Refining machine learning evaluation : statistical insights into model performance and fairness

Michaël Soumm

université PARIS-SACLAY

Refining Machine Learning Evaluation : Statistical Insights into Model Performance and Fairness

Affinage de l'Évaluation en Apprentissage Automatique : Perspectives Statistiques sur la Performance des Modèles et leur Équité

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°580 : sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Informatique
Graduate School : Informatique et sciences du numérique. Référent : Faculté des Sciences d'Orsay

Thèse préparée dans l'unité de recherche **LIST (Université Paris-Saclay, CEA)**, sous la direction de **Bertrand DELEZOIDE**, directeur de recherche, et le co-encadrement de **Adrian POPESCU**, ingénieur-chercheur

Thèse soutenue à Paris-Saclay, le 16 Décembre 2024, par

Michaël SOUMM

Composition du jury

Membres du jury avec voix délibérative

Titus ZAHARIA Professeur, Telecom SudParis, Institut Polytechnique de Paris	Président & Rapporteur
Mihai CIUC Professeur, Universitatea Politehnica Bucuresti	Rapporteur & Examinateur
Etienne BOURSIER Chargé de Recherche, INRIA & Université Paris-Saclay	Examinateur
Ana GARCIA SERRANO Professeure associée, Universidad Nacional de Educación a Distancia	Examinatrice

THESE DE DOCTORAT

NNT : 2024UPASG094

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Affinage de l'Évaluation en Apprentissage Automatique : Perspectives Statistiques sur la Performance des Modèles et leur Équité

Mots clés : Evaluation, Apprentissage profond, Apprentissage automatique, Analyse statistique, Performance, Équité

Résumé : Cette thèse aborde les limitations des méthodologies d'évaluation en apprentissage automatique en introduisant des approches statistiques rigoureuses adaptées de l'économétrie. À travers des applications dans trois domaines distincts de l'apprentissage automatique, nous démontrons comment les outils statistiques peuvent améliorer la robustesse, l'interprétabilité, et l'équité de l'évaluation des modèles. Dans l'apprentissage incrémental de classes, nous examinons l'importance des méthodes de pré-entraînement par rapport au choix de l'algorithme incrémental et montrons que celles-ci sont décisives dans les performance finales ; dans les systèmes de reconnaissance faciale, nous quantifions les biais démographiques et démontrons que des données synthétiques équilibrées démographiquement peuvent réduire significativement les disparités de performance entre les groupes ethniques ; dans les systèmes de recommandation, nous développons de nouvelles mesures basées sur la théorie de l'information pour analyser les variations de performance entre les profils d'utilisateurs, révélant que les méthodes d'apprentissage profond ne surpassent pas systématiquement les approches traditionnelles et soulignant l'importance des schémas comportementaux des utilisateurs. Ces résultats démontrent l'importance de la rigueur statistique dans l'évaluation de l'apprentissage automatique et fournissent des lignes directrices pratiques pour améliorer l'évaluation des modèles dans diverses applications.

Title : Refining Machine Learning Evaluation : Statistical Insights into Model Performance and Fairness

Keywords : Evaluation, Deep Learning, Machine Learning, Statistical analysis, Performance, Fairness

Abstract : This thesis addresses limitations in machine learning evaluation methodologies by introducing rigorous statistical approaches adapted from econometrics. Through applications in three distinct machine learning domains, we demonstrate how statistical tools can enhance model evaluation robustness, interpretability, and fairness. In class incremental learning, we examine the importance of pretraining methods compared to the choice of the incremental algorithm and show that these methods are crucial in determining final performance ; in face recognition systems, we quantify demographic biases and show that demographically-balanced synthetic data can significantly reduce performance disparities across ethnic groups ; in recommender systems, we develop novel information theory-based measures to analyze performance variations across user profiles, revealing that deep learning methods don't consistently outperform traditional approaches and highlighting the importance of user behavior patterns. These findings demonstrate the value of statistical rigor in machine learning evaluation and provide practical guidelines for improving model assessment across diverse applications.

Résumé Substantiel de la Thèse en Français

Cette thèse aborde les limitations des méthodologies d'évaluation en apprentissage automatique en appliquant des approches statistiques rigoureuses adaptées de l'économétrie. Nous démontrons comment les outils statistiques, en particulier la régression linéaire, l'ANOVA et la régression logistique, peuvent fournir une meilleure compréhension du comportement et de la performance des modèles à travers divers domaines d'apprentissage automatique. Nos travaux couvrent trois domaines distincts, chacun présentant des défis d'évaluation uniques et des opportunités d'innovation méthodologique.

Dans le domaine de l'apprentissage incrémental sans exemplaires, où les modèles doivent apprendre en continu de nouvelles classes sans accès aux données d'entraînement précédentes, notre analyse révèle plusieurs résultats importants. À travers une évaluation statistique approfondie sur plusieurs jeux de données et architectures, nous montrons que le choix de la méthode de pré-entraînement a un impact plus important sur la performance que le choix de l'algorithme d'apprentissage incrémental. Cette découverte remet en question les approches d'évaluation traditionnelles qui se concentrent principalement sur la comparaison des stratégies d'apprentissage incrémental. Nos résultats soulignent notamment que le pré-entraînement auto-supervisé peut considérablement améliorer les performances, particulièrement lorsque le modèle pré-entraîné est affiné sur les classes initiales.

Dans les systèmes de reconnaissance faciale, nous traitons la question des biais démographiques dans les algorithmes de vérification. En combinant régression logistique et ANOVA, nous quantifions précisément l'influence des différents facteurs démographiques sur la performance des modèles. Nous proposons DCFace, une nouvelle approche pour générer des données d'entraînement synthétiques démographiquement équilibrées, qui permet de réduire significativement les écarts de performance entre groupes ethniques. Notre analyse montre que les modèles entraînés sur ces données maintiennent une précision compétitive tout en réduisant considérablement les biais, bénéficiant particulièrement aux populations traditionnellement sous-représentées.

Dans le domaine des systèmes de recommandation, nous introduisons de nouvelles mesures fondées sur la théorie de l'information pour analyser les variations de performance selon les profils d'utilisateurs. Ces travaux comprennent le développement de Vis2Rec, un nouveau jeu de données pour la recommandation de visites touristiques. Notre analyse révèle que les méthodes d'apprentissage profond ne surpassent pas systématiquement les approches traditionnelles de factorisation matricielle. Nous proposons deux nouvelles mesures, la Surprise et la Surprise Conditionnelle, qui quantifient différents aspects du comportement des utilisateurs et fournissent un cadre générique pour évaluer l'efficacité des recommandations.

Cette thèse apporte plusieurs contributions méthodologiques qui s'étendent au-delà de ces domaines spécifiques. Elle montre comment les outils statistiques peuvent être adaptés pour fournir des évaluations plus nuancées et fiables des systèmes d'apprentissage automatique. Elle introduit des cadres pour quantifier l'importance relative de différents facteurs dans la performance des modèles, fournissant des méthodes pour isoler les effets causaux dans les

systèmes d'apprentissage complexes. Ces approches permettent une comparaison plus rigoureuse des modèles à travers différentes architectures et paradigmes d'entraînement tout en tenant compte des variables confondantes et des effets d'interaction.

Cette approche complète de l'évaluation de l'apprentissage automatique améliore non seulement notre compréhension du comportement des modèles et des disparités de performance, mais fournit également des idées pratiques pour améliorer les méthodologies d'évaluation à travers diverses applications d'apprentissage automatique.

Acknowledgements

Before starting, I want to thank everyone who helped make this work possible. What I thought would be a lonely journey at the start of this thesis turned into an adventure full of teamwork, new friendships, and support.

I want to deeply thank my thesis committee members: Pr. Titus Zaharia and Mihai Ciuc for serving as reviewers and taking the time to read and evaluate this manuscript with great care. Your detailed feedback and suggestions have greatly improved the quality of this work. I extend my gratitude to Dr. Etienne Boursier and Pr. Ana Garcia Serrano for agreeing to be part of the jury and for their valuable questions and insights during the defense. Your diverse expertise and perspectives have enriched this work immensely.

I want to express my deep thanks to Dr. Adrian Popescu, who co-supervised this thesis, for always being there, supporting me, and giving helpful feedback. You created a healthy work environment that helped me grow as a researcher. I also thank my thesis director, Dr. Bertrand Delezoide, for his guidance and support, and for letting me work on topics I truly cared about.

A big thank you to my colleagues, especially those I worked with on research papers: Dr. Eva Feillet, whose perfect organization, both in her workspace and mind, helped me structure my own; Dr. Grégoire Petit, whose endless energy lifted me up when things were tough; Mr. Alexandre Fournier-Montgieux, who showed me you could face any bad luck with a smile. These amazing people each brought their own brilliant ideas and have proven themselves as outstanding researchers.

