdPoly is that it achieves higher expressiveness with fewer parameters, representing a major improvement over traditional convolutional networks. This innovation has been successfully applied to both generative models, such as GANs, and discriminative models for identity verification.

For facial verification, the model was trained on MS1M-RetinaFace, using ResNet50 as the base architecture. The residual blocks in ResNet50 were converted into second-order residual blocks to construct ProdPoly-ResNet, while maintaining all other configurations of the baseline model. When evaluated on several datasets, amongst which is AgeDB-30, ProdPoly achieved the highest recorded accuracy at the time of this study, reaching 98.467%. Compared to the state-of-the-art ResNet50 + ArcFace model, it demonstrated a 0.24% improvement in accuracy, suggesting a slightly superior ability to handle aging-related variations.

Final Remarks and Research Direction

Despite the recent advancements introduced by GhostFaceNets and ProdPoly, the most widely adopted models in both industry and research continue to be FaceNet [47], ArcFace [13], SphereFace [36], and VGGFace2 [6]. FaceNet pioneered direct embedding learning through triplet loss, while ArcFace significantly improved feature discrimination by introducing the Additive Angular Margin Loss. SphereFace further enhanced class separation by incorporating hypersphere metrics, and VGGFace2 distinguished itself with a diverse dataset that allowed for improved generalization across different poses and age variations.

While newer models offer advantages in computational efficiency and parameter reduction, traditional models remain dominant due to their stability, strong benchmark performance, and widespread adoption in real-world applications. Given these considerations, ArcFace was selected for this study as the primary model, as it has consistently demonstrated strong performance across multiple benchmarks, including LFW, CFP-FP, and AgeDB-30. Its potent feature discrimination capabilities and ability to generalize across various recognition challenges establish it as a leading reference in the field of facial recognition.

2.1.4 ArcFace

ArcFace [13] is a face recognition framework that combines a deep convolutional neural network (DCNN) backbone with a hyperspherical embedding space, in which each face image is mapped to a normalized feature vector. Typically, architectures such as ResNet-50 or ResNet-100 [23] are employed as the backbone to extract rich facial representations, which are then projected onto a unit hypersphere through an L2-normalization layer. This normalization step ensures that comparisons between embeddings can be consistently expressed in terms of angular distances.

The core contribution of ArcFace lies in its training objective, the Additive Angular Margin Loss. In their paper, the authors propose this method as an improvement over previous loss functions such as Softmax, SphereFace, and CosFace, introducing an additional angular margin that directly corresponds to the geodesic distance on a hypersphere. This approach enables the model to learn more compact intra-class features and achieve better inter-class separation.

Formulation

Let us start from the conventional Softmax loss used for classification tasks, defined

as:

$$L_{\text{softmax}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}\right), \tag{2.2}$$

where:

- x_i is the output feature vector of sample i,
- W_j is the weight vector of class j,
- b_j is the bias associated to class j,
- y_i is the label corresponding to x_i,
- N is the batch size,
- n is the total number of classes.

Notably, this loss does not explicitly optimize for intra-class compactness or inter-class separation, which may result in performance gaps in situations with high intra-class variability, such as large pose or age differences within the same identity. For this reason, the authors propose a modification: applying L2-normalization to both the feature and weight vectors, removing the bias term for simplicity, and scaling the features by a factor s. This scale factor is a hyperparameter introduced to ensure numerical stability, since after normalization the magnitude of the logits becomes very small. By multiplying the logits by s, the values are rescaled to an appropriate range, which stabilizes optimization and improves convergence. This replaces $W_j x_i + b_j$ with:

$$sW_j^T x_i = s\|W_j\| \cdot \|x_i\| \cdot \cos(\theta_j), \tag{2.3}$$

and since W_j and x_i are normalized ($\|W_j\| = \|x_i\| = 1$), the scalar product depends only on the angle θ_j between them. The main innovation of ArcFace lies in introducing an additive angular margin m directly to the angle between the feature vector x_i and the corresponding class weight vector W_{y_i}. This margin is a fixed hyperparameter measured in radians, whose value was chosen empirically, to force a certain separation between classes. This results in the following formulation:

$$L_{\text{ArcFace}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{s\cdot\cos(\theta_{y_i}+m)}}{e^{s\cdot\cos(\theta_{y_i}+m)} + \sum\limits_{\substack{j=1 \\ j\neq y_i}}^{n} e^{s\cdot\cos(\theta_j)}}\right), \tag{2.4}$$

where:

- θ_{y_i} is the angle between the features vector x_i and the weight vector W_{y_i} of the ground truth,
- m is the additive angular margin,
- s is a scale factor.

This formulation effectively pushes samples of the same class into more compact clusters and increases inter-class separation, by directly optimizing the geodesic distance between identity embeddings on the surface of a hypersphere of radius s. Figure 2.4 illustrates the ArcFace loss.

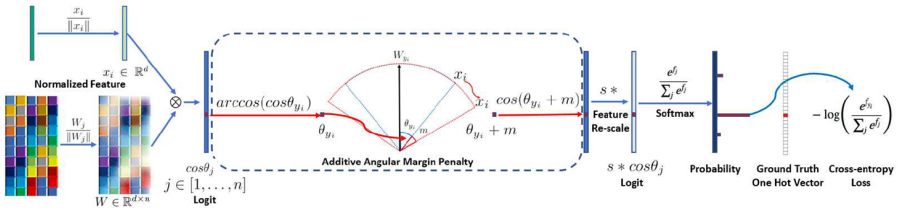


Figure 2.4: ArcFace loss, as presented by Deng et al. (2019) [13].

Compared to prior loss functions, ArcFace offers significant advantages in terms of training stability and geometric interpretability. SphereFace applies a multiplicative angular margin, where the angle between the feature vector and the class center is multiplied by an integer factor, as shown in Equation 2.5.

$$L_{\text{SphereFace}} = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{e^{\|x_i\|\cos(m\theta_{y_i,i})}}{e^{\|x_i\|\cos(m\theta_{y_i,i})} + \sum\limits_{\substack{j=1 \\ j\neq y_i}}^{n} e^{\|x_i\|\cos(\theta_{j,i})}}\right). \tag{2.5}$$

While this formulation is theoretically valid, it complicates training and may cause convergence instability. CosFace, on the other hand, introduces an additive margin in the cosine space by subtracting a constant from the cosine similarity between vectors, as presented in Equation 2.6.

$$L_{\text{CosFace}} = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{e^{s\cdot(\cos(\theta_{y_i})-m)}}{e^{s\cdot(\cos(\theta_{y_i})-m)} + \sum\limits_{\substack{j=1 \\ j\neq y_i}}^{n} e^{s\cdot\cos(\theta_j)}}\right), \tag{2.6}$$

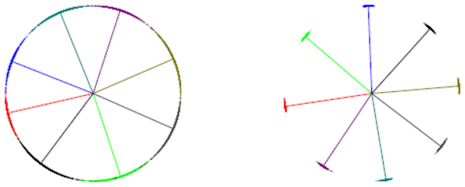
Although easier to implement, this method lacks a direct correspondence to geodesic distance on the hypersphere. In contrast, ArcFace applies an additive margin directly in the angular space, which involves adding a fixed angle before computing the cosine. This leads to a formulation that preserves geometric clarity, simplifies implementation, and has better empirical results.

In what follows, we adopt the same notation as in the previous equations; in particular, x_i denotes the feature vector and θ_{y_i} the angle between the feature vector and the class weight vector, and m the corresponding margin as stated above.

Results

The model was tested on multiple datasets and compared against other loss functions. Figure 2.5 shows a toy example under the softmax and ArcFace loss on 8 identities with 2D features. Table 2.2 presents results obtained by the authors, using ResNet50 trained on CASIA and comparing different loss functions. Table 2.3 shows a comparison of ArcFace (trained with ResNet100 on MS1MV2) against other state-of-the-art methods.

ArcFace introduces key improvements that have established it as a reference model in the field of facial recognition. By incorporating a fixed angular margin, it ensures



(a) Softmax (b) ArcFace

Figure 2.5: Comparison of Softmax and ArcFace losses in a toy example, as presented by Deng et al. (2019) [13].In this example, the authors trained two 2D embedding networks, one with Softmax loss and the other with ArcFace loss. The visualizations show that with Softmax the decision boundaries remain relatively ambiguous, with class clusters overlapping at the edges. By contrast, introducing an additive angular margin m during training forces a clearer separation between classes, resulting in more compact intra-class clusters and more distinct inter-class boundaries.

Model	**LFW(%)**	**CFP-FP(%)**	**Age-DB30(%)**
Softmax	99.08	94.39	92.33
SphereFace	99.42	94.38	91.7
CosFace	99.51	95.44	94.56
ArcFace (0.5 mg)	**99.53**	**95.56**	**95.15**

Table 2.2: Results of different loss functions on LFW, CFP-FP and Age-DB30, trained on CASIA with ResNet50, as presented by Deng et al. (2019) [13]. These results were obtained using the same backbone and datasets, varying only the loss function. Here the authors, applied a 0.5 additive angular margin in the ArcFace loss.

that each identity forms a compact cluster in the feature space while maintaining clear inter-class separation. This geometric approach not only enhances identity discrimination but also enables the learning of more robust and precise facial representations, achieving state-of-the-art performance across multiple datasets. Unlike earlier methods such as SphereFace, which required auxiliary loss functions to stabilize training, ArcFace achieves stable convergence with a simpler implementation and without significantly increasing computational cost. These characteristics have led to its widespread adoption and consolidation as one of the most effective and widely used solutions in face recognition tasks.

The main objective of this work is to evaluate the performance of ArcFace across

Method	**LFW(%)**	**YTF(%)**
VGG Face	98.95	97.30
SphereFace	99.42	95.0
CosFace	99.73	97.6
MS1MV2, R100, ArcFace	**99.83**	**98.02**

Table 2.3: Verification results on the most widely used benchmarks for evaluating unconstrained face verification on still images and videos , LFW and YTF respectively. The results, presented by Deng et al. (2019) [13], show that ArcFace, trained on the MS1MV2 dataset with a ResNet100 backbone, significantly outperforms previous methods such as SphereFace and CosFace across both datasets. These results highlight the effectiveness of the additive angular margin loss in enhancing the discriminative power of deep facial feature representations.

various data subsets in order to identify optimal sampling strategies for biometric image collection. The goal is to define a criterion that allows for a reduction in the amount of training data required by the model, optimizing performance while ensuring compliance with ethical and regulatory standards. The following section presents the sampling methods used in this study to select those data subsets.

2.2 Data Sampling

The following sections review a variety of sampling strategies that have been proposed in existing studies, including random sampling, proportional sampling (random within each identity), similarity-based sampling, distance-based sampling. Other sampling approaches have also been applied to facial recognition, including ethnicity-based and training-based strategies. However, these fall outside the main focus of this work.

2.2.1 Random Sampling

Random subsampling (RS) selects a subset of data under the assumption that each sample has an equal probability of being chosen, ensuring a uniform probability distribution. This approach simplifies the data selection process, making it computationally efficient for managing large datasets, and aims to achieve adequate representativeness and diversity, which can support the generalization capabilities of deep neural networks (DNNs). In facial recognition, where large-scale data is common, this efficiency is particularly beneficial.

RS is often favored for its low computational cost, which makes it an attractive choice for large-scale applications like facial recognition. However, its effectiveness compared to more sophisticated subsampling methods remains debated.

RS is particularly challenging when applied to imbalanced datasets. In [29], authors critique RS for not explicitly accounting for class distribution causing to preserve the original imbalance rather than correcting it. This can result in insufficient representation of minority classes and potential loss of critical information from underrepresented groups. This is especially problematic in facial recognition, where demographic balance is crucial for fairness and model accuracy. As highlighted [64], while RS is computationally practical, it may not provide the most balanced representation of the dataset compared to advanced techniques like stratified sampling, which explicitly manage class balance and diversity. Including comparisons with these alternative methods would better contextualize the strengths and weaknesses of RS.

RS can enhance generalization in some cases by focusing on representative subsets, potentially reducing overfitting and encouraging the learning of broader patterns. However, as shown in [14], this advantage is not universal; the generalization performance of models trained with RS can vary significantly. RS alters the error surface of neural networks, leading to different paths of optimization and the exploration of various local minima. According to the authors in [14], this variability can result in inconsistent convergence behaviors, where some models improve validation performance while others deteriorate depending on the specific subsample used.

In some cases, by reducing the size of the training dataset through RS applied to negative classes during training, as implemented in [3], it is possible to improve performance consistency across demographic groups. While not explicitly designed for fairness, this approach indirectly mitigates biases toward overrepresented classes by balancing the contribution of identiti es during optimization.

2.2.2 Stratified Sampling

Stratified sampling involves dividing data into homogeneous strata based on certain characteristics and selecting samples within each stratum using a specific mechanism. This method aims to ensure that units within a stratum are as homogeneous as possible, allowing a stratified sample, with the appropriate number of units from each stratum, to be representative of the total population [59].

This technique is primarily used in the training of deep neural networks to address class imbalance issues, improve the representativeness of data in training and test sets, and enhance model generalization. In these applications, strata are defined in various ways, depending on the context and the nature of the data.

In classification tasks, classes can be used as strata when the data are labeled, balancing the data between training and test sets to maintain the same structure [28]; sub-sampling the majority classes or over-sampling the minority classes to improve the predictive capacity of the minority classes [24, 34]; and sampling strata that provide more information to the model, focusing the training on areas where the model needs improvement [42].

In other cases, strata can be generated based on certain known variables, such as the confidence levels of the network's predictions or data clustering, allowing the sampling process to adapt to the characteristics and variability of each stratum [65].

Overall, the use of stratified sampling in neural networks is considered a key component for mitigating bias toward majority classes and improving model accuracy. This improvement arises from the ability to apply undersampling or oversampling within strata to achieve better class balance. And even when sampling proportionally within strata, stratified approaches ensure that all classes are represented in the training process, something that cannot be guaranteed through purely random sampling, making them particularly valuable in large-scale applications such as fraud detection, medical diagnosis, and other domains with highly imbalanced class distributions.

2.2.3 Maximin Sampling

This method is based on the selection of sampling points that maximize the minimum distance between them, promoting a uniform distribution of points in the design space. This algorithm is a sequential selection strategy, designed to construct a subset of elements that are maximally dispersed according to a predefined distance measure. Beginning with an arbitrary initialization, the algorithm proceeds iteratively by adding, at each step, the element that is farthest from the closest member of the already selected set, as explained in [7]. In this way, each newly chosen element maximizes its separation from the existing subset, ensuring that the final selection captures

the greatest degree of diversity.