More broadly, I thank everyone at LASTI, who welcomed me warmly and gave me a great place to work. Special thanks to Dr. Evan Dufraisse and Dr. Aboubacar Tuo, who became not just excellent colleagues but true friends I can count on.

To my friends, who have been there through it all, Adrien, Corentin, Kamil, Tom, Axel, Hugo, and many others who have enriched this journey with their presence, support, and friendship. Your constant encouragement, our shared laughs, late-night talks, and adventures have made these years not just bearable but truly enjoyable.

To my parents, Pacha and Galina, even though you didn't always understand what I was doing, you never stopped believing in me and supporting me. You're the ones who always pushed me to do my best and helped make me who I am today. To my sister, Alexandra, who's

been by my side my whole life, your support means everything. Our shared experiences and mutual respect have made my life so much richer, giving me a sense of normal life during all the thesis work. Your own journey, struggles, and wins keep reminding me of how strong our family is and what we stand for. Thank you for being part of this journey.

Finally, I want to thank Lou-Lélia, who has been an essential part of this journey. Your kindness, wisdom, and unwavering support have been a constant source of strength. Your ability to bring light to the darkest days, your understanding during the long work hours, and your genuine enthusiasm for my research have made this journey not just easier, but infinitely more meaningful. Thank you for sharing both the challenges and victories, for your patience during the stressful times, and for bringing balance and joy to this academic pursuit. Your presence has made this experience richer in ways words cannot fully express.

Table of Contents

When a measure becomes a target, it ceases to be a good measure.

— Marilyn Strathern's version of Goodhart's law

1 Introduction

1.1 The Evolution of Machine Learning Evaluation

The field of machine learning has undergone a remarkable transformation over the past few decades, evolving from simple statistical models to complex, multi-layered neural networks capable of tackling increasingly sophisticated tasks [Schmidhuber, 2015; Goodfellow et al., 2016]. This evolution has been driven by advances in computational power [Amodei and Hernandez, 2018], the availability of large-scale datasets [Sun et al., 2017b], and breakthroughs in model architectures and training techniques [LeCun et al., 2015; Vaswani et al., 2023]. As machine learning models have grown in complexity and capability, so too has the challenge of accurately evaluating their performance and understanding their behavior [Doshi-Velez and Kim, 2017; Lipton, 2017].

1.1.1 From Simple Metrics to Complex Performance Assessment

In the early days of machine learning, model evaluation often relied on straightforward metrics such as accuracy for classification tasks or mean squared error for regression problems [Sokolova and Lapalme, 2009]. These metrics provided a clear, easily interpretable measure of model performance that was sufficient for the relatively simple models of the time. However, as the field progressed, particularly with the advent of deep learning, the limitations of these basic metrics became increasingly apparent [Bouthillier et al., 2021].

Deep neural networks, with their ability to learn hierarchical representations from data, introduced new dimensions of complexity to the evaluation process. The sheer number of parameters in these models, often in the millions or billions, makes it challenging to understand how they arrive at their predictions. This "black box" nature of deep learning models, which prevents researchers from having strong theoretical guarantees of their models, necessitates more sophisticated evaluation approaches that could provide insights beyond simple performance numbers [Doshi-Velez and Kim, 2017].

The machine learning community responded to this challenge by developing various new metrics and evaluation techniques. For instance, in classification tasks, newer metrics like Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) emerged as a way to measure how well a model's predicted probabilities align with actual frequencies of correct predictions [Nixon et al., 2019]. In natural language processing, new evaluation

protocols were introduced to evaluate the quality of machine-generated text and go beyond standard metrics such as BLEU or ROUGE [Liu et al., 2016]. However, even these more advanced metrics and benchmarks often fail to capture the full complexity of model behavior, particularly in real-world applications where performance can vary significantly across different subgroups or contexts [Koh et al., 2020].

1.1.2 The Challenge of Comparing Diverse Models and Architectures

As machine learning has matured, it has given rise to a diverse ecosystem of model architectures, each with its own strengths and weaknesses. Convolutional Neural Networks (CNNs) [LeCun et al., 2010] have shown remarkable success in computer vision tasks, while Recurrent Neural Networks (RNNs) [Hochreiter and Schmidhuber, 1997], and more recently, Transformer models [Vaswani et al., 2023], have pushed the boundaries of natural language processing, and were then widely adopted in computer vision [Dosovitskiy et al., 2021]. This diversity, while driving progress in the field, has also made fair and meaningful comparisons between different approaches increasingly difficult.

The challenge of model comparison is further increased by the rise of transfer learning and pre-trained models [Tan et al., 2018]. Large-scale pre-trained models like BERT and GPT for natural language processing [Devlin et al., 2019; Achiam et al., 2023] or DINOv2 for computer vision [Oquab et al., 2023] have become the foundation for many state-of-the-art systems. These models, often trained on massive datasets with substantial computational resources, have raised questions about how to fairly evaluate and compare models that leverage pre-training versus those trained from scratch [Neyshabur et al., 2021].

Moreover, the growing trend of multi-modal models that can handle diverse types of input data (e.g., text, images, and audio) [Radford et al., 2021] has further complicated the evaluation landscape. Traditional metrics designed for single-modality tasks often fall short when assessing these complex, multi-faceted models. This has led to a growing recognition in the field that we need more holistic evaluation frameworks capable of providing meaningful comparisons across diverse model architectures and training paradigms [Ethayarajh and Jurafsky, 2021; Ribeiro et al., 2020].

1.1.3 The Need for Context-Aware Evaluation Methods

A significant challenge in modern machine learning evaluation is the need for methods that can account for the diverse contexts in which models are deployed. As machine learning systems are increasingly used in real-world applications, from healthcare diagnostics to financial forecasting to autonomous driving, it has become clear that performance in controlled benchmark settings does not always translate to real-world effectiveness [Hutchinson et al., 2022].

A model's performance can vary dramatically depending on the specific dataset it encounters, the task it is applied to, or the real-world scenario in which it operates [Taori et al., 2020]. Traditional evaluation methods often fail to capture these contextual nuances, leading to potential misunderstandings of a model's true capabilities and limitations. For example, the performance of pre-trained vision models is still frequently illustrated by their performance on the ImageNet dataset [Deng et al., 2009]. This tendency not only obscures the fact that ImageNet is a specific benchmark but also neglects the particularities of the pre-training dataset, which could have interesting properties for some downstream tasks. This can have serious consequences when these models are deployed in critical applications where failures can have significant real-world impacts [Amodei et al., 2016].

Furthermore, as society becomes more aware of the potential for AI systems to perpetuate or exacerbate existing biases, there is a growing demand for evaluation methods that can assess fairness and bias across different demographic groups or other relevant subpopulations [Mehrabi et al., 2021]. This requires moving beyond aggregate performance metrics to understand how models behave across diverse segments of the data [Mitchell et al., 2019].

The need for context-aware evaluation extends beyond just assessing performance and fairness. It also encompasses the ability to evaluate other crucial aspects of machine learning systems, such as their robustness to distribution shifts [Shen et al., 2021], their calibration (the alignment between predicted probabilities and true probabilities) [Guo et al., 2017], and their efficiency in terms of computational resources and energy consumption [Strubell et al., 2020]. As the field expands into more domains and industries, these evaluation methods must provide both accurate performance assessments and deeper insights into fairness, bias, generalizability, and efficiency across different scenarios and subgroups [Thomas et al., 2019; Zhang et al., 2023].

The evolution of machine learning evaluation calls for new approaches grounded in statistical principles – ones that can provide rigorous methods for assessing models across diverse architectures, tasks, and contexts, in a reproducible manner [Pineau et al., 2021]. These approaches should be capable of offering deeper insights into model behavior, going beyond surface-level performance metrics to provide a more comprehensive understanding of a model's strengths, weaknesses, and potential impacts when deployed in the real world [Molnar et al., 2020].

This thesis aims to address this pressing need by exploring how statistical tools can enhance machine learning evaluation. By leveraging techniques from econometrics and causal inference [Angrist and Pischke, 2009], we demonstrate how these analytical tools can be adapted to provide nuanced, interpretable insights into model performance across diverse contexts. Rather than proposing a single unified framework, we show how statistical principles can be applied flexibly to develop targeted evaluation approaches for different scenarios. Our goal is to move beyond simple performance comparisons to a deeper understanding of the factors that influence model behavior, ultimately contributing to the development of more robust, fair, and reliable machine learning systems.

1.2 Current Limitations in Model Evaluation

With the rise of more complicated machine learning models and setups, the limitations of traditional evaluation methods have become increasingly apparent. These limitations not only hinder our ability to accurately assess model performance but also pose significant challenges in ensuring the fairness, reliability, and interpretability of AI systems in real-world deployments.

1.2.1 Oversimplification Through Aggregate Metrics

One of the primary limitations in current model evaluation practices is the over-reliance on aggregate metrics. While measures such as overall accuracy, F1 score, or mean squared error provide a concise summary of model performance, they often mask nuances that can significantly impact a model's real-world effectiveness [Japkowicz, 2006].

Aggregate metrics can be particularly misleading when dealing with imbalanced datasets, a common occurrence in many real-world problems [He and Garcia, 2009]. For instance, in medical diagnosis tasks where positive cases are rare, a model could achieve high overall accuracy by

simply predicting the majority class, while failing to identify the crucial minority cases. This simple example is well-known and taught in most ML courses, which advocate for the use of more precise aggregate metrics, such as recall or precision. However, in scenarios where the imbalance is more subtle (many classes with different imbalance, unknown or corrupted labels, etc...), careful thinking must be put into model evaluation, and aggregate metrics must be treated cautiously [Northcutt et al., 2021].

In particular, aggregate metrics often fail to capture performance variations across different subgroups or data segments. In face recognition, a model might perform well on average but exhibit significant disparities in performance across different demographic groups or data distributions [Wang et al., 2019b]. This limitation becomes particularly problematic when evaluating models intended for diverse populations or varying environmental conditions, which poses fairness issues, a problem of its own that is tackled explicitly in this thesis.

The use of single-number metrics also tends to oversimplify the multi-faceted nature of model performance. In many applications, there are trade-offs between different aspects of performance (accuracy vs. fairness, performance on a single data vs. generalization) [Singh et al., 2021] that a single metric cannot adequately represent. This simplification can lead to suboptimal model selection and deployment decisions.

1.2.2 Challenges in Assessing Fairness and Bias

As AI systems increasingly influence decisions in areas such as biometrics, finance, and criminal justice, the need to assess and ensure fairness has become one of the new priorities in machine learning evaluation. However, evaluating fairness and bias in machine learning models presents significant challenges [Friedler et al., 2019].

One fundamental issue is the lack of consensus on how to define and measure fairness in different contexts [Jacobs and Wallach, 2019]. Various notions of fairness, such as demographic parity or equal opportunity [Hardt et al., 2016], have been proposed but are neither universally used nor let us assess fairness in a precise way. The choice of fairness metrics can significantly impact model evaluation and selection, yet there is often no clear guideline on which metrics are most appropriate for a given application [Gupta et al., 2020].

Furthermore, assessing fairness across different subgroups is complicated by the intersectionality of demographic attributes. Models that appear fair when evaluated along single demographic dimensions (e.g., gender or ethnicity separately) may still exhibit biases at the intersection of these attributes [Buolamwini and Gebru, 2018]. This intersectional fairness is challenging to measure and often requires larger, more diverse datasets than are typically available.

Another significant challenge lies in detecting and mitigating hidden biases. Models can learn subtle, unintended correlations in the training data that lead to biased outcomes, even when sensitive attributes are explicitly excluded from the input [Song and Shmatikov, 2019]. For instance, a loan approval model trained on historical data might learn to discriminate based on zip code or occupation, which can serve as proxies for protected attributes like race or gender, even when these sensitive attributes are explicitly removed from the training data. A model could develop a bias against teachers or social workers due to correlations in historical lending patterns, leading to systematically unfair treatment of certain professional groups. These hidden biases can be difficult to detect through standard evaluation methods and may only become apparent when the model is deployed in real-world settings.

Example: Limitations of Accuracy in Face Recognition

A face recognition system achieves 95% accuracy on a test set. However, this metric alone fails to reveal that:

- The system's performance varies significantly across different demographic groups.
- Most errors occur in low-light conditions, an important factor for real-world deployment.
- The system is overly confident in its incorrect predictions, potentially leading to errors in high-stakes applications.
- The test set lacks diversity in age groups, potentially overestimating the model's generalization ability.

Considered alone, the accuracy metric does not capture these insights, which are essential for understanding the model's true capabilities and limitations.

1.2.3 Lack of Interpretability in Evaluation Methods

In modern machine learning research, improvements in model performance often result from the simultaneous modification of multiple components: architecture changes, new training techniques, different training datasets or processing methods, and varying optimization strategies. While papers frequently report significant performance gains over previous approaches, it becomes increasingly difficult to pinpoint exactly which changes are responsible for these improvements [Bouthillier et al., 2021].

This challenge is particularly evident in recent breakthrough papers, where new architectures are often introduced alongside sophisticated training procedures and data augmentation techniques. For instance, when a new model achieves state-of-the-art performance, it's often unclear whether the improvement stems from the architectural innovation itself, the enhanced training methodology, or their synergistic interaction.

Example: Beating the State-of-the-Art

Consider a scenario where a new Deep Learning recognition model achieves a 5% increase in accuracy over the state-of-the-art. This improvement could be attributed to multiple factors:

- A new neural network architecture
- An increased number of parameters
- A larger or more diverse training dataset
- New data augmentation techniques
- A different optimization algorithm

Without rigorous analysis, determining the true source of improvement is challenging. Statistical methods can help isolate the effects of each factor. This allows for statements such as: *"Controlling for model size and dataset characteristics, our novel architecture contributes to a 2.3% increase in accuracy".*

The research community currently lacks systematic ways to efficiently quantify the relative importance of each modification. When comparing two approaches, differences in implementation details, hyperparameter choices, and computational budgets can all affect the final performance, making it challenging to make fair comparisons or draw definitive conclusions about which components are most crucial. The ideal solution would be to perform an ablation study

on every combination of modifications, exponentially increasing the number of experiments needed for a comprehensive study.

This limitation in our evaluation methodology has important implications for research direction and resource allocation. Without clear understanding of which components drive performance improvements, researchers might focus their efforts on less impactful modifications or unnecessarily complicate their models with components that provide marginal benefits.

1.3 Towards a Better Evaluation of ML

As we saw, there is a pressing need for more comprehensive approaches to assessing machine learning models. This thesis explores how statistical tools, drawing inspiration from econometrics and statistical analysis [Angrist and Pischke, 2009; Crepon and Jacquemet, 2010; Wooldridge, 2013], can enhance model evaluation. Through distinct contributions, we demonstrate how statistical principles can provide deeper insights into model behavior and performance across diverse contexts while addressing many limitations discussed in the previous section. The diversity of ML evaluation scenarios suggests that rather than seeking a single universal framework, we should embrace methodological pluralism guided by statistical rigor in evaluating AI models.

1.3.1 Adapting Statistical Approaches for Machine Learning Evaluation

The complexity of modern ML systems requires evaluation methods that can capture nuanced relationships between various factors and model performance. By adapting statistical approaches commonly used in econometrics, we can develop a more rigorous and insightful evaluation framework for machine learning.

This approach centers on modeling the relationships between raw performance metrics and the multiple factors that might influence them. These factors could include model architecture choices, data characteristics, training procedures, deployment contexts, or any variable that may be correlated to performance. By systematically analyzing these relationships, we can move beyond simple performance comparisons to a more nuanced understanding of what drives model behavior. This framework can be applied at the algorithm level, enabling researchers to compare the performance of multiple approaches accurately without the need to do exhaustive ablation studies [Meyes et al., 2019]. Applied at the data level, it allows us to analyze performance variations between different data points and draw practical conclusions on how each data point attribute affects performance.

Moreover, this approach enables us to explore the interplay between different aspects of machine learning systems. For example, we can investigate how the relationship between model size and performance varies depending on the amount of training data available, or perform intersectional fairness analyses.

The framework also allows for quantifying the relative importance of different factors. This can help practitioners prioritize their efforts in model development and optimization, focusing on the areas that are likely to yield the most significant improvements [Bello et al., 2021].

1.3.2 Statistical Rigor in Machine Learning Analysis

Ensuring statistical rigor is crucial for drawing reliable conclusions from our analyses. This involves not only identifying relationships between variables but also assessing their statistical significance and practical importance.

By applying concepts of statistical significance, we can differentiate between genuine effects and random fluctuations due to data sampling. This is particularly important in the context of machine learning, where the complexity of models and the variability in performance across different datasets can make it challenging to discern meaningful patterns [Slack et al., 2021; Varoquaux and Colliot, 2023].