This results in effective coverage and minimizes redundancies in the training of models. This approach has been shown to improve data efficiency by prioritizing informative and diverse samples, which helps improve generalization and reduces the need for large labeled datasets [49]. The method exposes the model to a balanced representation of all regions of the input space [51], which is especially helpful when training networks on high-dimensional problems and complex data structures.

One key benefit of the maximin method is its ability to reduce redundancy in training data. By selecting points that are as far apart from each other as possible, it reduces the likelihood of choosing points that do not provide new information to the learning process of the neural network, thereby improving training times and model generalization [30, 46].

Furthermore, this approach ensures that no region of the space is underrepresented, which contributes to improved model performance by avoiding biases toward specific areas of the space. This is particularly useful in applications such as physical modeling or simulations of complex systems, where experimental data can be costly or difficult to obtain [43].

Compared to random sampling, maximin is more advantageous when the data have a well-defined cluster structure, although it involves an additional computational cost due to the need to calculate distances, which becomes even more significant in high-dimensional spaces [27].

Overall, according to some authors [49], maximin enhances the efficiency of neural network models, enabling the development of more generalizable solutions across a wide range of applications. However, according to [4], it falls short of capturing the dataset's most relevant traits.

2.2.4 k-means Sampling

The k-means method is a clustering technique that organizes data into k predefined homogeneous groups by minimizing the weighted sum of intra-cluster variances, with each cluster represented by a prototype or centroid that serves as a model for its group [41]. This process is iterative, continuing until the centroids converge.

Sampling methods based on the k-means algorithm have been effectively employed to select representative instances from large datasets, preserving estimation accuracy close to that of full-sample approaches while significantly reducing computational cost. [35]. k-means clustering can be considered a good proxy for selecting representative samples because it groups the data into clusters and identifies their centroids. The elements that are closest to these centroids reflect the most typical examples of each group. In this way, k-means could help reduce redundancy while still covering the main structure of the data. When applied to sampling, it allows for the selection of elements that preserve the key characteristics of the original dataset [16].

This method has also been successfully applied in fields such as intrusion detection, where it helps reduce redundant samples within the training datasets [54]. For instance, in intrusion detection, k-means is used to select the most representative sam-

ples, which significantly reduces training time without compromising the quality of the processed information.

In audio data processing [5], the authors note that the way k-means clustering groups similar samples often matches how humans perceive sound. This means it can help uncover hidden patterns in the data and improve how the data is labeled and interpreted.

However, for large datasets, the iterative nature of k-means can result in significant computational cost. While k-means is known for its efficiency in reducing redundancy and summarizing data, its effectiveness can be limited by the initial random placement of centroids or the chosen number of clusters, both of which can influence the quality of the final clustering [50, 67].

Moreover, limitations may arise if the data is not well-distributed or if the number of clusters is not adequately matched to the data's variability. Nonetheless, these challenges can be mitigated when used in conjunction with methods such as Synthetic Minority Oversampling Technique (SMOTE), which helps balance the dataset [17, 54].

Chapter 3

Methodology

This chapter presents the methodological framework adopted for this work, detailing the algorithms and experimental design decisions that support the implementation. The focus is placed on the selected model, as well as on the sampling techniques used to optimize data selection used for training it. Additionally, a dedicated subsection analyzes the characteristics of the embeddings used in certain sampling methods, as they play a key role in the interpretation of the results and supporting the conclusions drawn from the experiments.

3.1 Implementation Details

3.1.1 Model

The experiments in this work are based on the ArcFace model proposed by Deng et al. [13], which serves as the foundation for the subsampling analyses conducted. ArcFace introduces an additive angular margin loss that enhances inter-class separation and intra-class compactness, significantly improving facial recognition performance across diverse scenarios as described in Section 2.1.

For this study, adaptations were made following an unofficial implementation by Kuan-Yu Huang and Ali Fayzi [26], which apply minor variations from the original model. This implementation employs ResNet50 and MobileNet as backbones, pretrained with ImageNet, evaluating each under both CCROP[1] and non-CCROP settings. The authors of this work obtained the results shown in Table 3.1.

Training Parameters

Embedding Dimension: The generated embeddings have a fixed dimensionality of 512, providing rich and discriminative representations for each identity in the feature

[1]CCROP indicates whether a center crop was applied to the image during preprocessing, typically focusing on the central facial region.

space [36, 61, 63, 66].

Angular Margin and Scale: The additive angular margin was set to m=0.5, while the scale factor was configured to s=64. These settings optimize the geodesic distance between features on a hypersphere, promoting class separation [61].

Learning Rate: A learning rate of 0.1 was employed, except in one of the experiments where it was reduced to 0.001, to ensure model convergence.

Batch Size: The batch size was set to 128.

Network Architecture The architecture is based on ResNet50 as the backbone with the incorporation of the additive angular margin layer, which introduces the ArcFace concept.

Convolutional Layers: ResNet50 [23] is a 50-layer deep convolutional neural network that employs residual connections, allowing the input of a layer to bypass one or more intermediate layers and be added directly to the output. The network introduces residual blocks, as shown in Figure 3.1, each consisting of three convolutional layers with batch normalization and a ReLU activation function after each layer. The input is then added to the output of the block, preserving information flow and mitigating vanishing gradient issues. The architecture begins with a convolution and max-pooling stage, followed by four groups of residual blocks with increasing depth and feature dimensionality (3, 4, 6, and 3 blocks, respectively), and concludes with global average pooling and a fully connected layer.

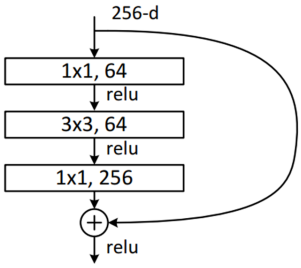


Figure 3.1: Residual block for ResNet50.

L2 Normalization: The output features and class weights are normalized with respect to the L2 norm, ensuring all vectors reside on a unit hypersphere.

Additive Angular Margin: An additive angular margin is applied to the angle between each sample's normalized feature and its corresponding class weight.

Fully Connected Layer and Scaling: The logits resulting from the angular margin operation are rescaled before being passed to the softmax function for classification.

This structure ensures that the network-generated representations remain highly discriminative, even under significant variability in input images.

The training and test datasets are as described in Subsection 3.1.2.

Backbone	CCROP	LFW	AgeDB-30	CFP-FP
ResNet50	FALSE	99.42	95.32	92.56
MobileNetV2	FALSE	99.13	91.62	91.5
ResNet50	TRUE	99.38	95.13	94.87
MobileNetV2	TRUE	98.88	91.58	93.19

Table 3.1: Results on ArcFace with the unofficial implementation by Kuan-Yu Huang and Ali Fayzi [26].

3.1.2 Dataset Description

This study employed four datasets. One for training, one for validation, and two for testing purposes. The dataset used for training includes facial images with high variability in terms of pose, illumination, and age, ensuring the model learns robust and discriminative representations. Additionally, standard benchmarks were used for validation and for testing performance in real-world conditions.

TRAINING DATASET

MS1M-Arcface

- Release Year: 2016
- Description: Considered the world's largest dataset for facial recognition, it contains images of celebrities collected from the internet.
- Number of Images: 5,822,653
- Number of Identities: 85,742
- Note: The MS1M-ArcFace dataset, has been widely used for training facial recognition models. However, it is no longer publicly accessible due to privacy concerns and regulatory issues. Despite these restrictions, it remains a benchmark dataset in facial recognition research[2].

VALIDATION DATASET

Celebrities in FrontalProfile in the Wild (CFP-FP)

- Release Year: 2016
- Description: Pairs of frontal and profile pose images of celebrities.
- Number of Images: 7,000
- Number of Identities: 500
- Download Link: `http://www.cfpw.io/`

TESTING DATASETS

Labeled Faces in the Wild (LFW)

[2] Although MS1M-ArcFace is no longer officially available, in this study, the MS1M-ArcFace is used for experimental purposes, ensuring compliance with data protection regulations and avoiding unauthorized redistribution.

- Release Year: 2007
- Description: A database of faces collected from the web, designed for studying the problem of facial recognition under unconstrained environments[3].
- Number of Images: 13,233
- Number of Identities: 5,749
- Download Link: `http://vis-www.cs.umass.edu/lfw/`

AgeDB30

- Release Year: 2017
- Description: A manually collected dataset in "in-the-wild" settings.
- Number of Images: 16,488
- Number of Identities: 568
- Download Link: `https://ibug.doc.ic.ac.uk/resources/agedb/`

Image Preprocessing

All preprocessing and data formatting steps were implemented following the unofficial ArcFace implementation by Kuan-Yu Huang and Ali Fayzi [26].

Images in all datasets were transformed to the *Align_112_112* format by resizing to 128x128 pixels, applying random horizontal flip, adjusting saturation within limits of 0.6 to 1.4, adjusting brightness within limits of -0.4 to 0.4, and finally normalizing the values.

The training set, consisting of 5.8 million images, was converted to the *tfrecord* format, a binary storage format that allows for more efficient image processing. Meanwhile, the test sets were downloaded in *blp* format, an optimized compression for binary data, reducing dataset size and speeding up processing.

3.1.3 System Setup

Experiments involving ImageNet embeddings were conducted in a Jupyter Notebook environment hosted on a virtual machine instance (*deeplearning-2-vm*) on Google Cloud Platform. The machine was configured as an e2-standardx16 with 16 virtual CPUs (running on Intel Broadwell-based physical hardware) and 16 GB of RAM. It featured a 250 GB balanced persistent disk using a SCSI interface. The system was based on Debian GNU/Linux 10 (buster), running kernel version 4.19.0-25-cloud-amd64 on an x86_64 architecture.

All other experiments were conducted on a local desktop PC. This machine was equipped with a 13th Gen Intel® Core™ i5-13400F processor (16 threads), 64 GB of RAM, and an NVIDIA GeForce RTX™ 3060 GPU (used only in a subset of tests). The storage setup included a 447.1 GB SATA SSD and a 3.6 TB external USB HDD, for a total capacity of 4 TB. The system ran Ubuntu 24.04.1 LTS, with kernel version 6.8.0-52-generic on an x86_64 architecture.

[3]The images are collected in real world conditions, without controlled settings.

3.2 Experimental Design

These experiments explore the impact of various subsampling strategies on model performance by evaluating multiple subsets drawn from the original training dataset. Each subsample is generated using a distinct sampling methodology, enabling a comparative analysis of verification accuracy and the identification of optimized data collection patterns. For the majority of experiments, 10% of the original dataset—approximately 582,000 images—was used, except in cases where the selected identities did not contain a sufficient number of images or where specific test configurations required adjustments. The exact number of images varied slightly depending on the sampling technique and rounding procedures applied during the selection process.

3.2.1 Sampling Methods

Given that the primary objective is to identify sampling strategies that could be applied a priori in real-world facial data collection scenarios, the selected methods prioritize different criteria, including data representativeness, randomness, the number of images per identity, and the total number of distinct identities.

The experiments are designed to assess whether these subsampling approaches yield statistically significant differences in model performance, thereby informing best practices for efficient and effective data acquisition. The chosen sampling techniques were selected based on their relevance and potential applicability within the context of biometric data collection frameworks, particularly when facing restrictions on data quantity or identity-level consent. A major motivation behind this work is to explore how the number of identities required can be reduced, because of said data collection restrictions, without significantly compromising model performance. As a result, several experiments focus specifically on identity selection.

Additionally, some sampling methods incorporate oracle-type information that may not be available in real deployment scenarios. While this may appear redundant or unrealistic, the intention is to investigate whether there are inherent patterns or relationships in the data that could inform more effective data selection strategies. This approach helps uncover potential signals that could guide the development of more practical and approximate strategies for selecting data in future applications.

Sampling on Images

Uniform Random Sampling of Images

In this case, simple random sampling was used to select 10% of the images from a facial recognition database. This method involves selecting a fraction of the dataset at random, without taking into account any underlying characteristics of the data. This ensures that each image in the original dataset has the same probability of being selected, meaning the sampling is not biased by factors such as facial characteristics or class membership.

From a statistical perspective, random sampling ensures that the selected subset retains the same general characteristics as the complete set. However, since samples

are selected independently of visual features, there is a risk that some important classes or variations within the original data may not be represented in the selected subset, which may also lead to a loss of crucial image variations. The implementation details of this sampling method are provided in Algorithm 1.

Performed Tests: 10% randomly selected images from the complete dataset.

Algorithm 1 Random Sampling of 10% of Images

Input: Image dataset $\mathcal{D}$, subset size $s = 0.10 \times |\mathcal{D}|$
Output: Subset of images $\mathcal{S}$

Initialize an empty set $\mathcal{S}$ of size s to store selected images
Generate a list of unique random indices R of size s such that $R \subseteq \{1, 2, \ldots, |\mathcal{D}|\}$
for each index i in R **do**
    Select the image corresponding to position i in $\mathcal{D}$
    Add the selected image to the set $\mathcal{S}$
end for
return Subset $\mathcal{S}$

Stratified Sampling

Stratified sampling is a technique where the dataset is divided into groups or "strata". In this context, the identities are treated as strata because no additional metadata (such as ethnicity, gender or age) is available for further stratification, and each identity is assumed to be internally homogeneous. A subset of images from each identity is selected proportionally to the number of images that identity holds, until 10% of the total dataset is reached. The goal is to ensure that each stratum is represented in the final sample according to its proportion in the original dataset.

The process involves two main steps: first, calculating the proportion of images to be selected from each identity based on the number of images they contain relative to the total dataset; second, randomly selecting the images from each identity according to that proportion. Finally, a subset containing 10% of the total images is constructed, respecting the distribution of the identities. Details of the implementation are further given in Algorithm 2.

This method ensures that the proportion of images per identity is maintained in the subset, which is beneficial to avoid bias in class representation. It ensures that identities with more images remain well-represented in the subset, while those with fewer images are not overlooked in the sampling process.

Performed Tests: 10% randomly selected images from each identity in the same proportion they are in the original dataset.

k-Means on Images

In this case, the k-means algorithm is applied to the embeddings of images, as presented in Subsection 3.2.4, within each class to form clusters. Subsequently, the images closest to the centroids of these clusters are selected as representatives of each group. This is based on the assumption that selecting images closest to the centroids will capture the most relevant diversity within each class.

Algorithm 2 Stratified Sampling of 10% of Images

Input: Image dataset $\mathcal{D}$, subset size $s = 0.10 \times |\mathcal{D}|$, number of images in identity i n_i

Output: Subset of images $\mathcal{S}$

Identify the set of unique identities $\mathcal{I}$ in $\mathcal{D}$
for each index i in $\mathcal{I}$ **do**
    Randomly select $0.10 \times n_i$ images from n_i images in identity i
    Add the selected images to the set $\mathcal{S}$
end for
return Subset $\mathcal{S}$

Given that the average number of images per class is 68, we computed 7 clusters within each class using the k-means algorithm. From each cluster, the image closest to the centroid was selected to approximate 10% of the images per class, thus achieving a representative 10% sampling of the entire dataset. Algorithm 3 details the specific steps of this method. This approach yielded a slightly smaller number of images due to the selection of 7 clusters per class, as not all classes contain at least 7 images or clusters.

Applying this method to each identity yields a subset of images that represents the structural diversity within the class, in terms of intra-class variability, and helps eliminate redundancies

Performed Tests: 10% selected images embeddings, using k-means within each identity, over all identities.

Algorithm 3 Sampling of 10% of Images using k-Means over classes

Input: Full dataset embeddings $\mathcal{E}$ with unique identities $\mathcal{I}$, subset size $s = 0.10 \times |\mathcal{E}|$

Output: Subset of images $\mathcal{S}$

Calculate the number of clusters $k = 0.10 \times \frac{|\mathcal{I}|}{|\mathcal{E}|}$
for i in $\mathcal{I}$ **do**
    Calculate k-Means with k clusters over $\mathcal{E}_i$
    Select the images closest to each cluster centroid
    Add the selected images to the set $\mathcal{S}$
end for
return Subset $\mathcal{S}$

Comparison of image sampling strategies applied to a toy dataset

To illustrate the behavior and implications of different image sampling strategies, a simple toy example is shown in Figure 3.2 using three alternative selection methods. The first visualization depicts uniform random sampling applied across the entire dataset, serving as a baseline agnostic to class distribution. The second plot demonstrates stratified sampling, where samples are drawn from each class in proportion to their frequency in the original dataset, ensuring that all classes are represented according to the underlying data distribution. Finally, the third plot presents k-means sampling within each identity, where clustering is performed independently for each class, with k being the number of images to select from each class, and the samples

closest to the cluster centroids are selected.

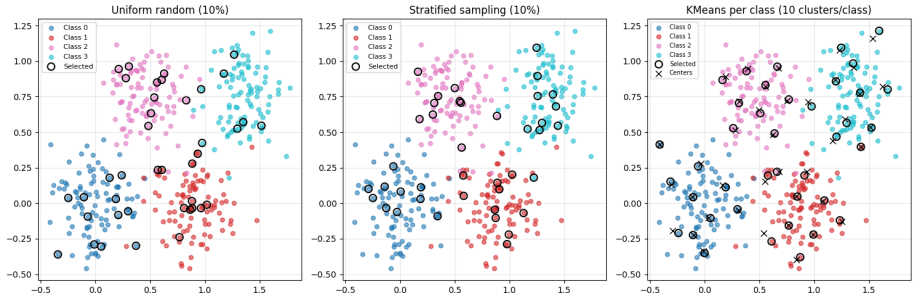


Figure 3.2: Comparison of image sampling strategies applied to a toy dataset. From left to right: uniform random sampling across the entire dataset; stratified sampling, maintaining original class distribution; and k-means sampling within each identity, where clustering is used to select representative samples closest to cluster centroids.

Sampling on Identities

Selection of Identities with the Most and Least Images

Both methods are based on selecting a subset of identities, but they differ in their selection criteria. In the first approach, identities with the highest number of images were prioritized, and images were sampled until 10% of the dataset was reached, while in the second approach, 10% of the total images were selected from the identities with the fewest images.

The selection of images from identities with the highest number of images provides the smallest possible number of identities from the dataset while maintaining the 10% subsample. However, this approach does not contribute to the generalization of the model.

On the other hand, selecting images from identities with the fewest images gives greater weight to underrepresented identities, but valuable information from the dominant classes used to train the model may be lost in the process.

However, these tests, outlined in Algorithm 4, give us a benchmark of model behavior on identities with more and less representativity in the original dataset.

Performed Tests: 10% of the total images, selecting them either from the identities with the most images or from those with the least.

Uniform Random Sampling of Identities

In this instance, a random subset of identities from the database is first selected until the total of 10% of the dataset is reached. Across ten test scenarios, the proportion of selected identities varied, ranging from 5% to 50% of all classes, to evaluate the impact of identity subset size on sampling performance.

This approach ensures that the subset contains images representing a diversity of identities while maintaining the total number of images at a reduced level. This

Algorithm 4 Sampling of 10% of Images from Identities with the Most/Least Images

Input: Image dataset $\mathcal{D}$, subset size $s = 0.10 \times |\mathcal{D}|$
Output: Subset of images $\mathcal{S}$

Identify the set of unique identities $\mathcal{I}$ in $\mathcal{D}$
Sort unique identities $R = \mathcal{I}$ in descending/ascending order of the number of images
while $|\mathcal{S}| =< s$ **do**
    Select all images in order
    Add the selected images to the set $\mathcal{S}$
end while
return Subset $\mathcal{S}$

method guarantees that image selection within each identity reflects the distribution of the original dataset. Thus, identities with more images will be better represented in the subset, which may improve model performance by maintaining class proportionality. However, identities with fewer images may contribute very few samples, potentially reducing the model's robustness for those identities.

The purpose of these tests is to evaluate how the trade-off between the number of identities and the number of images per identity affects model accuracy when the selected identities are chosen at random. By varying the proportion of identities while keeping the total number of images constant, the experiment aims to evaluate whether diversity or depth in identities, has a greater influence, and to compare these effects with those obtained from alternative sampling strategies. The procedure used to apply this sampling strategy is shown in Algorithm 5.

Performed Tests: Ten ten test scenarios were conducted. In each scenario, 10% of the total images (approximately 582.000 images) were proportionally selected from a randomly chosen subset of identities. The size of this identity subset varied across experiments, representing 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, and 50% of the total number of classes in the dataset.

Algorithm 5 Sampling of 10% of Images from p% classes

Input: Image dataset $\mathcal{D}$, subset size $s = 0.10 \times |\mathcal{D}|$, identities percentage p
Output: Subset of images $\mathcal{S}$

Identify the set of unique identities $\mathcal{I}$ in $\mathcal{D}$
Calculate the number of identities to take $n = p \times |\mathcal{I}|$
Generate a list of unique random indices R of size n such that $R \subseteq \{1, 2, \ldots, |\mathcal{I}|\}$
Calculate the proportion of images to take from each identity $a = \frac{s}{\sum_1^R n_r}$
for each index i in R **do**
    Randomly select $a \times n_r$ images
    Add the selected images to the set $\mathcal{S}$
end for
return Subset $\mathcal{S}$

k-Means on Identities

This sampling strategy follows the same general rationale as the Uniform Random Sampling of Identities approach, but instead of selecting identities purely at random, it introduces a structured criterion based on similarity in the embedding space. In this context, k-means clustering sampling serves as a proxy for selecting representative identities in real-world scenarios where data acquisition aims to capture the most informative samples.

Each identity is represented by the mean embedding of its images, as previously shown in Subsection 3.2.4, and these representations are clustered using k-means, with the number of clusters corresponding to the desired number of identities. Finally, images from the identities closest to the cluster centroids are selected proportionally until 10% of the total data is reached. Implementation steps can be found in Algorithm 6.

This procedure allows for grouping identities in a way that those selected represent the common and distinctive facial features in the dataset, capturing global variations and eliminating redundancies.

The quality of the selected subset is dependent on how well the clusters are formed and initialized. If the clustering is not optimal, the selected identities may not capture all the relevant variability.

Performed Tests: Experiments conducted using 10% of the images, proportionally selected from identity subsets representing 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, and 50% of the total classes. These subsets were selected using k-means clustering over the mean embeddings of each identity.

Algorithm 6 Sampling of 10% of Images using k-Means over classes

Input: Image dataset $\mathcal{D}$, Full dataset embeddings $\mathcal{E}$ with unique identities $\mathcal{I}$, subset size $s = 0.10 \times |\mathcal{D}|$, identities percentage p
Output: Subset of images $\mathcal{S}$

Calculate the number of clusters $k = p \times |\mathcal{I}|$
for i in $|\mathcal{I}|$ **do**
    Calculate the mean embedding $\mathcal{E}_i$
    Add the selected embeddings to the set $\mathcal{E}_I$
end for
Calculate k-Means with k clusters over $\mathcal{E}_I$
Select the k classes closest to each cluster centroid and store the embeddings in $\mathcal{E}_k$
Calculate the proportion of images to take from each identity $a = \frac{s}{\sum_0^{|\mathcal{E}_k|} n_k}$
for i in $|\mathcal{E}_k|$ **do**
    Randomly select $a \times n_i$ images from $\mathcal{D}_k$
    Add the selected images to the set $\mathcal{S}$
end for
return Subset $\mathcal{S}$

Greedy Maximin on Identities

This sampling strategy applies a structured criterion in the embedding space to maximize identity diversity, with the objective of capturing the widest possible range of identity variability with a limited number of samples.

The Greedy Maximin method is an approach used to select a subset of identities based on the distances between the mean embeddings of identities, as defined in Subsection 3.2.4, prioritizing those that maximize the distance from previously selected identities.

For consistency, the initialization was fixed at the identity indexed at zero, as the choice of starting point does not significantly affect the statistical outcome of the selection, as stated in [27], then a percentage p% of identities is iteratively selected. In each step, the identity that maximizes the minimum distance to the already selected identities is chosen, ensuring that the selected identities are as distant as possible from each other in the embedding space. Once the identities are selected, the images are proportionally selected based on the number of images per identity until 10% of the total data is reached.Its implementation follows the structure illustrated in Algorithm 7.

By varying the value of p%, the number of selected identities can be controlled, allowing us to study how the number of selected identities affects the representation of the images, while keeping the total number of images fixed at 10%.

Performed Tests with ArcFace Embeddings: Ten configurations were tested in which 10% of the images were proportionally selected from identity subsets chosen using the Greedy Maximin algorithm. These subsets included 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, and 50% of the total classes.

Performed Tests with ImageNet Embeddings: Seven experiments conducted, in which 10% of the images were proportionally selected from identity subsets chosen using the Greedy Maximin algorithm, based on ImageNet embeddings. The selected class subsets represented 10%, 15%, 20%, 25%, 30%, 35%, and 40% of the total number of identities.

Additional Experiments

To better understand the effects of **distance-based sampling strategies**, three additional selection methods were employed: (1) selecting the 10% of identities closest to the overall dataset mean, (2) selecting the 10% farthest from the mean, and (3) selecting the 10% of identities closest to the first identity in the dataset (taken as a random initialization), based on embedding distance. These methods were designed to examine how the distribution of identity embeddings relative to a reference point influences model performance.

In parallel, to evaluate the impact of using full identity information under **data quantity constraints**, a second set of experiments was conducted. In this case, 15% of the total identities were selected using the same strategies as before (Maximin sampling based on ImageNet and ArcFace embeddings, as well as Random Sampling) but all available images for each selected identity were included, rather than limiting the dataset to 10% of total images (approximately 582,000 images). This setup enables a direct comparison between shallow sampling under data-limited conditions, and deep sampling within the same identities, helping to isolate the effect of identity depth on model performance.

Finally, to inform future research and practical deployment scenarios, each sampling strategy was also evaluated under a **class-balanced setup**, in which the same number of images was selected from each identity (whenever possible). This contrasts

Algorithm 7 Sampling of 10% of Images using Greedy Maximin over classes

Input: Image dataset $\mathcal{D}$, Full dataset embeddings $\mathcal{E}$ with unique identities $\mathcal{I}$, subset size $s = 0.10 \times |\mathcal{D}|$, identities percentage p

Output: Subset of images $\mathcal{S}$

Calculate the number of classes to take $c = p \times |\mathcal{I}|$
for i in $|\mathcal{I}|$ **do**
    Calculate the mean embedding $\mathcal{E}_i$
    Add the selected embeddings to the set $\mathcal{E}_I$
end for
Calculate d_{IxI} as the distance matrix of $\mathcal{E}_I$
Select the initial identity $\mathcal{I}_0$ and add it to the permutation set $\mathcal{E}_c$
for each remaining class identity j up to c **do**
    Find the identity $\mathcal{I}_j$ furthest from all identities currently in $\mathcal{E}_c$
    Add $\mathcal{I}_j$ to the permutation set $\mathcal{E}_c$
    Update the covering radius λ_{j-1} as the maximum distance from $\mathcal{I}_j$ to any identity in $\mathcal{E}_c$
    Update the minimum distances from each remaining identity to the set $\mathcal{E}_c$
end for
Calculate the proportion of images to take from each identity $a = \frac{s}{\sum_0^{|\mathcal{E}_c|} n_c}$
for i in $|\mathcal{E}_c|$ **do**
    Randomly select $a \times n_i$ images from $\mathcal{D}_i$
    Add the selected images to the set $\mathcal{S}$
end for
return Subset $\mathcal{S}$

with the original experiments, where images were sampled proportionally to their occurrence in the full dataset. By enforcing equal representation across 35% of identities, this analysis aims to isolate the effects of class imbalance from those of sample selection.

Comparison of identity sampling strategies applied to a toy dataset

To provide an intuitive understanding of the identity-level selection strategies evaluated in this study, a simple toy example is shown in Figure 3.3 comparing three representative approaches. The first visualization illustrates uniform random sampling of identities, where identities are selected without considering their relationships in the embedding space. The second example applies the Greedy Maximin method, which prioritizes selecting identities that are maximally distant from each other, thereby promoting diversity and wide coverage of the feature space. Finally, the third example demonstrates k-means identity sampling, where clustering is performed on the mean embeddings of each identity, and those closest to the cluster centroids are chosen as representative. The different spatial coverage offered by each sampling method can be visualized in the plots.