Beyond mere statistical significance (which quantifies the *uncertainty* of our conclusions), measuring effect sizes (in other terms, the *strength* of our conclusions) allows us to understand the magnitude of different factors' impacts. This is crucial for prioritizing which aspects of a model or training process to focus on for improvement. It also helps communicate the practical significance of findings to stakeholders who may not have a deep technical understanding of the models. For example, when analyzing a recommendation system's performance, we might find that both increasing model size and improving data quality have statistically significant effects. However, effect size analysis might reveal that data quality improvements lead to a 15% performance gain, while doubling the model size only yields a 2% improvement. This quantitative comparison helps practitioners make informed decisions about resource allocation, suggesting in this case that investing in data quality would be more impactful than scaling up model architecture.

1.3.3 Balancing Comprehensive Analysis with Practical Applicability

While a comprehensive statistical framework offers many benefits, it is central to balance this rigor with practical applicability in the fast-paced field of machine learning.

One key challenge is ensuring the interpretability of results for ML practitioners who may not have extensive statistical training. This involves developing clear guidelines for applying the framework and interpreting its outputs. Visualization tools and intuitive metrics can be crucial in making the insights derived from complex analyses accessible and actionable [Doshi-Velez and Kim, 2017; Molnar et al., 2020].

Statistical tools need to be thoughtfully adapted while maintaining core analytical principles across different machine learning tasks. For instance, while the specific metrics may differ between classification tasks (focusing on confusion matrices and class-specific performance) and generative modeling tasks (requiring distributional analysis), the underlying statistical approaches for quantifying effects and uncertainties remain similar. This balance between task-specific adaptation and common statistical foundations helps ensure rigorous evaluation across diverse machine learning applications, while avoiding the limitations of either overly rigid or completely disparate evaluation approaches.

There's also the challenge of managing the complexity of the evaluation process. While a more comprehensive analysis can provide deeper insights, it may also require more time and resources. Striking the right balance between depth of analysis and practical feasibility is crucial for ensuring the framework's utility in real-world ML development processes.

By addressing these challenges, a unified statistical framework for machine learning evaluation has the potential to significantly enhance our understanding of model behavior and performance. It can provide a more solid foundation for comparing different approaches, assessing fairness and robustness, and ultimately developing more reliable and trustworthy AI systems. This approach aligns with the broader movement towards more responsible and transparent AI

development, supporting the creation of systems that are not only high-performing but also fair, interpretable, and reliable across diverse real-world contexts.

1.4 Thesis Outline

1.4.1 Application of the Evaluation Framework Across Applications

The proposed statistical evaluation tools presented in this thesis are applied to three domains of ML. Their selection was driven by both the diversity of evaluation challenges they present and their complementary nature in terms of analysis requirements. Each domain offers unique opportunities to demonstrate different aspects of our evaluation methodology while addressing significant challenges in machine learning research.

Chapter 2 provides essential background and reviews related work across our domains of interest. This chapter establishes the current state of evaluation practices in machine learning and examines the specific challenges and existing approaches in the three selected application domains: class-incremental learning, face recognition, and recommender systems. This review highlights the need for more rigorous evaluation methodologies across these seemingly disparate domains.

Chapter 3 presents our **first contribution**: a methodological framework adapting econometric tools for machine learning evaluation. This chapter introduces how regression analysis, significance testing, and effect size measurements can be adapted to address machine learning evaluation challenges. We carefully explain how these tools can be applied to different types of variables and outcomes, laying the groundwork for the applications that follow.

Chapter 4 explores the statistical evaluation of exemplar-free class-incremental learning (EFCIL), an interesting first application domain due to its many varying factors within a highly controlled experimental setting. In EFCIL, models must learn new classes sequentially without access to previous training data. The domain involves multiple potentially interacting components – from pre-training strategies and architectural choices to data characteristics – making it an excellent candidate for studying the attribution of performance gains. Within this context, we present our **second contribution**: the first large-scale systematic study of pre-training strategies for EFCIL. Through careful experimental design and statistical analysis, we quantify the relative importance of different factors, providing concrete guidance for practitioners.

Chapter 5 focuses on evaluating face recognition systems, a domain that presents unique challenges in fairness evaluation. Unlike our EFCIL study, where we directly compare models, fairness analysis in face recognition requires understanding complex interactions between demographic variables and their impact on model performance between different identities. The intricate nature of these relationships demands sophisticated evaluation approaches that can capture both performance and fairness aspects. In this context, we present two contributions: our **third contribution** is a novel data generation method, leading to the DCFace dataset, while our **fourth contribution** is a rigorous approach to fairness assessment using statistical tools. Through our analysis, we provide fine-grained insights into demographic biases and their interactions, demonstrating how our proposed controlled generation method outperforms other datasets in both fairness and performance.

Chapter 6 addresses evaluation in recommender systems, a domain that introduces yet another dimension of evaluation complexity through its need for user-centric modelization. Here, the challenge lies in understanding why certain users are more difficult for recommender systems to handle than others. This domain allows us to analyze user behavior patterns and their relationship with model performance. Within this context, we present two key

contributions: our **fifth contribution** is Vis2Rec, a new dataset specifically designed for visit recommendation tasks, while our **sixth contribution** develops novel coherence measures that quantify the inherent difficulty of recommendation tasks for specific users. By introducing new behavioral measures and studying their impact on different recommendation algorithms, we demonstrate how our approach can provide insights into the fundamental characteristics of the data that influence model performance.

Finally, Chapter 7 synthesizes our findings across these domains, identifying common patterns and insights that emerge from applying our framework. This concluding chapter also discusses the broader implications of our work and suggests promising directions for future research in machine learning evaluation.

1.5 List of Publications

The work presented in this thesis has led to the publication of the following works. Equal contribution is denoted by †.

1.5.1 Articles Published in International Conferences

- M. Soumm, A. Popescu, and B. Delezoide. Vis2rec: A large-scale visual dataset for visit recommendation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2987–2997, January 2023
- G. Petit†, M. Soumm†, E. Feillet†, A. Popescu, B. Delezoide, D. Picard, and C. Hudelot. An analysis of initial training strategies for exemplar-free class-incremental learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1837–1847, January 2024
- A. Fournier-Montgieux†, M. Soumm†, A. Popescu, B. Luvison, and H. L. Borgne. Toward fairer face recognition datasets, 2024. URL `https://arxiv.org/abs/2406.16592` *Accepted as a conference paper at the Conference on Applications of Computer Vision (WACV), to be held in March 2025*

1.5.2 Articles Under Review in International Conferences

- M. Soumm†, A. Fournier-Montgieux†, A. Popescu, and B. Delezoide. Quantifying user coherence: A unified framework for cross-domain recommendation analysis, 2024. URL `https://arxiv.org/abs/2410.02453`

1.5.3 Preprints

- M. Soumm. Causal inference tools for a better evaluation of machine learning, 2024. URL `https://arxiv.org/abs/2410.01392`

The pursuit of knowledge and understanding is not a straightforward path, but a journey filled with complexities and challenges. Each step forward brings new insights and deeper questions. It is in this endless quest for learning that we truly discover the vast potential of the human mind.

— Carl Sagan

2
Background and State-ot-the-Art

Contents

2.1 Introduction

In the last two decades, rapid advancements in deep learning have revolutionized numerous fields of machine learning, including computer vision, natural language processing, generative modeling, recommender systems, and much more. Even though the first theoretical tools of these domains were introduced in the XX$^{\text{th}}$ century, the increase in the volume of available data, computing power, and people working in the field led to the exponential development of many branches of deep learning. Alongside these developments, evaluation methodologies

have also evolved, providing valuable insights into model performance and contributing to the iterative improvement of algorithms. Researchers have developed sophisticated metrics, benchmark datasets, and evaluation frameworks that have significantly enhanced our ability to assess and compare different models. However, as the complexity and capabilities of deep learning models continue to grow, so too do the challenges in their evaluation. There is an ongoing need to ensure that evaluation methods keep pace with model advancements, providing rigorous and meaningful assessments that go beyond surface-level performance indicators and offer deeper insights into the true capabilities and limitations of these models.

This chapter examines the current state of deep learning evaluation. We begin with an overview of deep learning, setting the context for a detailed examination of evaluation methods in each specialized area. This exploration will reveal the limitations of current evaluation approaches and introduce new methods that offer more thorough assessments of model performance. We then focus on three applications: class-incremental learning, face recognition, and recommender systems. By exploring these domains, we aim to highlight the challenges and considerations necessary for developing robust and reliable deep learning systems.

As we explore each domain, this chapter prepares the ground for the detailed studies that form the core of this thesis. Our goal is to review existing practices and propose new approaches to evaluation that can lead to more transparent, reproducible, and impactful deep learning research and applications.