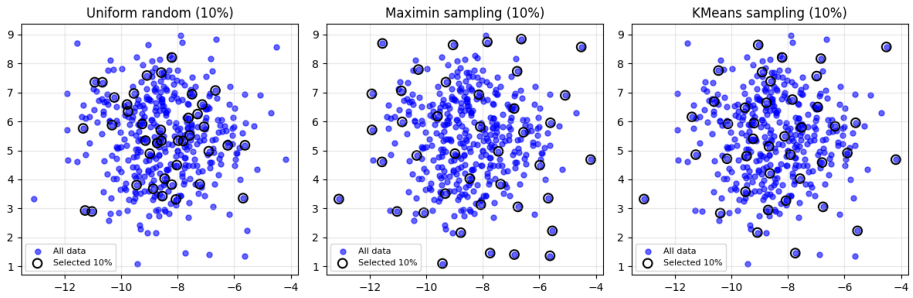


Figure 3.3: Comparison of identity sampling strategies applied to a toy dataset. From left to right: uniform random sampling, Greedy Maximin, and k-means identity sampling, where each method exhibits a distinct spatial coverage of identities.

3.2.2 Validation Protocol

The CFP-FP dataset was used as a validation benchmark to select the best-performing model. Rather than applying early stopping, the model was trained for a fixed range of 15 epochs. Among these, the checkpoint corresponding to the highest verification accuracy on CFP-FP was selected. The model parameters from this epoch were then used to evaluate performance on the LFW and AgeDB-30 datasets. This procedure ensured that the final results reflected the configuration with the strongest generalization performance on CFP-FP, without influencing the training process itself.

3.2.3 Limitations

Due to time and computational resource constraints, each experimental configuration was executed only once. This decision prevents the estimation of inter-run variability measures, such as standard deviation or confidence intervals, and limits the possibility of conducting formal statistical significance tests between methods. To mitigate this limitation, the analysis was complemented with additional metrics, including Receiver Operating Characteristic (ROC) curves and the calculation of the Area Under the Curve (AUC). The consistency observed in the shape of the ROC curves and their AUC values is used as supplementary evidence to support method comparisons, even in the absence of multiple experimental repetitions.

3.2.4 Embeddings

For k-Means clustering, the greedy maximin algorithm, and data visualization, a reduced feature vector of the images (embeddings) was used. These embeddings were generated using two distinct methods:

1. Pretrained MobileNetV2 on ImageNet: This model was used to obtain a feature vector that remains independent of the specific model used in this study. This approach provides a general embedding that is not biased toward facial recognition data.

2. ResNet50 with ArcFace layer, without softmax: The facial recognition model used in this study, ResNet50 with an ArcFace layer and the original training weights, excluding the softmax layer, was employed. This method generates a feature vector representative of a space specialized in facial features.

Despite the redundancy of using ArcFace, it was included to evaluate the effect of a model specifically trained on faces in the resulting feature space. In both cases, the resulting embedding for each image has dimensions of (1,512).

We expect the embeddings obtained from ArcFace to exhibit higher discriminative power, as they are derived from a model explicitly trained for facial recognition and optimized for identity separation in the embedding space. In contrast, the embeddings produced by the ImageNet-pretrained MobileNetV2 are not specialized for this task and are therefore expected to show lower discriminability among different identities. To evaluate this assumption, the intra-class and inter-class distance distributions in both embedding spaces are analyzed, allowing us to compare how well each representation distinguishes between images of the same and different identities.

To analyze the distribution of these feature vectors as a representation of the data, basic statistical measures were computed, and histograms were generated for intra-class and inter-class distances of the embeddings.

For statistics and histograms, the global average distance of all pairwise distances between points in each class was calculated. The statistics are presented in Table 3.2. Figures 3.4 and 3.5 show, for both methods, histograms of global average intra-class distances, compared to global average inter-class distances for three identities, depicting a greater separability in ArcFace embeddings compared to ImageNet. For ArcFace embeddings, the Kolmogorov–Smirnov test statistic ranged from 0.99 to 1.0, indicating near-complete separability between intra- and inter-class distance distributions. For ImageNet embeddings, the statistic ranged from 0.22 to 0.53, reflecting moderate separability. This implies that distance-based sampling techniques are more informative when applied to ArcFace embeddings. However, in both cases the p-value was effectively zero, confirming that the differences between the distributions are statistically significant.

Dist.	**ImageNet**		**ArcFace**	
	Intra	**Inter**	**Intra**	**Inter**
Min.	0	$2.96E-06$	0	$1.56E-06$
Max.	12.3474	15.3829	36.3245	53.9541
Mean	4.4839	1.9930	23.6595	35.4338
Var	0.6321	0.6578	10.4802	9.1640

Table 3.2: Statistics for Intra-class and Inter-Class distances for ImageNet and ArcFace Embeddings.

Furthermore, the Uniform Manifold Approximation and Projection (UMAP) [38] technique was employed for dimensionality reduction to enable the visualization of the distributions and selected classes in the feature space. UMAP is a non-linear dimensionality reduction algorithm that constructs a graph representation of the data and then optimizes a low-dimensional embedding while maintaining the original data topology, preserving both local and global structures of high-dimensional data while optimizing for interpretability in lower-dimensional spaces. The first type of visual-

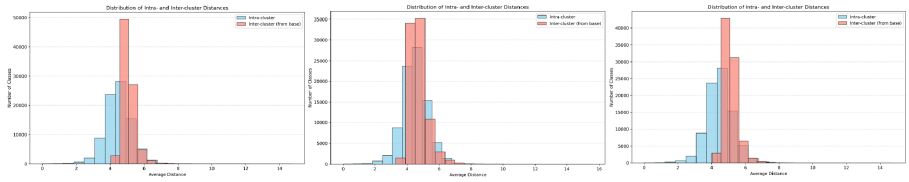


Figure 3.4: Feature space distributions for ImageNet embeddings. Blue columns represent global average of Intra-Class distances, and red columns global average of Inter-Class distances for 3 different randomly selected classes.

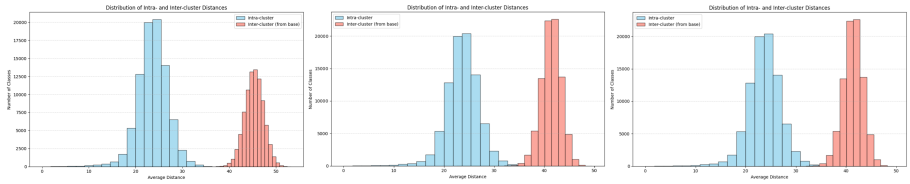


Figure 3.5: Feature space distributions for ArcFace embeddings. Blue columns represent global average of Intra-Class distances, and red columns global average of Inter-Class distances for three different randomly selected classes.

ization is shown in Figure 3.6, which highlights in red the 8,745 selected classes (10% of total IDs in the dataset), obtained using Random sampling methods on the embeddings extracted from ImageNet and ArcFace respectively. This comparison allows to evaluate how different feature extraction methods influence the structure of the embedding space, as well as data selection techniques.

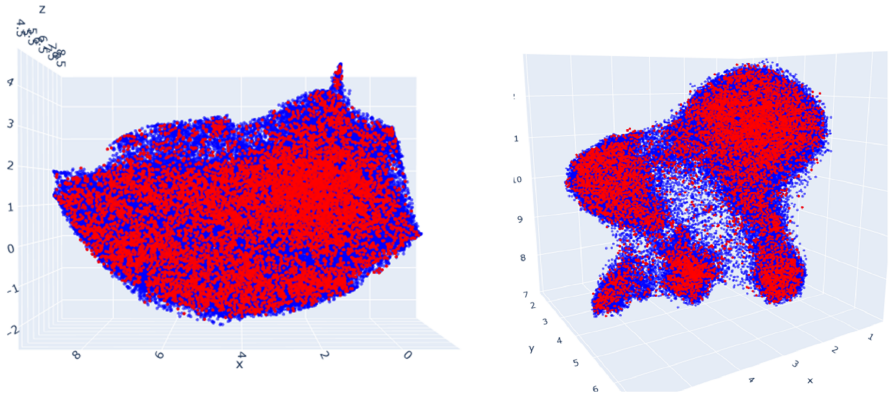


Figure 3.6: UMAP projection of ImageNet Embeddings (left) and Arcface Embeddings (right), highlighting in red the 8,745 classes randomly selected.

Lastly, the second visualization, in Figure 3.7, presents the distribution of embeddings for the entire set of images corresponding to 20 selected classes from the dataset, both for ImageNet (left) and ArcFace (right). Notably, the visualization reveals that embeddings obtained from ArcFace provide a greater degree of class separability, making inter-class boundaries more distinct. In contrast, embeddings derived from ImageNet do not exhibit the same level of separation, suggesting that models explicitly trained for facial recognition, such as ArcFace, generate feature representations that

enhance identity discrimination. These results are consistent with our expectations, confirming that embeddings produced by a task-specific model (face discrimination) are more separable than those obtained from a general-purpose model (generic image discrimination).

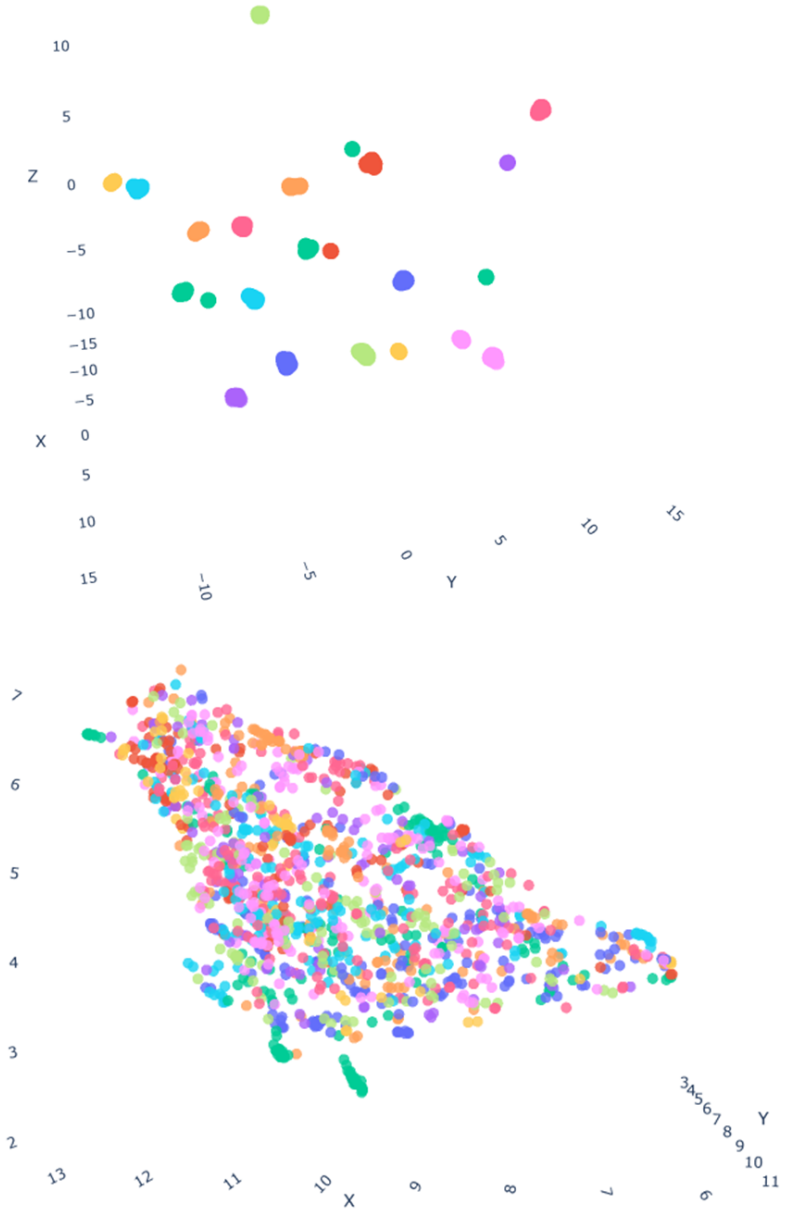


Figure 3.7: Twenty classes Embeddings from ArcFace (top) and ImageNet (bottom).

Chapter 4

Results

A total of 53 experiments are conducted, with model performance evaluated using accuracy metrics on the LFW, and AgeDB-30 datasets, using CFP-FP for optimal epoch and iteration selection as explained in Subsection 3.2.2.

The results are categorized into two groups: experiments where sampling was applied at image level, and those where it was applied at class (identity) level. The former aims to assess the impact of data reduction independently of identities, while the latter explores methods for reducing the number of identities with minimal loss in model accuracy.

4.1 Sampling on Images

Uniform random sampling, stratified sampling, and k-means clustering within each identity were applied to the full set of 85,742 identities. Table 4.1 presents the optimal epoch and step selected through validation, as well as the test results for each experiment.

Stratified sampling yields a modest improvement over uniform random sampling on the CFP-FP and LFW datasets but performs worse on AgeDB-30. This suggests that while stratified sampling may better preserve certain subgroup distributions, it is less effective in capturing age-related variability. In contrast, k-means sampling achieves the highest performance across all benchmarks, pointing to potential redundancy within classes that clustering helps mitigate. The latter could be visualized on the comparison of Figures 4.1 and 4.2.

Additionally, embeddings derived from ArcFace consistently outperform those based on ImageNet, underscoring the superior representational capacity and generalization ability of ArcFace in facial recognition tasks.

Furthermore, when comparing the k-means with ArcFace embeddings sampling scenario to the full-dataset training, we observe that reducing the data volume by

Description	Images	IDs	Validation		Test on Optimal Step	
			Epoch-Step	CFP-FP	Acc LFW	Acc AgeDB30
Uniform Random Sampling	582,265	85,742	$E14 - 3876$	88.13	98.1	84.75
Stratified Sampling	543,686	85,742	$E13 - 1036$	89.34	98.2	83.68
k-means on images (IN)	594,312	85,742	$E15 - 988$	90.44	98.48	87.52
k-means on images (AF)	593,605	85,742	$E15 - 1998$	**92.03**	**98.72**	**87.53**
Full Database	5,822,653	85,742	N/A	94.87	99.38	95.13

Table 4.1: *Validation and test performance for different image-based sampling strategies*. Each row shows a different data selection method used to train the ArcFace model with MS1M-Arcface dataset. All methods use approximately 10% of the complete dataset, except for the full database test (last row), presented as an upper bound reference. The "Images" and "IDs" columns report the number of images and identities used respectively. The "Validation" columns show the epoch and training step at which the best validation performance on the CFP-FP benchmark was achieved, along with its corresponding accuracy. The "Test on Optimal Step" columns report the model test accuracy on LFW and AgeDB-30 benchmarks, using the checkpoint selected via CFP-FP validation. k-means was applied on both ImageNet (IN) and ArcFace (AF) embeddings. Results indicate that using k-means clustering on ArcFace embeddings leads to the best performance, with the largest performance gap compared to the full dataset appearing on the AgeDB-30 benchmark.

Figure 4.1: Example - Complete set of images in the selected ID.

Figure 4.2: Example - k-means selected images. The selected subset demonstrates how clustering minimizes redundancy by preserving representative facial variations from the full set in Figure 4.1.

90% results in only a minor decrease in accuracy, approximately 1% for LFW, 3% for CFP-FP, and at most 8% for AgeDB-30.

These findings highlight the effectiveness of sampling based on clustering as a strategy for data reduction when there are no restrictions on identity selection. This method can capture key structural patterns within each identity and select a representative subset of images that allows for a meaningful reduction in data volume while maintaining the diversity and information necessary for acceptable model performance. This embedding-guided approach relies on oracle-like information and may not be directly applicable in real-world biometric data collection scenarios, yet it provides valuable insights insights into how more practical sampling methods could be designed, where **selecting the most representative images from a group provides greater informational value and efficiency**.

4.2 Sampling on Classes

Class sampling experiments aimed to determine both the most effective identity selection methods and the optimal number of identities for model training, under the constraint of a fixed total number of images.

To evaluate identity selection methods, the following were used: uniform random class selection, representative class selection via k-means, and selection of the most distinct identities via Maxmin. Results of these methods are primarily presented in Subsection 4.2.1 for 30% of identities.

Additionally, in Subsection 4.2.2, for the previously described random, Maximin, and k-means sampling methods, experiments were conducted varying the number of selected classes.

Further tests were designed to evaluate the impact of class quantity by selecting identities with the highest or lowest number of images, while maintaining a fixed dataset size. Also, to examine the influence of data distribution, extreme selection strategies were applied, including choosing images furthest from and closest to the dataset mean, as well as a minimum-distance selection from class zero, functionally analogous to an inverse Maximin approach. Additional experiments were also performed to assess the effects of class balance compared to proportional sampling.

4.2.1 Identity Selection Methods

To evaluate selection techniques, a subset comprising 30% of the identities in the original dataset was used, yielding a subset of 25,722 classes as an initial base. The tested methods included uniform random sampling, representative identity selection using k-means clustering, and Maximin sampling based on inter-class distance. For both k-means and Maximin approaches, selection was performed using the mean embedding vector of each identity. The embedding vectors used for calculation were also obtained as explained in Subsection 3.2.4.

Table 4.2 presents the optimal epoch and step selected through validation, as well as model performance in tests for each method.

Description	Images	IDs	Validation		Test on Optimal Step	
			Epoch-Step	CFP-FP	Acc LFW	Acc AgeDB30
k-means on IDs - 30 % (IN)	582,139	25,722	$E10-4077$	90.66	98.15	86.6
k-means on IDs - 30 % (AF)	582,102	25,722	$E14-2889$	**91.99**	98.42	87.22
Maxmin on IDs - 30 % (IN)	581,558	25,722	$E14-2941$	90.81	98.12	86.4
Maxmin on IDs - 30 % (AF)	582,304	25,722	$E14-1863$	91.47	98.23	85.38
Random Sampling on IDs - 30 %	581,901	25,722	$E14-3902$	91.4	**98.6**	**87.32**

Table 4.2: *Validation and test results for method comparison in class based sampling*. To compare purely the effect of different methods, each row shows the results for a fixed 30% of all classes (approximately 26k identities) evaluated with Maximin, and k-means (both based on ImageNet and ArcFace embeddings) and Random Sampling. Results show k-means with ArcFace embeddings achieve the best CFP-FP score, while Random Sampling performs slightly better on LFW and AgeDB-30 tests.

Description	Images	IDs	Validation		Test on Optimal Step	
			Epoch-Step	CFP-FP	Acc LFW	Acc AgeDB30
IDs with highest number of images	582,265	3,598	$E4-4356$	62.16	75.2	53.93
IDs with lowest number of images	582,265	24,930	$E13-3424$	90.21	98.03	85.4

Table 4.3: *Validation and test results for classes with the highest and lowest number of images*. Both subsets contain the same total number of images (582k) but differ in identity diversity. Models trained on more identities with fewer images each significantly outperform those trained on fewer, overrepresented identities, highlighting the importance of inter-class diversity over intra-class volume.

For the CFP-FP dataset, k-means sampling with ArcFace embeddings yields the highest performance. However, for the LFW and AgeDB-30 datasets, uniform random sampling results in better outcomes. These findings indicate that none of the tested identity selection methods consistently outperforms the others across all benchmarks for this number of selected identities. Also, selection techniques based on embeddings also tend to perform slightly better when using ArcFace embeddings compared to those generated with ImageNet.

4.2.2 Evaluating Class Quantity

Identities with the Highest and Lowest Image Counts

Training with multiple images of the same individual helps the model learn variations in lighting, pose, and expression. Conversely, training with a large number of identities promotes learning the overall variability and heterogeneity of the population. This experiment was conducted to assess the trade-off between these two factors under a fixed training data budget. Model performance was analyzed by selecting identities based on image frequency: one subset included the 3,598 identities - 4% of total in the dataset - with the most images, while the other consisted of the 24,930 - 29% of identities - with the fewest.

As shown in Table 4.3, training on the identities with the highest image count, produce the weakest results, notably lower than those obtained in prior experiments. This result suggests limited generalization when the model is trained on a small number of overrepresented classes (representing only 4% of all identities), which may not adequately reflect the overall feature distribution. In contrast, the performance obtained using the least-represented identities closely matches that of the broader 30%

subset in Table 4.2. **These extreme-case experiments confirm the widely accepted understanding that the number of identities is more critical than the number of images per identity**. When identities with the largest number of images are selected, the model risks overfitting to dominant classes. Conversely, selecting identities with fewer images increases diversity and promotes generalization, though at the expense of intra-class robustness.

Varying Class Quantity

In this series of experiments, the number of identities was systematically varied while keeping the total image count approximately constant (10% of images in the full dataset), using Maximin, k-means and random selection strategies. **The goal was to determine whether distance-based or representative sampling provides measurable benefits over purely random selection across different levels of class diversity**.

Identity proportions ranged from 5% to 50% of all identities in the dataset, covering a spectrum of class diversity, always with approximately 582,000 images in total. The subsets were constructed in a nested manner, such that each smaller subset is fully contained within the larger ones (e.g., the 45% subset is contained within the 50% subset, the 40% within the 45%, and so on). Given the consistently superior performance of ArcFace embeddings compared to ImageNet, only a limited number of trials using ImageNet embeddings were performed, and only for the Maximin strategy and for comparative purposes.

The overlap of identities coincidentally selected by Maximin (AF), k-means (AF), and random sampling for 15% identities, is illustrated in the Venn diagram shown in Figure 4.3.This diagram reveals that the three sampling strategies capture largely distinct portions of the identity space. Out of 12,861 identities per subset, less than 3% are common to all methods. Pairwise overlaps are also low, nearing 18% between Maximin and k-means, and 15% between random sampling and each of the other two methods. These low intersections confirm that these strategies are non-redundant and complementary in exploring the dataset.

Results are presented in Tables 4.4, 4.5, 4.6 and 4.7.

Description	Images	IDs	Validation		Test on Optimal Step	
			Epoch-Step	CFP-FP	Acc LFW	Acc AgeDB30
Maxmin on IDs - 10 % - (IN)	377,605	8,574	$E15-1700$	85.79	95.57	79.33
Maxmin on IDs - 15 % - (IN)	582,464	12,861	$E15-1300$	87.87	96.13	81.25
Maxmin on IDs - 20 % - (IN)	582,625	17,148	$E11-3490$	89.93	97.57	83.27
Maxmin on IDs - 25 % - (IN)	586,642	21,435	$E13-3004$	90.27	98.07	86.62
Maxmin on IDs - 30 % - (IN)	581,558	25,722	$E14-2941$	90.81	98.12	86.4
Maxmin on IDs - 35 % - (IN)	582,134	30,010	$E12-3983$	91.04	**98.45**	86.15
Maxmin on IDs - 40 % - (IN)	582,426	34,297	$E13-2400$	**91.41**	98.3	**87.47**

Table 4.4: *Validation and test results for Maximin on Imagenet embeddings*. Experiments were conducted using identity subsets ranging from 10% to 40% of the total dataset, with the total number of images kept approximately constant. Results show that increasing the proportion of identities generally improves performance, with the best outcomes observed for subsets containing 35% and 40% of the total identities.

The results are also visualized in Figure 4.4 , where the performances of all methods are compared for each test set. The plots also include the results from the full dataset

Description	Images	IDs	Validation		Test on Optimal Step	
			Epoch-Step	CFP-FP	Acc LFW	Acc AgeDB30
Maxmin on IDs - 5 % - (AF)	382,431	4,287	$E11-2130$	86.41	95.73	79.55
Maxmin on IDs - 10 % - (AF)	582,172	8,574	$E15-1328$	88.03	97.12	80.7
Maxmin on IDs - 15 % - (AF)	582,590	12,861	$E15-3286$	89.43	97.7	83.38
Maxmin on IDs - 20 % - (AF)	582,520	17,148	$E15-1286$	89.19	98.03	81.62
Maxmin on IDs - 25 % - (AF)	582,150	21,435	$E13-424$	90.11	97.97	83.48
Maxmin on IDs - 30 % - (AF)	582,304	25,722	$E14-1863$	91.47	98.23	85.38
Maxmin on IDs - 35 % - (AF)	582,255	30,010	$E15-3328$	91.24	98.63	87.12
Maxmin on IDs - 40 % - (AF)	581,614	34,297	$E11-3570$	92.34	98.63	86.97
Maxmin on IDs - 45 % - (AF)	581,837	38,584	$E15-4370$	**92.64**	98.83	**87.95**
Maxmin on IDs - 50 % - (AF)	582,595	42,871	$E14-1837$	92.51	**98.87**	88.48

Table 4.5: *Validation and test results for Maximin on ArcFace embeddings*. Experiments were conducted with identity subsets ranging from 5% to 50% of the total dataset. Performance generally improves as the proportion of identities increases, with the best results achieved at 45% and 50%. These configurations also yield the highest overall accuracies on CFP-FP and LFW across all experiments.

Description	Images	IDs	Validation		Test on Optimal Step	
			Epoch-Step	CFP-FP	Acc LFW	Acc AgeDB30
Random Sampling on IDs - 5 %	291,176	4,287	$E12-986$	85.87	95.75	80.08
Random Sampling on IDs - 10 %	581,320	8,574	$E14-1967$	88	96.35	82.48
Random Sampling on IDs - 15 %	582,566	12,861	$E13-3388$	90.14	97.75	83.73
Random Sampling on IDs - 20 %	578,693	17,149	$E14-4227$	89.57	97.78	85.75
Random Sampling on IDs - 25 %	581,988	21,435	$E15-3356$	87.86	97.95	84.67
Random Sampling on IDs - 30 %	581,901	25,722	$E14-3902$	91.4	98.6	87.32
Random Sampling on IDs - 35 %	582,334	30,010	$E15-3314$	90	98.15	86.9
Random Sampling on IDs - 40 %	585,657	34,297	$E14-1525$	89.71	98.4	86.95
Random Sampling on IDs - 45 %	581,535	38,584	$E15-3398$	**92.17**	**98.78**	**89.17**
Random Sampling on IDs - 50 %	581,942	42,871	$E15-1356$	92.06	98.6	88.62

Table 4.6: *Validation and test results for Random Sampling*. Experiments were conducted using subsets ranging from 5% to 50% of the total identities, maintaining the total number of images. Performance improves steadily with larger identity quantity, reaching its best results at 45% for all benchmarks, and also achieves the highest global accuracy on the AgeDB-30 benchmark across all experiments.

Description	Images	IDs	Validation		Test on Optimal Step	
			Epoch-Step	CFP-FP	Acc LFW	Acc AgeDB30
k-means Sampling on IDs - 5 %	209,813	4,287	$E15-1054$	71.2	86.9	64.27
k-means Sampling on IDs - 10 %	498,082	8,574	$E11-1090$	88.51	96.71	83.33
k-means Sampling on IDs - 15 %	581,962	12,861	$E11-1540$	90.36	97.95	83.47
k-means Sampling on IDs - 20 %	582,218	17,149	$E12-2972$	90.79	98	84.9
k-means Sampling on IDs - 25 %	582,333	21,435	$E12-961$	91.34	98.25	85.97
k-means Sampling on IDs - 30 %	582,102	25,722	$E14-2889$	**91.99**	98.42	87.22
k-means Sampling on IDs - 35 %	582,126	30,010	$E15-4342$	91.53	**98.67**	86.22
k-means Sampling on IDs - 40 %	582,678	34,297	$E15-3272$	91.44	98.52	87.9
k-means Sampling on IDs - 45 %	584,212	38,584	$E15-3104$	91.63	98.38	86.38
k-means Sampling on IDs - 50 %	582,013	42,871	$E14-1902$	91.66	98.57	**88.00**

Table 4.7: *Validation and test results for k-means sampling*. Identity subsets ranging from 5% to 50% of the dataset were selected using k-means clustering on ArcFace embeddings. Unlike other sampling strategies, k-means achieves its best performance on the CFP-FP and LFW benchmarks at a lower subset size (30% and 35% respectively), indicating earlier performance saturation. Nonetheless, the best AgeDB-30 accuracy is obtained at 50%, suggesting that larger identity diversity still benefits generalization on an age-related dataset.