2.2 General Challenges in Modern Machine Learning Evaluation

The rapid evolution of deep learning has led to increasingly complex models and training paradigms, creating new challenges in model evaluation and analysis [Goodfellow et al., 2016; Zhang et al., 2019]. In this section, we present the state-of-the-art deep approaches in machine learning and analyze some of the limitations in their evaluation.

2.2.1 Deep Learning: A Brief Overview

2.2.1.1 Evolution and Key Concepts

The field of deep learning has seen remarkable progress over the past decades, with each advancement bringing new possibilities and challenges for model evaluation [LeCun et al., 2015; Schmidhuber, 2015].

Important Milestones Relevant to Evaluation Challenges

- 1986: Backpropagation algorithm [Rumelhart et al., 1986]: Enabled training of multi-layer networks, increasing model complexity
- 1998: LeNet-5 [Lecun et al., 1998]: Demonstrated convolutional neural networks (CNNs) for digit recognition, an early step toward modern deep learning architectures
- 2012: AlexNet [Krizhevsky et al., 2012]: Demonstrated the power of deep CNNs, leading to a surge in large-scale vision models
- 2014: Generative Adversarial Networks [Goodfellow et al., 2014a]: Introduced new challenges in evaluating generative models
- 2017: Transformer architecture [Vaswani et al., 2023]: Sparked the development of large language models, raising questions about evaluation at scale

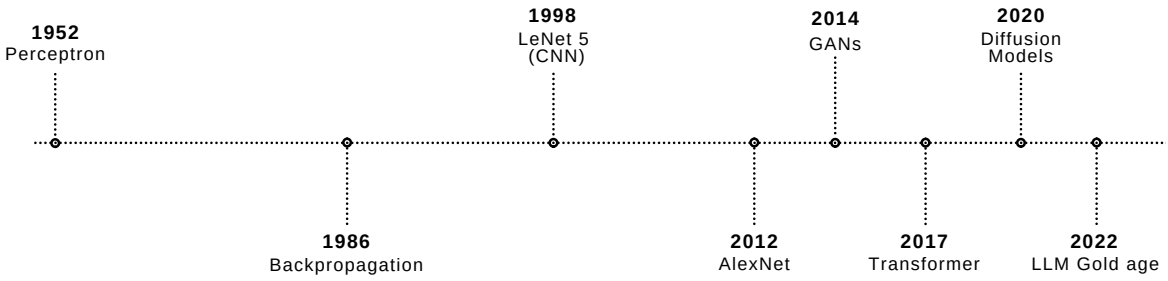


Figure 2.1: A brief historical overview of some important developments in deep learning.

These milestones have not only advanced the capabilities of deep learning but have also highlighted the need for more sophisticated evaluation techniques [Doshi-Velez and Kim, 2017; Lipton, 2017]. As models have grown in complexity, traditional evaluation metrics have often proven insufficient to capture nuanced aspects of performance and generalization [Hand, 2006; Sokolova and Lapalme, 2009].

The fundamental components of deep learning, such as neurons, layers, and activation functions, form the building blocks of modern architectures [Goodfellow et al., 2016]. However, the interactions between these components and their collective behavior have created systems whose performance can be challenging to predict and evaluate systematically [Zhang et al., 2021]. Despite significant progress in the field, we have not yet developed a comprehensive mathematical theory that can fully explain deep networks or provide guarantees of their performance [Shalev-Shwartz and Ben-David, 2014; Arora et al., 2018]. As a result, the primary method for studying the performance of deep networks remains empirical studies [Belkin et al., 2019b; Rahaman et al., 2019]. This reliance on empirical evaluation underscores the need for robust and well-designed frameworks to conduct these studies effectively. This complexity and the lack of theoretical foundations underscore the importance of developing robust evaluation methodologies, which is a central theme in our research and in recent literature [Bouthillier et al., 2021; Hendrycks et al., 2021].

Researchers have proposed various approaches to address these evaluation challenges, including more rigorous statistical testing [Demšar, 2006], the use of multiple diverse datasets [Torralba and Efros, 2011], and the development of task-specific evaluation metrics [Zhang et al., 2020a]. However, there remains a significant need for comprehensive and standardized evaluation frameworks that can keep pace with the rapid advancements in deep learning architectures and methodologies [Ethayarajh and Jurafsky, 2021].

2.2.1.2 Common Architectures

The choice of neural network architecture is primarily driven by the type and structure of the data being processed. Different architectures have been developed to effectively capture and leverage the inherent patterns and relationships within various data types [LeCun et al., 2015]. This section briefly outlines some common architectures and their applications.

Examples of Architectures for Different Data Types

- Unstructured low-dimensional data: Multi-Layer Perceptrons (MLPs)
- Image data: Convolutional Neural Networks (CNNs), Vision Transformers (ViTs)
- Sequential data: Long Short-Term Memory (LSTM), Transformers

- Graph-structured data: Graph Neural Networks (GNNs)

Multi-Layer Perceptrons (MLPs) are the most basic form of deep neural networks, suitable for unstructured data where there's no inherent spatial or temporal relationship between features [Rumelhart et al., 1986]. While simple, MLPs have been successfully applied to various tasks and serve as building blocks for more complex architectures [Goodfellow et al., 2016].

Convolutional Neural Networks (CNNs) have been the main architecture used for computer vision tasks for nearly a decade [Krizhevsky et al., 2012]. CNNs exploit the spatial structure of image data through local connectivity patterns and weight sharing, enabling them to learn hierarchical representations of visual features [LeCun et al., 2010]. The success of CNNs in image classification, object detection, and segmentation tasks has made them a standard choice for image-related problems [He et al., 2016b].

Vision Transformers (ViTs) represent a more recent development in image processing architectures [Dosovitskiy et al., 2021]. Adapting the transformer architecture originally designed for natural language processing, ViTs treat images as sequences of patches and apply self-attention mechanisms to capture global dependencies. ViTs have shown competitive performance with CNNs on various vision tasks, particularly when pre-trained on large datasets [Touvron et al., 2021].

Other notable architectures include **Recurrent Neural Networks (RNNs)** and their variants like **Long Short-Term Memory (LSTM)** networks, which are designed to handle sequential data by maintaining an internal state [Hochreiter and Schmidhuber, 1997]. These have been widely used in natural language processing tasks, although they have been increasingly replaced by transformer-based models in recent years [Vaswani et al., 2023]. For graph-structured data, **Graph Neural Networks (GNNs)** have emerged as a powerful tool [Wu et al., 2021b]. GNNs can learn representations of nodes, edges, and graphs, making them suitable for tasks such as node classification, link prediction, and graph classification [Zhang et al., 2020b]. The diversity of these architectures reflects the complexity of real-world data and the tasks we aim to solve. However, this diversity also presents challenges in evaluation, as different architectures may require different evaluation strategies and metrics [Yang et al., 2020]. Moreover, as hybrid architectures and novel designs continue to emerge [Tan et al., 2019], developing comprehensive and fair evaluation frameworks becomes increasingly important to accurately assess and compare model performance across different architectural choices [Bouthillier et al., 2021].

2.2.1.3 Learning Paradigms

Deep learning encompasses various learning paradigms, each suited to different types of tasks and data availability. This section discusses three primary learning paradigms: supervised learning, unsupervised and self-supervised learning, and generative models.

Supervised Learning: Supervised learning has been the dominant paradigm in deep learning, where models are trained on labeled datasets to learn a mapping from inputs to outputs [Goodfellow et al., 2016]. This approach has led to significant breakthroughs in various domains, including image classification [He et al., 2016b], object detection [Ren et al., 2016], and machine translation [Bahdanau et al., 2016].

Examples of Supervised Learning Tasks

- Classification: Assigning input to predefined categories [Krizhevsky et al., 2012]
- Regression: Predicting continuous values [Lathuiliere et al., 2020]
- Sequence-to-sequence learning: Mapping input sequences to output sequences [Sutskever et al., 2014]

While supervised learning has achieved remarkable success, it faces challenges. While raw data is increasingly widely available, labeling it can be expensive and time-consuming to obtain [Sun et al., 2017b]. Moreover, the loss landscapes of supervised models tend to be sharper [Lee et al., 2024], which tends to favor overfitting, a phenomenon where the model memorizes training data rather than learning generalizable patterns [Zhang et al., 2021]. Transferability of supervised features, either to out-of-distribution samples [Quionero-Candela et al., 2009; Liu et al., 2022], or to different domains [Yosinski et al., 2014], remains a very active topic of research.

Unsupervised and Self-Supervised Learning: Unsupervised learning aims to discover hidden structures in unlabeled data [Hinton and Salakhutdinov, 2006]. Self-supervised learning, a subset of unsupervised learning, creates supervisory signals from the data itself, allowing models to learn meaningful representations without explicit labels [Jing and Tian, 2020].