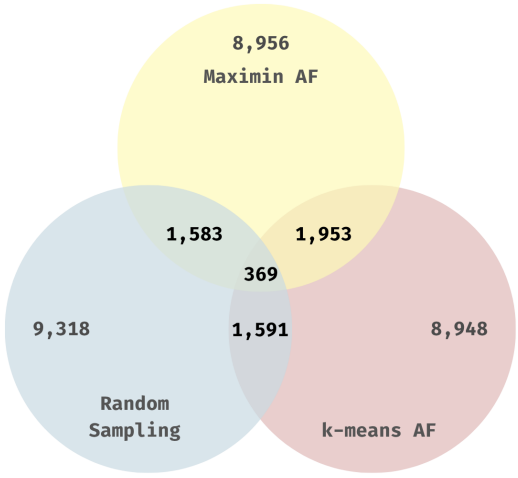


Figure 4.3: Venn Diagram for 15% of IDs selected through k-means, Maximin and Random Sampling. The diagram shows that all three methods select largely distinct identity subsets, with less than 18% overlap between them, confirming their complementarity in exploring the dataset.

implementation, obtained in [26], as well as random image selection from all identities in the dataset (calculated in Section 4.1) as references of maximum and minimum thresholds, respectively.

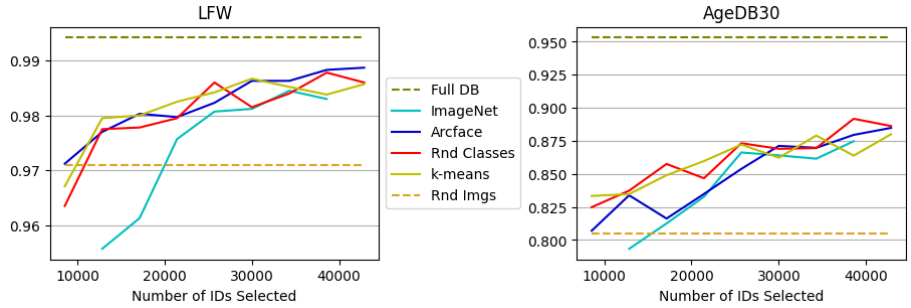


Figure 4.4: Accuracy comparison of Random Class Selection, Maximin sampling with ImageNet and ArcFace embeddings, and k-means sampling with ArcFace embeddings. For reference, the plot also includes the full-dataset training accuracy as an upper benchmark, and the random image sampling across all identities as a lower benchmark.

Across all evaluated conditions, increasing the number of identities, at the cost of fewer images per identity, consistently leads to improved performance, regardless of the sampling strategy. **However, beyond a certain point, accuracy tends to stall or even decline**. For CFP-FP, peak performance is observed at 45% for both random sampling and Maximin using ArcFace embeddings, 40% for Maximin with ImageNet embeddings, and 30% for k-means. In the LFW dataset, the best results occur at 35% for Maximin with ImageNet, 50% for Maximin with ArcFace, 45% for random sampling, and 35% for k-means. For AgeDB-30, performance peaks at 50% for both Maximin on ArcFace and k-means, 45% for random sampling, 40% for Maximin

on ImageNet.

As we can see in Figure 4.5, if we exclude the 5% IDs test, considering it incomparable due to the insufficient number of images within that proportion of identities, we observe that for identity subsets below 30%, k-means generally outperforms both Random Sampling and Maximin across the three benchmark datasets. However, this trend does not hold for larger identity subsets: in those cases, Maximin achieves superior results on CFP-FP and LFW, while Random Sampling shows better performance on AgeDB-30.

One possible explanation for this behavior is that, **as the sample size decreases, the quality and informativeness of the selected examples become increasingly important**. In such constrained scenarios, methods like k-means prioritize the representativeness of each class and could yield more representative and informative subsets. This suggests that **k-means may serve as a stronger selection mechanism in low-resource settings**, where preserving variance and reducing redundancy is important for model generalization. Conversely, as the number of identities increases and the dataset becomes more comprehensive, the marginal benefit of such selection diminishes, and broader coverage strategies like Maximin or even random selection may be equivalent or even outperform k-means sampling.

Maximin vs Random Sampling

IDs	CFP-FP	LFW	AgeDB-30
0.10	0.03	0.80	-2.16
0.15	-0.79	-0.05	-0.42
0.20	-0.42	0.26	-4.82
0.25	2.56	0.02	-1.41
0.30	0.08	-0.38	-2.22
0.35	1.38	0.49	0.25
0.40	2.93	0.23	0.02
0.45	0.51	0.05	-1.37
0.50	0.49	0.27	-0.16

k-means vs Random Sampling

IDs	CFP-FP	LFW	AgeDB-30
0.10	0.58	0.37	-1.03
0.15	0.24	0.20	-0.31
0.20	1.36	0.22	-0.99
0.25	3.96	0.31	1.54
0.30	0.65	-0.18	-0.11
0.35	1.70	0.53	-0.78
0.40	1.93	0.12	1.09
0.45	-0.59	-0.40	-3.13
0.50	-0.43	-0.03	-0.70

k-means VS Maximin

IDs	CFP-FP	LFW	AgeDB-30
0.10	0.55	-0.42	3.26
0.15	1.04	0.26	0.11
0.20	1.79	-0.03	4.02
0.25	1.36	0.29	2.98
0.30	0.57	0.19	2.16
0.35	0.32	0.04	-1.03
0.40	-0.97	-0.11	1.07
0.45	-1.09	-0.46	-1.79
0.50	-0.92	-0.30	-0.54

Figure 4.5: Comparative performance of subsampling methods, across the three benchmark datasets, expressed as percentage differences. Excluding the 5% case since results are unrepresentative for this analysis, we can see that for identity subsets under 30% k-means performs best, while for larger subsets Maximin wins on CFP-FP and LFW and Random sampling performs best on AgeDB-30.

Furthermore, while Maximin applied to ArcFace embeddings achieves the highest overall performance on CFP-FP (92.64%) and LFW (98.87%), Random Sampling obtains the best score on AgeDB-30 (89.17%). Taken together, these results are not sufficient to establish that any single subsampling method is consistently preferable to the others.

However, since all curves remain above the lower benchmark (corresponding to random image selection without considering identity information, represented by the lower dotted line in Figure 4.4), we can conclude that accounting for an optimal number of identities does yield better performance than identity-agnostic random sampling.

Additonal Experiments

Distance based sampling

In this case, three distance-based sampling strategies were evaluated: selecting the 10% of identities closest to the global embedding mean, the 10% farthest from it, and the 10% closest to the first identity in the dataset. These methods aimed to analyze how the spatial distribution of identities in the embedding space, affects model performance. Part of the motivation for these experiments was to explore whether, beyond the number of identities included, it is also relevant to ensure that the selected samples provide a reasonable coverage of the overall embedding space. Selected identities for (1) and (2) can be visualized in Figure 4.6, for the ArcFace embeddings as defined in Subsection 3.2.4, which displays a UMAP [38] projection of the ArcFace embeddings, where the 10% of identities farthest and the 10% closest from the dataset mean are shown in red are highlighted in red on the left and right panels respectively. This Figure contrasts with the more uniformly distributed sample as shown in Figure 3.6 for random sampling shown in Subsection 3.2.4.

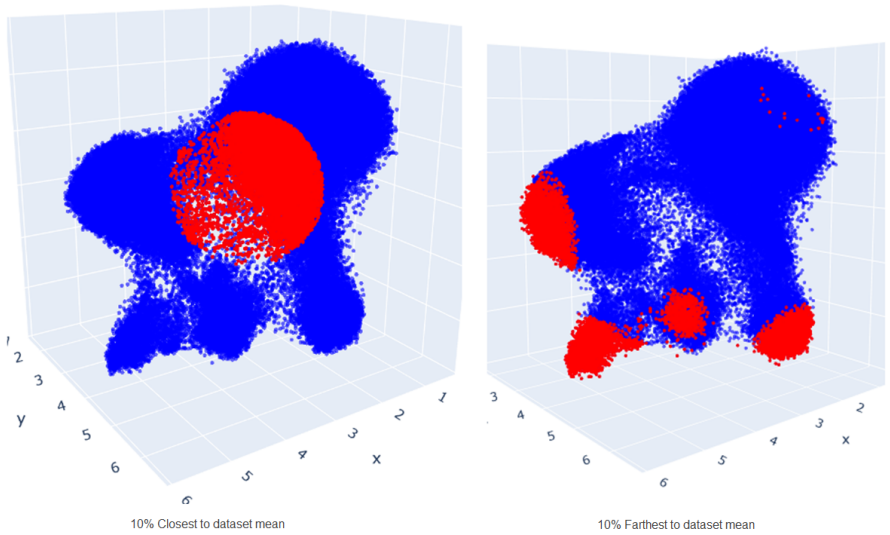


Figure 4.6: UMAP [38] projection of Arcface Embeddings, highlighting in red the 10% of identities closest to dataset mean (right), and 10% of identities farthest to dataset mean (left).

Results in Table 4.8 show that, overall, these methods perform similarly to other strategies using the same number of identities. Compared to best previously obtained results for 10% of identities, Maximin and k-means, while k-means and Maximin show strong slightly better results on CFP-FP and LFW, AgeDB-30 achieves its best performance when using identities closest to the dataset mean, suggesting that age-related features may be better captured by identities with more average or representative embeddings. However, the sample based on the farthest distances from the dataset mean yields significantly lower performance, indicating that while there may not be a consistently preferred selection method, there could be suboptimal configurations. In particular, the selection of the farthest identities does a poor job covering the images/identity manifold, leading to a lower performance.

Description	Images	IDs	Validation		Test on Optimal Step	
			Epoch-Step	CFP-FP	Acc LFW	Acc AgeDB30
Closest to dataset mean 10% IDs	582,125	8,574	$E14 - 1889$	87.27	96.53	79.8
Farthest to dataset mean 10% IDs	582,227	8,574	$E2 - 1452$	73.39	88.68	63.7
Closest to ID 0 10% IDs	319,534	8,574	$E15 - 2056$	87.01	96.11	**83.93**
Maxmin on IDs - 10 % - (AF)	582,172	8,574	$E15 - 1328$	88.03	**97.12**	80.7
k-means Sampling on IDs - 10 %	498,082	8,574	$E11 - 1090$	**88.51**	96.71	83.33

Table 4.8: *Validation and test results for alternative distance-based identity sampling strategies using 10% of the dataset.* Methods include selections based on proximity to the dataset mean, as well as to a reference identity, compared to Maximin and k-means.While Maximin and k-means achieve slightly higher accuracy on CFP-FP and LFW, AgeDB-30 performs best when using identities closest to ID 0. Conversely, selecting farthest identities leads to consistently lower performance.

Description	Images	IDs	Validation		Test on Optimal Step	
			Epoch-Step	CFP-FP	Acc LFW	Acc AgeDB30
Maximin 15% full IDs (IN)	624,048	12,861	$E14 - 625$	88.41	96.67	81.37
Maximin 15% full IDs (AF)	1,101,310	12,861	$E14 - 3625$	87.76	96.45	81.93
Random 15% full IDs	872,493	12,861	$E15 - 4576$	**89.89**	**98.00**	**85.67**
Maxmin on IDs - 15% - (IN)	582,464	12,861	$E15 - 1300$	87.87	96.13	81.25
Maxmin on IDs - 15% - (AF)	582,590	12,861	$E15 - 3286$	89.43	97.7	83.38
Random Sampling on IDs - 15%	582,566	12,861	$E13 - 3388$	**90.14**	**97.75**	**83.73**

Table 4.9: *Validation and test results comparing performance between data-limited configurations (using only 10% of the dataset) and full-identity sampling (using all available images for the same selected 15% of identities).* Each row shares the same set of identities, selected via Maximin (on ImageNet or ArcFace embeddings) and Random Sampling. Results reveal mixed effects of increasing image depth per identity, suggesting that while deeper identity data may help in some cases, it can also introduce redundancy or overfitting.

Full IDs vs 10% of data

To analyze the impact of identity depth on model performance, additional experiments were conducted using the set of 15% of the total identities, but including all available images for each. These configurations were compared with the data-limited scenarios where total images were constrained to 10% of the total dataset.

Interestingly, in most test cases, the same 15% identities with a higher image count, result in lower performance compared to configurations with fewer images per identity, which can be seen in Table 4.9. Notably, the performance trends vary across sampling strategies: for Maximin sampling based on ImageNet embeddings, accuracy improves with more images, indicating room for representational gain. In contrast, for ArcFace-based Maximin sampling, adding more images leads to a performance drop, suggesting that the initial subset was already highly representative and additional data may have introduced redundancy or overfitting. Random sampling shows mixed results, further highlighting the complex interaction between sample depth and identity diversity.

Balanced data

Finally, one of the last experiments involved constructing a dataset balanced across identities. Specifically, for each of the 30 thousand identities selected by the different methods, we attempted to allocate the same number of images per identity whenever possible.

Table 4.10 shows that, while the results on LFW were moderately comparable, for CFP-FP and AgeDB-30 the balanced setting yielded improvements of 1.3% and

Description	Images	IDs	Validation		Test on Optimal Step	
			Epoch-Step	CFP-FP	Acc LFW	Acc AgeDB30
Maximin 35% balanced IDs (AF)	582,265	30,010	$E11-3520$	92.17	**98.85**	87.35
Random 35% balanced IDs	582,265	30,010	$E15-328$	**92.43**	98.58	**87.72**
k-means 35% balanced IDs (AF)	582,265	30,010	$E12-3972$	91.76	98.43	86.93
Maxmin on IDs - 35 % - (AF)	582,255	30,010	$E15-3328$	91.24	98.63	87.12
Random Sampling on IDs - 35 %	582,334	30,010	$E15-3314$	90	98.15	86.9
k-means Sampling on IDs - 35 %	582,126	30,010	$E15-4342$	91.53	98.67	86.22

Table 4.10: *Validation and test results for balanced vs stratified sampling on 35% IDs*. Each row shares the same set of identities, selected via Maximin, and Random and k-means sampling. Balanced identities apply an equal number of images per identity, to the contrary of proportional sampling applied throughout this work. Results suggest balancing identities improves global performance, though not consistently for all datasets and sampling methods.

0.7% on average, respectively, suggesting that a more even distribution of data across identities enhances model performance compared to proportional sampling applied on all previous tests. However, the improvements were not consistent across all sampling methods and benchmark datasets.

4.2.3 Robustness metrics comparison for identity selection under Maximin, random sampling and k-means

To complement the overall accuracy and AUC results, the models using the True Positive Rate at fixed False Positive Rate (TPR@FPR) metric are also evaluated. This provides a more precise view of verification performance under different conditions. Unlike general metrics, TPR@FPR measures how well the system correctly recognizes genuine pairs (TPR) while limiting the acceptance of impostor pairs (FPR), reflecting the balance between security and usability. By comparing results across different FPR thresholds (0.01, 0.001, and 0.0001) and identity proportions, we can see how each sampling strategy affects robustness. The following results on LFW and AgeDB-30 show these effects.

Maximin	**LFW**			**AgeDB-30**		
TPR@FPR	**0.01**	**0.001**	**0.0001**	**0.01**	**0.001**	**0.0001**
5% IDs	0.8532	0.6363	0.1332	0.2076	0.0447	0.0122
10% IDs	0.9427	0.8197	0.288	0.2161	0.075	0.0127
15% IDs	0.9611	0.8467	0.3399	0.2703	0.0982	0.0192
20% IDs	0.9667	0.8827	0.3086	0.2507	0.071	0.0198
25% IDs	0.9645	0.902	0.3456	0.3211	0.0993	0.0123
30% IDs	0.9775	0.946	0.4744	0.3418	0.1547	0.0257
35% IDs	0.9833	0.9307	0.4784	0.4445	0.2543	0.0595
40% IDs	0.9827	0.9353	0.3639	0.4117	0.1737	0.0396
45% IDs	0.9863	0.9425	0.302	0.5285	0.265	0.0436

Table 4.11: TPR@FPR at different operating points for the Maximin sampling strategy on the LFW and AgeDB-30 datasets under stratified setups, evaluated across identity proportions ranging from 5% to 50%.

LFW

On average, the k-means sampling strategy achieves the best overall performance and exhibits the smallest degradation as the FPR requirement is more restrictive.

Random Samp.	LFW			AgeDB-30		
TPR@FPR	**0.01**	**0.001**	**0.0001**	**0.01**	**0.001**	**0.0001**
5% IDs	0.868	0.6533	0.1995	0.1639	0.0477	0.0097
10% IDs	0.9291	0.8312	0.293	0.2573	0.0863	0.017
15% IDs	0.9557	0.8233	0.234	0.3118	0.0413	0.0068
20% IDs	0.9577	0.8647	0.2652	0.4098	0.162	0.0199
25% IDs	0.9647	0.883	0.3058	0.4349	0.148	0.0249
30% IDs	0.9833	0.9303	0.3312	0.4223	0.1843	0.0172
35% IDs	0.9738	0.8727	0.3348	0.4892	0.1437	0.0392
40% IDs	0.9777	0.869	0.2752	0.5046	0.2387	0.0475
45% IDs	0.9843	0.9507	0.3394	0.5441	0.2033	0.0472
50% IDs	0.9809	0.9063	0.3228	0.5361	0.224	0.0589

Table 4.12: TPR@FPR at different operating points for the Random sampling strategy on the LFW and AgeDB-30 datasets under stratified setups, evaluated across identity proportions ranging from 5% to 50%.

k-means	LFW			AgeDB-30		
TPR@FPR	**0.01**	**0.001**	**0.0001**	**0.01**	**0.001**	**0.0001**
5% IDs	0.4814	0.1782	0.1138	0.0329	0.0092	0.0013
10% IDs	0.9212	0.7557	0.2882	0.2532	0.081	0.0117
15% IDs	0.9591	0.8493	0.2727	0.2443	0.077	0.0169
20% IDs	0.9708	0.9047	0.2758	0.3897	0.123	0.0137
25% IDs	0.973	0.938	0.398	0.4113	0.1277	0.021
30% IDs	0.9796	0.952	0.4175	0.4054	0.0794	0.0254
35% IDs	0.9833	0.9432	0.3525	0.3673	0.0997	0.0243
40% IDs	0.9819	0.939	0.4138	0.4913	0.2027	0.0287
45% IDs	0.981	0.9543	0.4194	0.4105	0.1063	0.0209
50% IDs	0.9836	0.9507	0.3087	0.504	0.197	0.0196

Table 4.13: TPR@FPR at different operating points for the k-means sampling strategy on the LFW and AgeDB-30 datasets under stratified setups, evaluated across identity proportions ranging from 5% to 50%.

However, unlike the general accuracy results, it performs worse at low identity proportions, likely due to the limited representativeness when few identities are available. Random Sampling generally yields the weakest and most inconsistent results across different FPR levels, although the differences among methods are relatively small, suggesting that for a simpler and more homogeneous dataset such as LFW, the choice of sampling strategy has a limited effect on model robustness.

AgeDB-30

In contrast, Random Sampling provides the best results on average for AgeDB-30, achieving competitive TPR values even with a smaller proportion of training identities compared to k-means and Maximin. This pattern aligns with its stronger performance in general accuracy metrics. For datasets characterized by greater intra-class variability, such as AgeDB-30, more structured sampling approaches do not necessarily yield an advantage, as the high diversity of facial attributes diminishes the benefits of controlled or cluster-based selection strategies.

4.2.4 ROC and metrics comparison for 35% identity selection

In addition, a comparative analysis of the ROC curves and associated metrics is presented for models trained with 35% of the total identities under both stratified and balanced setups. By examining AUC, EER, and TPR@FPR values, we can better understand the differences in model robustness and generalization under varying data distributions.

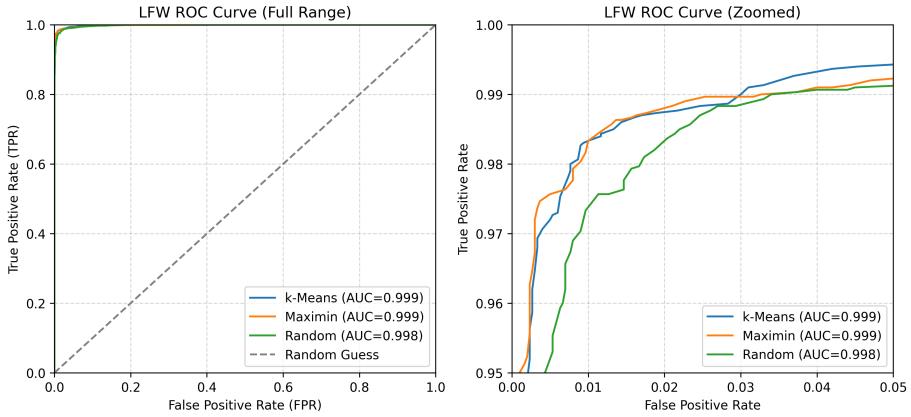


Figure 4.7: ROC curve on LFW for k-means, Maximin, and Random Sampling under a **stratified** setup, trained with 35% of total identities. On the right panel, the upper left zoomed region of the ROC curve is presented for further clarity.

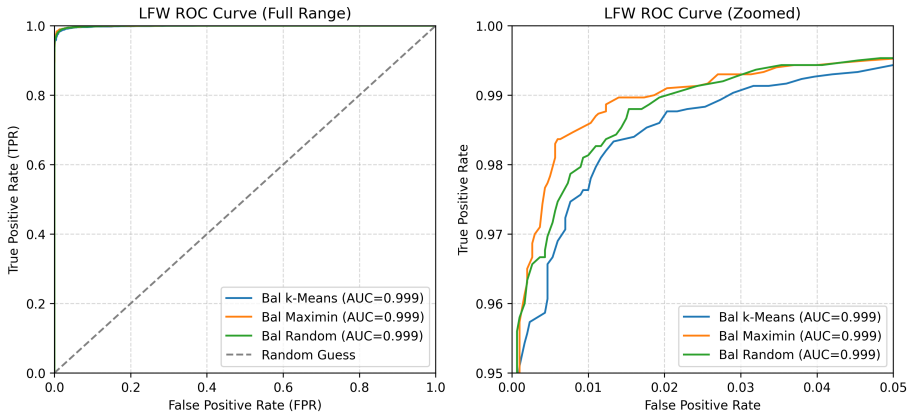


Figure 4.8: ROC curve on LFW for k-means, Maximin, and Random Sampling under a **balanced** setup, trained with 35% of total identities. On the right panel, the upper left zoomed region of the ROC curve is presented for further clarity.

On LFW, all sampling strategies across both datasets performed very well across both setups, with AUC values near 0.999 and very low Equal Error Rates (EERs) below 1.8%, as we can see in Figures 4.7 and 4.8, as well as Table 4.14. This means that all the models can distinguish between matching and non-matching pairs almost perfectly. When allowing only 1% of false positives, all methods achieve TPRs close to 0.98. Although the class-balanced setup produces slightly better results, we can

LFW	35% Stratified			35% Balanced		
	k-means	Maximin (AF)	Random	k-means	Maximin (AF)	Random
AUC	0.9989	0.9987	0.9984	0.9987	0.9991	0.9991
EER	0.0142	0.0137	0.0183	0.016	0.0123	0.0145
TPR@FPR=0.01	0.9833	0.9833	0.9733	0.9763	0.9847	0.9813

Table 4.14: Accuracy metrics on LFW for 35% identity selection under stratified and class-balanced setups. Results are shown for k-means, Maximin (AF), and Random Sampling. All configurations yield high accuracy, with AUCs near 0.999, low EERs and TPR@FPR=0.01 close to 0.98. Class-balanced setups slightly improve performance, but differences between strategies remain marginal, confirming robust model behavior across sampling methods.

Age-DB30	35% Stratified			35% Balanced		
	k-means	Maximin (AF)	Random	k-means	Maximin (AF)	Random
AUC	0.9343	0.9472	0.9427	0.9434	0.9482	0.9493
EER	0.1400	0.1267	0.1313	0.1302	0.1265	0.1195
TPR@FPR=0.01%	0.3603	0.4377	0.485	0.4487	0.4713	0.4697

Table 4.15: Accuracy metrics on Age-DB30 for 35% identity selection under stratified and class-balanced setups. Results are shown for k-means, Maximin (AF), and Random Sampling. While all methods achieve AUC scores close to 0.95, the class-balanced setup yields improvements in EER and TPR@FPR=0.01, indicating better generalization under age variation despite the overall performance still being modest.

see that for this test dataset, the choice of sampling strategy has minimal impact on verification performance, as even random sampling yields near-optimal results, and balancing classes does not substantially change overall accuracy. Since AUC scores are consistently high and performance differences between sampling strategies or data balancing are minimal, it suggests that LFW may not be a sufficiently challenging dataset for this task.

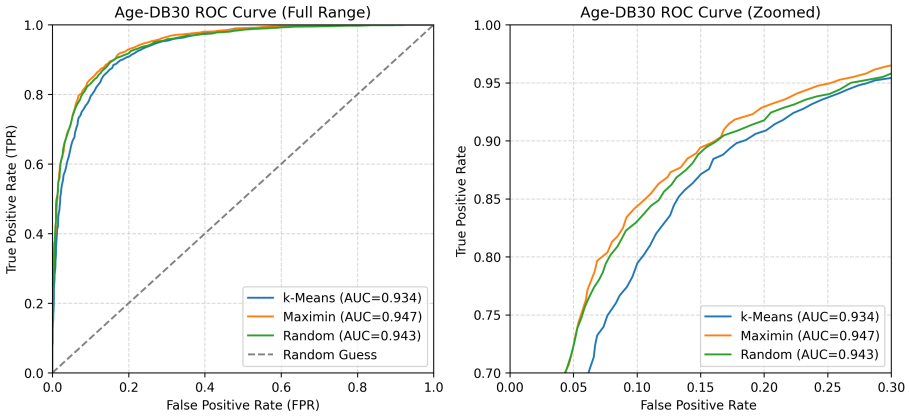


Figure 4.9: ROC curve on AgeDB-30 for k-means, Maximin, and Random Sampling under a **stratified** setup, trained with 35% of total identities. On the right panel, the upper left zoomed region of the ROC curve is presented for further clarity.

On the other hand, AgeDB-30, which introduces greater age-related variability, shows in Figures 4.9 and 4.10, and also on Table 4.15, a slightly clearer performance gap across metrics, and a modest improvement and better consistency across methods when balancing classes. While all experiments still achieve relatively high AUCs (around

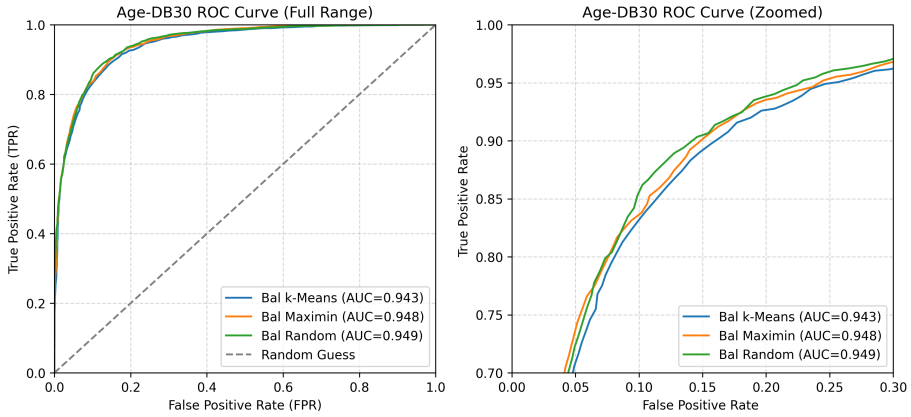


Figure 4.10: ROC curve on AgeDB-30 for k-means, Maximin, and Random Sampling under a **balanced** setup, trained with 35% of total identities. On the right panel, the upper left zoomed region of the ROC curve is presented for further clarity.

0.94), the values increased by roughly one point for balanced data, and EER decreased on average. The TPR@FPR=1% metric also improves notably (about 8%), indicating that equalizing the number of samples per identity enhances the model's ability to generalize across varying age representations.

Overall, for datasets with lower variability, such as LFW, there is no added value perceived from data selection strategies, and balancing data does not seem to improve a great deal. Under more challenging conditions, different sampling strategies can lead to marginal performance gains, yet the differences are not drastic. However, balancing the dataset offers more consistent benefits, enhancing inter-age generalization and improving true positive rates at low false positive rates, which are critical for real-world applications that demand extremely low tolerance for false alarms. Notably, random sampling often matches or outperforms more computationally demanding methods. This, combined with the consistent benefits of class balancing, suggests that a more effective and practical approach may be to ensure that the selected images contribute balanced and diverse examples, than the method used to select the identities.

This page has been intentionally left blank.

Chapter 5

Conclusion

This study analyzed the impact of different subsampling strategies for facial recognition model training, based on their eventual applicability to data acquisition, to test the hypothesis that there exists a subsampling method capable of consistently producing better results across all evaluations. Through 53 controlled experiments, both image-level and class-level sampling methods were evaluated on benchmark datasets CFP-FP, LFW, and AgeDB-30.