Example of Self-Supervised Learning Tasks

- Contrastive learning: Learning representations by contrasting similar and dissimilar samples [Chen et al., 2020a]
- Masked autoencoding: Reconstructing masked portions of input data [He et al., 2022]
- Predictive coding: Learning to predict future or missing parts of the input [van den Oord et al., 2019]

Unsupervised and self-supervised learning have gained significant attention due to their ability to leverage large amounts of unlabeled data. In computer vision, contrastive learning methods like SimCLR [Chen et al., 2020a] have shown impressive results in learning visual representations. For natural language processing, masked language modeling, as used in BERT [Devlin et al., 2019], has become a standard pre-training technique.

These approaches have shown promise in learning robust and transferable representations, often matching or surpassing supervised pre-training in downstream tasks [Chen et al., 2020b]. They are computationally attractive since they have a reduced dependence on labeled data. However, evaluating the quality of learned representations and their transferability to downstream tasks remains an active area of research [Nguyen et al., 2020].

Generative Models: Generative models aim to learn the underlying distribution of the data, enabling the generation of new samples [Goodfellow et al., 2014a]. These models have applications in various domains, including image synthesis, text generation, and data augmentation.

Main Generative Model Architectures

- Generative Adversarial Networks (GANs) [Goodfellow et al., 2014a]
- Variational Autoencoders (VAEs) [Kingma and Welling, 2022]
- Diffusion Models [Ho et al., 2020b]
- Flow-based Models [Papamakarios et al., 2021]

Generative models have made significant strides in recent years. GANs have revolutionized image synthesis, producing highly realistic images [Karras et al., 2019]. VAEs offer a probabilistic approach to generative modeling, providing both generation and inference capabilities [Kingma and Welling, 2022]. More recently, diffusion models have shown impressive results in image and audio generation [Dhariwal and Nichol, 2021]. In recent years, large-scale generative models have gained widespread attention and adoption beyond the research community. Notable examples include ChatGPT, a language model capable of engaging in human-like conversations [Achiam et al., 2023], and DALL-E, an image generation model that can create visual content from textual descriptions [Ramesh et al., 2021]. These models have become accessible to the general public, leading to their integration into various aspects of daily life and work.

The diversity of learning paradigms in deep learning has led to a rich landscape of models and applications. However, this diversity also presents challenges in evaluation:

Comparing models across different learning paradigms: It is often difficult to directly compare supervised, unsupervised, and generative models due to their different objectives and data requirements [Misra et al., 2021]. For instance, how do we fairly compare a supervised image classifier with a self-supervised representation learning model?

Assessing the quality of unsupervised representations: Evaluating the quality of learned representations in unsupervised learning is non-trivial [Tsitsulin et al., 2023]. Proxy tasks and downstream performance are often used, but these may not fully capture the richness of the learned representations.

Evaluating the fidelity and diversity of generated samples: For generative models, metrics like Inception Score and Fréchet Inception Distance are commonly used, but they have known limitations [Borji, 2018]. Balancing fidelity (quality of generated samples) with diversity (variety of generated samples) remains a challenge.

Measuring the transferability of learned representations: While unsupervised and self-supervised learning aim to learn transferable representations, quantifying this transferability across different domains and tasks is challenging [Neyshabur et al., 2021]. How do we measure the "generality" of learned representations?

Robustness to distribution shifts: Evaluating how well models perform on out-of-distribution data is crucial for real-world applications [Hendrycks and Dietterich, 2019]. This is particularly challenging for unsupervised and generative models.

Computational efficiency and scalability: As models grow larger, evaluating their efficiency and scalability becomes increasingly important [Strubell et al., 2020]. How do we balance performance gains with computational costs?

As the field continues to evolve, developing comprehensive evaluation frameworks that can address these challenges across different learning paradigms remains a critical area of research [Ethayarajh and Jurafsky, 2021; Bouthillier et al., 2021], that will be tackled more specifically in the following sections.

2.2.2 Dataset Challenges

The datasets used in deep learning research and applications often present significant challenges that can impact model performance and generalization. Two key issues in this domain are the misalignment with real-world practices and the inherent variability and bias in datasets.

2.2.2.1 Misalignment with Real-World Practices

Many benchmark datasets used in deep learning research do not accurately represent the complexities and distributions encountered in real-world applications [Quionero-Candela et al., 2009]. This misalignment can lead to overly optimistic performance estimates and poor generalization when models are deployed in practice.

Example: ImageNet vs. Real-World Images

The ImageNet dataset, while groundbreaking for computer vision research, often contains high-quality, centered images of objects. In contrast, real-world applications may encounter low-resolution, poorly lit, or partially obscured images, leading to a significant performance drop when models trained on ImageNet are deployed in practical scenarios [Taori et al., 2020].

Controlled environments vs. noisy real-world data: Most benchmark datasets are collected under controlled conditions, ensuring high-quality, well-labeled data. However, real-world scenarios often involve noisy, ambiguous, or imperfect data. This discrepancy can lead to models that perform well on clean benchmark data but struggle with the complexities of real-world inputs [Hendrycks and Dietterich, 2019]. For instance, a facial recognition system trained on studio-quality images may fail when confronted with low-light or off-angle faces in practical applications.

Temporal shifts in data distribution: Real-world data distributions often change over time, a phenomenon known as concept drift [Gama et al., 2014]. Datasets used in research, however, are typically static snapshots. This temporal mismatch can result in models that quickly become outdated or less effective as the underlying data distribution evolves. For example, a sentiment analysis model trained on social media data from 2010 may not accurately capture current language usage and sentiment expressions.

Domain-specific nuances not captured in general datasets: General-purpose datasets often fail to capture the nuances and specificities of particular domains or applications. This can lead to models that perform well on broad tasks but struggle with domain-specific challenges [Pan and Yang, 2009]. For instance, a general object detection model might perform poorly in specialized fields like medical imaging or satellite imagery, where domain-specific features and contexts are crucial.

Difficulty in simulating rare scenarios: Many real-world applications involve rare but highly important events or scenarios that are challenging to represent adequately in training datasets. This is particularly crucial in safety-critical systems, where the ability to handle rare edge cases can be a matter of life and death [Amodei et al., 2016]. For example, autonomous driving datasets may not sufficiently represent rare traffic scenarios or extreme weather conditions, leading to potentially dangerous failures in real-world deployment.

Summary of Challenges in Dataset-Practice Alignment

- Controlled environments vs. noisy real-world data
- Temporal shifts in data distribution
- Domain-specific nuances
- Difficulty in simulating rare but critical scenarios

To mitigate these challenges, researchers are increasingly focusing on developing more representative datasets and evaluation protocols that better reflect real-world conditions. Koh et al. [2020] introduced the WILDS benchmark to address this issue, providing a collection of datasets that reflect real-world distribution shifts. These datasets aim to bridge the gap between controlled research environments and the challenges faced in practical applications. In recent years, efforts have been made to create larger, more diverse datasets that better capture real-world complexity. One notable example is LAION (Large-scale Artificial Intelligence Open Network), which has produced datasets like LAION-5B, a massive dataset of 5.85 billion CLIP-filtered image-text pairs [Schuhmann et al., 2022]. LAION datasets are designed to be more representative of real-world data, including a wide range of image qualities, styles, and contexts. These datasets have been instrumental in training large-scale vision-language models and generative models like Stable Diffusion [Rombach et al., 2022]. However, while LAION and similar large-scale datasets offer greater diversity and scale, they also present new challenges. These include increased potential for biases due to web-scraped content, difficulties in content moderation at scale, and the need for more sophisticated filtering and quality control methods. Furthermore, using such massive datasets raises important ethical and legal questions regarding data provenance, copyright, and privacy [Birhane and Prabhu, 2021].

As the field progresses, it becomes increasingly important to develop datasets and evaluation methodologies that not only capture the complexity of real-world scenarios but also address the ethical and societal implications of large-scale data collection and model training.

2.2.2.2 Variability and Bias in Datasets

Dataset variability and bias present significant challenges in deep learning, affecting model performance, fairness, and generalization capabilities.

Dataset Bias

Dataset bias refers to systematic errors or imbalances in training data that can lead to unfair or inaccurate model predictions, particularly for underrepresented groups or scenarios.

Mehrabi et al. [2021] provide a comprehensive overview of various types of biases that can occur in machine learning datasets, including:

- Historical bias: Reflecting societal prejudices in the data
- Representation bias: Underrepresentation of certain groups or scenarios
- Measurement bias: Inconsistencies in data collection across different groups
- Aggregation bias: Combining distinct groups into a single category, losing important distinctions

Fabbrizzi et al. [2022] further explores biases specifically in visual datasets, categorizing them into three main types: label bias, negative set bias, and scene/background bias. Their work highlights how these biases can significantly impact the performance and fairness of computer vision models.