Some of the methods are based on reduced feature representations of the images. Although such oracle-based information is not directly applicable to data collection, these results emphasize the importance of using domain specific embeddings, like those produced by ArcFace, over more generic representations. The use of discriminative embeddings makes the identities more separable in the feature space, allowing sampling methods to select more meaningful and representative data.

At the image level, uniform random, stratified, and k-means sampling within identities, were compared. Stratified sampling provides only a modest improvement over uniform random sampling for LFW and CFP-FP and even underperforms on AgeDB-30, indicating limited benefit for age-related variability. By contrast, k-means sampling on images consistently achieves the highest performance across benchmarks, revealing that clustering can mitigate redundancy within classes, capturing representative variations, critical for model robustness. Additionally, embeddings derived from ArcFace consistently, but not significantly, outperforms those based on ImageNet, confirming the importance of high-quality feature representations.

At the class level, random selection, k-means clustering of identity embeddings, and Maximin sampling were evaluated. While Maximin with ArcFace embeddings achieve the best overall performance on CFP-FP and LFW, Random Sampling outperforms them on AgeDB-30. Furthermore, the best method varies depending on the fixed number of identities, indicating that no universally superior class-level selection strategy can be confirmed. Experiments that vary the number of identities further revealed that increasing class diversity even with fewer images per identity, tended to improve performance. This finding reinforces the widely accepted principle that broader identity coverage plays a more pivotal role than per-class image depth. However, this effect holds only up to a certain threshold, beyond which performance stalls or start to de-

cline. Optimal proportions also vary by dataset and method, generally between 30% and 50% of total identities. At smaller identity numbers, k-means sampling showed superior performance, indicating that representativeness becomes more crucial when identities are limited. In contrast, when larger subsets are available, methods such as Maximin or even random selection achieve comparable or better results. These findings confirm that while image representativeness is critical in low-resource settings, diversity of identities plays a decisive role in overall model generalization. Finally, for more challenging benchmark datasets, such as AgeDB-30, random sampling outperforms the other methods, also indicating that more complex methods might not provide additional benefits to performance.

Additional experiments reveal that identities closer to the dataset mean tend to yield better results for AgeDB-30, as they capture more representative and generalizable features, while selecting outlier identities leads to poorer performance. Also, when comparing full-identity and limited-data configurations, adding more images per identity in most cases reduces performance due to redundancy, emphasizing that diversity is more valuable than depth. Finally, balancing the number of images per identity improves generalization, confirming that evenly distributed identity representation contributes to slightly better performance.

These conclusions are reinforced by the evaluation metrics and ROC analyses, which show that sampling behavior varies across datasets. For LFW, characterized by low variability, different sampling strategies produce minimal differences, with k-means giving the most stable results under stricter FPR thresholds. In contrast, for AgeDB-30, which has greater intra-class diversity, random sampling yields the best average results, indicating that more sophisticated selection methods provide limited advantage when the dataset already contains such variability. ROC analysis also shows that while data selection strategies have little impact on homogeneous datasets, they can yield modest gains under more challenging conditions. Class balancing consistently improves generalization and TPR@FPR, which is crucial for real-world biometric applications. In general, these findings confirm that maintaining diversity and balance across identities contributes more to model robustness than the complexity of the selection strategy itself.

Overall, a key finding of this work is that, **in general, random sampling at the identity level yields competitive results, and therefore can be considered as an extremely valid method for selecting identities for training, over more complex ones**, particularly if the goal is to get greater intra-class variability. This reduces the costs associated with data collection, storage, and processing. Also, **in some cases, k-means may be more suitable when the number of identities is very limited and high intra-class variability is not required**. These outcomes are particularly relevant in ethically constrained data collection scenarios, where the volume of biometric data is limited by consent requirements. Nevertheless, once consent from a subject has been obtained, clustering can serve as a valuable guide for final image selection per identity, enhancing efficiency while maintaining representativeness.

Future research could investigate sampling strategies that explicitly account for demographic attributes, in order to evaluate their effects on fairness and generalization, which can be complementary with the strategies explored in this study. Another promising direction is the use of synthetic data generation techniques to enhance efficiency and reduce the costs associated with large-scale data collection. Furthermore, the combination of class-level and image-level sampling methods should be examined, as their joint optimization may yield more balanced and representative training sets.

Finally, further studies should place greater emphasis on the statistical validation of results through multiple experimental repetitions and extend the analysis to a broader range of face recognition datasets and models to strengthen the robustness and generalizability of the findings.

This page has been intentionally left blank.

Bibliography

[1] Mohamad Alansari, Oussama Abdul Hay, Sajid Javed, Abdulhadi Shoufan, Yahya Zweiri, and Naoufel Werghi. Ghostfacenets: Lightweight face recognition model from cheap operations. *IEEE Access*, 11:35429–35446, 2023.

[2] Vítor Albiero, Kai Zhang, and Kevin Bowyer. How does gender balance in training data affect face recognition accuracy? In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 09 2020.

[3] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, and Ying Fu. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1445–1449, October 2021.

[4] Kyriakos Axiotis, Vincent Cohen-Addad, Monika Henzinger, Sammy Jerome, Vahab S. Mirrokni, David Saulpic, David P. Woodruff, and Michael Wunder. Data-efficient learning via clustering-based sensitivity sampling: Foundation models and beyond. *ArXiv*, abs/2402.17327, 2024.

[5] Boris Bergsma, Marta Brzezinska, Oleg V. Yazyev, and Milos Cernak. Cluster-based pruning techniques for audio data. *arXiv preprint arXiv:2309.11922*, 2023.

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A Dataset for Recognising Faces across Pose and Age . In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74, Los Alamitos, CA, USA, May 2018. IEEE Computer Society.

[7] R.G. Casey and G. Nagy. An autonomous reading machine. *IEEE Transactions on Computers*, C-17(5):492–503, 1968.

[8] Jason Chan. Facial recognition technology and ethical issues. *Proceedings of the Wellington Faculty of Engineering Ethics and Sustainability Symposium*, 07 2022.

[9] Valeriia Cherepanova, Steven Reich, Samuel Dooley, Hossein Souri, John Dickerson, Micah Goldblum, and Tom Goldstein. A deep dive into dataset imbalance and bias in face identification. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 229–247, New York, NY, USA, 2023. Association for Computing Machinery.

[10] Grigorios G. Chrysos, Stylianos Moschoglou, Giorgos Bouritsas, Jiankang Deng, Yannis Panagakis, and Stefanos Zafeiriou. Deep polynomial neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:4021–4034, 2020.

[11] Colin Clifford, Tamara Watson, and David White. Two sources of bias explain errors in facial age estimation. *Royal Society Open Science*, 5:180841, 10 2018.

[12] Papers With Code. State-of-the-art in face recognition, 2024. Available at `https://paperswithcode.com/task/face-recognition`, Accessed: 9-Mar-2025.

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[14] Cody Dennis, Andries Engelbrecht, and Beatrice Ombuki-Berman. An analysis of the impact of subsampling on the neural network error surface. *Neurocomputing*, 466, 09 2021.

[15] Samuel Dooley, Rhea Sukthanker, John Dickerson, Colin White, Frank Hutter, and Micah Goldblum. Rethinking bias mitigation: Fairer architectures make for fairer face recognition. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 74366–74393. Curran Associates, Inc., 2023.

[16] K. M. Faraoun and A. Boukelif. Neural networks learning improvement using the k-means clustering algorithm to detect network intrusions. *INFOCOMP Journal of Computer Science*, 5(3):28–36, Sep. 2006.

[17] Gebrekiros Gebreyesus Gebremariam, J. Panda, and S. Indu. Blockchain-based secure localization against malicious nodes in iot-based wireless sensor networks using federated learning. *Wireless Communications and Mobile Computing*, 2023.

[18] Ruohan Gong and Zuqi Tang. Training sample selection strategy applied to cnn in magneto-thermal coupled analysis. *IEEE Transactions on Magnetics*, PP:1–1, 02 2021.

[19] Sixue Gong, Xiaoming Liu, and Anil K. Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX*, page 330–347, Berlin, Heidelberg, 2020. Springer-Verlag.

[20] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[21] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III*, volume 9907 of *Lecture Notes in Computer Science*, pages 87–102. Springer, 2016.

[22] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1577–1586, 2020.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[24] Paulina Hensman and David Masko. The impact of imbalanced training data for convolutional neural networks. *Degree Project in Computer Science, KTH Royal Institute of Technology*, 2015.

[25] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In Erik Learned-Miller, Andras Ferencz, and Frédéric Jurie, editors, *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, Marseille, France, October 2008. Inria. ECCV 2008 Workshop.

[26] Kuan-Yu Huang and Ali Fayzi. Arcface-tf2. `https://github.com/peteryuX/arcface-tf2`, 2021. GitHub repository.

[27] Omar Ibrahim, James Keller, James Bezdek, and Mihail Popescu. Experiments with maximin sampling. In *Experiments with Maximin Sampling*, pages 1–7, 07 2020.

[28] Justin Johnson and Taghi Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6:27, 03 2019.

[29] Justin Johnson and Taghi Khoshgoftaar. The effects of data sampling with deep learning and highly imbalanced big data. *Information Systems Frontiers*, 22, 10 2020.

[30] M.E. Johnson, L.M. Moore, and D. Ylvisaker. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26(2):131–148, 1990.

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

[32] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[33] Yann LeCun, Koray Kavukcuoglu, and Clement Farabet. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 253–256, 2010.

[34] Han S. Lee, Minseok Park, and Junmo Kim. Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning. *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3713–3717, 2016.

[35] Li-li Li, Heng Xiao, Yu Wang, et al. Osk: Optimal subsampling method based on k-means clustering for imbalanced big data. *Research Square*, 2023. Preprint, Version 1, posted 05 December 2023.

[36] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6738–6746, 2017.

[37] Yanbiao Ma, Licheng Jiao, F. Liu, Shuyuan Yang, Xu Liu, and Lingling Li. Predicting and enhancing the fairness of dnns with the curvature of perceptual manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[38] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

[39] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: The first manually collected, in-the-wild age database. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1997–2005, 2017.

[40] National Academies of Sciences, Engineering, and Medicine. *Facial Recognition Technology: Current Capabilities, Future Prospects, and Governance*. The National Academies Press, Washington, DC, 2024.

[41] F. Nielsen. *Introduction to HPC with MPI for Data Science*. Undergraduate Topics in Computer Science. Springer International Publishing, 2016.

[42] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and S. Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys*, 51:1–36, 09 2018.

[43] Luc Pronzato. Minimax and maximin space-filling designs: some properties and methods for construction. In *Minimax and maximin space-filling designs: some properties and methods for construction*, 2017.

[44] Isadora Rezende. Facial recognition in police hands: Assessing the 'clearview case' from a european perspective. *New Journal of European Criminal Law*, 11:203228442094816, 08 2020.

[45] Joseph P. Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Raymond Fu, and Samson Timoner. Face recognition: Too bias, or not too bias? *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–10, 2020.

[46] Alexander Schein and Michael Gee. Greedy maximin distance sampling based model order reduction of prestressed and parametrized abdominal aortic aneurysms. *Advanced Modeling and Simulation in Engineering Sciences*, 8, 12 2021.

[47] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[48] Evan Selinger and Brenda Leong. Ethical implications of facial recognition technology. In *Advances in AI and Data Ethics*, chapter 5, pages 123–145. Oxford University Press, 2023.

[49] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

[50] Vinita Shrivastava, Mohdilyas Khan, and Vijay K Chaudhari. Neural network learning improvement using k-means clustering algorithm to improve the performance of web traffic mining. In *2011 3rd International Conference on Electronics Computer Technology*, volume 1, pages 78–82. IEEE, 2011.

[51] Vin Silva and Gunnar Carlsson. Topological estimation using witness complexes. *Proc. Sympos. Point-Based Graphics*, 06 2004.

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015.

[53] Soumyadip Sengupta, Jun Cheng Chen, Carlos Castillo, Vishal M. Patel, Rama Chellappa, David W. Jacobs. Frontal to profile face verification in the wild. In *IEEE Conference on Applications of Computer Vision*, February 2016.

[54] Duo Sun, Lei Zhang, Kai feng Jin, Jiasheng Ling, and Xiaoyuan Zheng. An intrusion detection method based on hybrid machine learning and neural network in the industrial control field. *Applied Sciences*, 2023.

[55] Yehui Tang, Kai Han, Jianyuan Guo, Chang Xu, Chao Xu, and Yunhe Wang. Ghostnetv2: Enhance cheap operation with long-range attention. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 9969–9982. Curran Associates, Inc., 2022.

[56] Silvana Tayler, Marcelo Fiori, and Javier Preciozzi. Optimization of data collection in facial recognition models through subsampling strategies. In *2025 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2025.

[57] Philipp Terhorst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales, Julian Fierrez, and Arjan Kuijper. A comprehensive study on face recognition biases beyond demographics. *IEEE Transactions on Technology and Society*, 3:16–30, 2021.

[58] The European Parliament and the Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016. Accessed: 2025-02-18.

[59] Steven K. Thompson. *Sampling*. John Wiley & Sons, Hoboken, NJ, USA, 3rd edition, 2012.

[60] Muhammad Habib ur Rehman, Chee Sun Liew, Assad Abbas, Prem Prakash Jayaraman, Teh Ying Wah, and Samee U Khan. Big data reduction methods: a survey. *Data Science and Engineering*, 1(4):265–284, 2016.

[61] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[62] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 692–702, 2018.

[63] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 499–515, Cham, 2016. Springer International Publishing.

[64] Kejin Wu and Dimitris N Politis. Scalable subsampling inference for deep neural networks. *ACM/IMS Journal of Data Science*, 2(1):1–29, 2025.

[65] Zhuo Wu, Zan Wang, Junjie Chen, Hanmo You, Ming Yan, and Lanjun Wang. Stratified random sampling for neural network test input selection. *Information and Software Technology*, 165:107331, 2024.

[66] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5419–5428, 2017.

[67] Qian Zhou and Bo Sun. Adaptive k-means clustering based under-sampling methods to solve the class imbalance problem. *Data and Information Management*, 8(3):100064, 2024.

Esta es la última página.
Compilado el Monday 18th May, 2026.
http://fing.edu.uy/