These biases can lead to models that perform poorly on minority groups or reinforce existing societal inequalities when deployed in real-world applications.

Examples of Bias in Visual Datasets

- Gender Bias in Facial Recognition: Facial recognition systems trained on datasets with predominantly light-skinned male faces have shown significantly higher error rates for women and people with darker skin tones [Mehrabi et al., 2021].

- Occupation Bias: In image datasets, images of doctors are more likely to show men, while images of nurses are more likely to show women, reinforcing gender stereotypes [Fabbrizzi et al., 2022].
- Geographic Bias: Datasets like ImageNet are predominantly composed of images from Western countries, leading to poor performance on objects and scenes common in other parts of the world [Fabbrizzi et al., 2022].

Dataset variability, on the other hand, refers to the inherent differences in data distributions across different domains, time periods, or geographic locations. This variability can lead to challenges in model generalization and robustness [Shen et al., 2021]. For instance, a model trained on images from one geographic region may perform poorly when applied to images from another region due to differences in lighting conditions, architecture styles, or cultural contexts [Fabbrizzi et al., 2022]. These challenges can be addressed in various ways:

Diverse and representative data collection: This involves actively seeking out data from underrepresented groups and scenarios. Fabbrizzi et al. [2022] suggest using targeted data collection methods to ensure balanced representation across different demographic groups, geographic locations, and cultural contexts. This may include collaborating with diverse communities to gather more inclusive data. Gebru et al. [2021] propose creating "datasheets for datasets" to document the composition, collection process, and intended uses of datasets, which can help identify and address representational issues.

Careful data annotation and quality control: Improved annotation processes can help mitigate label bias. Fabbrizzi et al. [2022] recommend using diverse annotator teams and implementing rigorous quality control measures. This might involve multiple rounds of annotation, cross-validation of labels, and explicit consideration of cultural and contextual factors in the annotation process. Northcutt et al. [2021] highlight the importance of identifying and correcting label errors in test sets to ensure accurate model evaluation.

Bias-aware preprocessing and augmentation techniques: These techniques aim to balance datasets and reduce existing biases. Methods such as resampling, reweighting, and synthetic data generation can help address representation biases [Shorten and Khoshgoftaar, 2019b]. For visual data, techniques like style transfer or domain randomization can help increase variability and improve model robustness [Fabbrizzi et al., 2022].

Regular dataset audits and updates: Periodic assessments of datasets can help identify and address biases over time. This includes analyzing the distribution of different attributes, checking for outdated or inappropriate content, and ensuring the dataset remains relevant as societal norms and visual cultures evolve [Fabbrizzi et al., 2022]. Birhane and Prabhu [2021] demonstrate the importance of such audits by revealing troubling issues in popular computer vision datasets.

Development of domain-specific datasets for specialized applications: Generic datasets often fail to capture the nuances of specific domains. Creating specialized datasets for areas like medical imaging, satellite imagery, or industrial inspection can help address domain-specific biases and variabilities [Shen et al., 2021]. For instance, Wang et al. [2020b] discuss the creation of a tailored dataset for COVID-19 diagnosis from chest X-rays, highlighting the importance of domain expertise in dataset development.

Summary of Strategies to Address Dataset Variability and Bias

- Diverse and representative data collection
- Careful data annotation and quality control
- Bias-aware preprocessing and augmentation techniques
- Regular dataset audits and updates
- Development of domain-specific datasets for specialized applications

Addressing these challenges requires a multi-faceted approach, including more diverse and representative data collection, careful preprocessing and augmentation techniques [Shorten and Khoshgoftaar, 2019b], and the development of robust evaluation methods that can detect and quantify biases in both datasets and trained models. Buolamwini and Gebru [2018] demonstrate the effectiveness of such comprehensive approaches in addressing gender and racial bias in commercial facial analysis systems. Furthermore, as Fabbrizzi et al. [2022] emphasize, it is crucial to consider the entire machine learning pipeline, from data collection to model deployment, to effectively mitigate biases and account for dataset variability.

2.2.3 Metric-Related Limitations

The evaluation of deep learning models often relies on a set of standard metrics that, while useful, may not fully capture the complexity and nuances of model performance. This section discusses the limitations of current evaluation practices, focusing on the overreliance on aggregate metrics, the lack of task-specific evaluation criteria, and the challenges in assessing model robustness.

2.2.3.1 Overreliance on Aggregate Metrics

Many deep learning evaluations heavily rely on aggregate metrics that provide a single summary statistic of model performance. In particular, for classification problems, the cross-entropy loss used during model training is hard to interpret. While these metrics offer a convenient way to compare models, they often hide important nuances in model behavior. Usual aggregate metrics for classification include:

Accuracy: Proportion of correct predictions

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

where TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives

F1 Score: Harmonic mean of precision and recall

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

where

$$\textbf{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \textbf{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Mean Average Precision (mAP): Average precision across all recall levels

$$\text{mAP} = \frac{1}{C} \sum_{c=1}^{C} \text{AP}(c)$$

where C is the number of classes, and AP(C) is the average precision for class c

Area Under the ROC Curve (AUC-ROC): Model's ability to distinguish between classes

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(t) \cdot \frac{d\text{FPR}(t)}{dt} dt$$

where TPR is the True Positive Rate and FPR is the False Positive Rate:

$$\text{TPR}(t) = \frac{\text{TP}(t)}{\text{TP}(t) + \text{FN}(t)}, \quad \text{FPR}(t) = \frac{\text{FP}(t)}{\text{FP}(t) + \text{TN}(t)}$$

and t is the classification threshold

The problem with aggregate metrics lies in their inherent nature of condensing complex performance characteristics into a single value. While these metrics provide a convenient way to compare models, they often oversimplify the nuanced behavior of machine learning systems, particularly in real-world scenarios with diverse data distributions. Sokolova and Lapalme [2009] highlight that such simplification can misinterpret a model's true capabilities and limitations. This issue is especially pronounced in heterogeneous datasets scenarios, where performance can vary significantly across different subgroups or in specific, often critical, edge cases. For instance, a model might achieve high overall accuracy while performing poorly on minority classes or failing catastrophically in rare but important situations. This discrepancy between aggregate performance and fine-grained behavior can lead to unexpected and potentially harmful outcomes when models are deployed in real-world applications. Several specific limitations of aggregate metrics are that they:

Mask performance disparities across subgroups: they often hide significant variations in model performance across different demographic groups or data subsets. Buolamwini and Gebru [2018] demonstrated this issue in facial recognition systems, where overall high accuracy masked substantially lower performance for women and people with darker skin tones. This masking effect can perpetuate and even exacerbate existing biases and inequalities when such systems are deployed.

Fail to capture rare but critical errors: In many real-world applications, rare events can have disproportionately high importance. Amodei et al. [2016] discuss this in the context of AI safety, where a model's performance on infrequent but critical scenarios (e.g., edge cases in autonomous driving) can be more important than its average performance. Aggregate metrics typically fail to highlight these crucial edge cases, potentially leading to oversight of critical failures.

May not align with task-specific goals: Different applications often have unique performance requirements that aren't well-captured by standard metrics. For instance, in medical diagnosis, false negatives (missing a disease) might be more costly than false positives, but this nuance is lost in balanced metrics like accuracy. Powers [2011] emphasize the importance of choosing evaluation metrics that align with the specific goals and constraints of the task at hand.

Can be misleading when classes are imbalanced: In datasets with significant class imbalance, metrics like accuracy can be misleading. A model that always predicts the majority class might achieve high accuracy but be practically useless. Jeni et al. [2013] demonstrate this issue in emotion recognition tasks, where class imbalances are common. They propose using metrics like the Matthews Correlation Coefficient (MCC) or balanced accuracy to provide a more reliable performance assessment in such scenarios.

Addressing these limitations requires a more comprehensive approach to model evaluation. This may involve disaggregated analysis across different subgroups [Mitchell et al., 2019], focused evaluation on critical edge cases [Thomas et al., 2019], the use of multiple complementary metrics

[Powers, 2011], and the development of task-specific evaluation criteria that better align with real-world performance requirements [Li et al., 2018]. By adopting such multifaceted evaluation strategies, researchers and practitioners can gain a more nuanced and reliable understanding of model performance, leading to more robust and fair AI systems.

Summary of Limitations of Aggregate Metrics

Aggregate metrics:

- Mask performance disparities across subgroups
- Fail to capture rare but critical errors
- May not align with task-specific goals
- Can be misleading when classes are imbalanced

To address these limitations, researchers are increasingly advocating for more comprehensive evaluation approaches that consider performance across various subgroups and scenarios. This includes disaggregated analysis, where performance is reported separately for different demographic groups or data subsets. Recent work by Zhang et al. [2023] proposes using stratified performance metrics to better capture model behavior across different data slices. Kearns et al. [2018] introduce the concept of "fairness gerrymandering" and suggest techniques to prevent models from exploiting loopholes in aggregate fairness metrics. Building on this, Dwork et al. [2021] propose a framework for multicalibration, which ensures predictive parity across a wide range of overlapping subgroups. To address task-specific goals, Lipton [2017] argues for developing contextual evaluation metrics that align more closely with domain-specific objectives. For imbalanced datasets, Grandini et al. [2020] provides a comprehensive analysis of various metrics and proposes guidelines for choosing appropriate evaluation measures based on the specific characteristics of the problem. These advancements in metric design and application represent a significant step toward more nuanced and reliable model evaluation in deep learning, helping to uncover hidden biases and performance discrepancies that may be obscured by traditional aggregate metrics.

2.2.3.2 Lack of Task-Specific Evaluation Criteria

Standard evaluation metrics often fail to capture task-specific requirements and nuances. Different applications may have unique performance criteria that are not well-represented by generic metrics.

Examples of Task-Specific Evaluation Needs

- **Medical diagnosis:** In medical applications, the consequences of false negatives (missing a disease) can be far more severe than false positives. Therefore, evaluation metrics that prioritize sensitivity (recall) over specificity may be more appropriate. McClish [1989] propose using partial area under the ROC curve (pAUC) to focus on high-sensitivity regions, particularly relevant for screening tests where missing a positive case is costly.
- **Recommender systems:** accuracy is not the only factor in a successful recommender system. Herlocker et al. [2004] argue that user satisfaction often depends on factors like diversity, novelty, and serendipity of recommendations. Metrics such as intra-list diversity [Ziegler et al., 2005] and coverage [Ge et al., 2010] have

been developed to capture these aspects, providing a more holistic evaluation of recommender system performance.

- **Autonomous driving:** In safety-critical applications like autonomous driving, traditional metrics like average accuracy can be misleading. Borg et al. [2018] propose a framework for evaluating autonomous driving systems that emphasizes performance in rare scenarios.
- **Image memorability:** Predicting which images are memorable to humans is a unique challenge that requires specialized evaluation approaches. Cohendet et al. [2018], in the context of the MediaEval Benchmarking Initiative for Multimedia Evaluation, introduce an evaluation protocol designed explicitly for image memorability prediction tasks, considering the ranking of images based on their memorability, providing a more comprehensive assessment of model performance in this domain.

The lack of task-specific evaluation criteria can lead to models that optimize for the wrong objectives, potentially compromising their real-world utility. To address this, researchers are developing more specialized evaluation frameworks tailored to specific domains and applications [Ribeiro et al., 2020]. These frameworks often combine multiple metrics and incorporate domain knowledge to provide a more nuanced and relevant assessment of model performance. For instance, in natural language processing, Liu et al. [2016] reviews various task-specific metrics for dialogue systems, highlighting how these specialized metrics capture aspects of language quality and coherence that general-purpose metrics might miss. Similarly, in computer vision, Behzadi-Khormouji and Oramas [2023] propose task-specific evaluation protocols for visual reasoning tasks, emphasizing the need to assess both accuracy and the model's ability to provide human-interpretable explanations for its decisions.

2.2.3.3 Challenges in Evaluating Model Robustness

Assessing the robustness of deep learning models is crucial in standard evaluation practices. Robustness, in the context of machine learning, refers to a model's ability to maintain consistent and reliable performance under various conditions or perturbations that may differ from its training environment. However, it remains often overlooked in standard evaluation practices, and encompasses several aspects.

Generalization to out-of-distribution data: This refers to a model's ability to perform well on data that differs significantly from its training distribution. Hendrycks and Dietterich [2019] introduce benchmarks for evaluating out-of-distribution generalization, highlighting how models often struggle with distribution shifts. Techniques like domain adaptation [Wang and Deng, 2018] and robust optimization [Sinha et al., 2017] have been proposed to improve performance on out-of-distribution data.

Resistance to adversarial attacks: Adversarial attacks involve carefully crafted perturbations to input data that can cause models to make incorrect predictions. Goodfellow et al. [2014b] first demonstrated the vulnerability of deep neural networks to such attacks. Evaluating adversarial robustness often involves measuring a model's accuracy under various types of attacks, such as those proposed by Goodfellow et al. [2014b] and Carlini and Wagner [2017]. Defensive techniques like adversarial training [Shafahi et al., 2019] aim to improve model resistance to these attacks.

Stability under input perturbations: This aspect focuses on a model's ability to maintain consistent predictions when inputs are slightly modified in ways that shouldn't affect

the outcome. Zheng et al. [2016] propose evaluation methods for assessing model stability, including measuring prediction consistency under small input transformations. Techniques like data augmentation [Shorten and Khoshgoftaar, 2019a] and regularization methods [Zhang et al., 2018] can help improve stability.

Consistency across different random initializations: This refers to the variability in model performance due to randomness in initialization and training. Dodge et al. [2020] demonstrate that the impact of random seeds can be significant, especially for smaller datasets. Evaluating this aspect often involves training models with multiple random seeds and analyzing the distribution of results. Ensemble methods [Lakshminarayanan et al., 2017] can help mitigate inconsistencies across initialization.

Performance under various levels of noise or data corruption: This aspect evaluates how well models perform when input data is corrupted or noisy, which is common in real-world scenarios. Hendrycks and Dietterich [2019] introduce a benchmark for corruption robustness, simulating various types of noise and corruptions. Techniques like noise injection during training [You et al., 2019] and robust loss functions [Barron, 2019] have been proposed to improve robustness to noise and corruption.

Summary of Model Robustness Related Challenges

- Generalization to out-of-distribution data
- Resistance to adversarial attacks
- Stability under input perturbations
- Consistency across different random initializations
- Performance under various levels of noise or data corruption

Evaluating robustness is challenging because it requires testing models under a wide range of conditions, many of which may not be represented in standard test sets. Furthermore, there's often a trade-off between robustness and other performance metrics, making it difficult to compare models solely based on standard benchmarks [Taori et al., 2020].

The recent evaluation approaches provide a more comprehensive view of model performance but also add complexity to the evaluation process, highlighting the need for standardized robustness evaluation protocols in the deep learning community. While current evaluation metrics provide valuable insights, they often fall short in capturing the full spectrum of model performance, task-specific requirements, and robustness. Addressing these limitations requires a more holistic approach to evaluation that combines standard metrics with task-specific criteria, disaggregated analysis, and robust testing methodologies.

2.2.4 Complexity in Performance Attribution

As deep learning models become more sophisticated, attributing performance gains to specific factors becomes increasingly challenging. This complexity arises from the interplay of various elements in the deep learning pipeline, making it difficult to isolate the impact of individual components.

2.2.4.1 Interplay of Data, Model Size, and Architecture

The performance of deep learning models is influenced by a complex interplay of factors, including the quantity and quality of training data, model size, and architectural choices.

Data quantity and quality: Larger datasets often lead to better performance, but the relationship is not always linear [Sun et al., 2017a]. The quality and diversity of data can be as important as quantity. Barz and Denzler [2020] demonstrate that carefully curated smaller datasets can sometimes outperform larger, noisier ones. Moreover, Jiang et al. [2020] show that the distribution of data across classes and the presence of hard examples significantly impact model performance, highlighting the importance of data quality beyond mere quantity.

Model size: Increasing model size (e.g., number of parameters) generally improves performance, but with diminishing returns and increased computational cost [Kaplan et al., 2020]. Brown et al. [2020] showcase the impressive capabilities of extremely large language models, but also highlight the computational challenges they pose. On the other hand, Frankle and Carbin [2019] propose the "lottery ticket hypothesis," suggesting that smaller subnetworks within large models might be responsible for most of the performance, complicating the relationship between model size and performance.

Architecture: Different architectural choices (e.g., convolutional vs. transformer models) can lead to varying performance across tasks [Dosovitskiy et al., 2021]. The success of transformer architectures in both natural language processing [Vaswani et al., 2023] and computer vision [Dosovitskiy et al., 2021] has challenged long-held assumptions about optimal architectures for different domains. Zoph et al. [2018] demonstrate that neural architecture search can discover novel architectures that outperform human-designed ones, further complicating the attribution of performance gains to specific architectural choices.

Challenges in Performance Attribution

- Difficult to isolate the impact of individual factors
- Performance gains may be due to complex interactions rather than single improvements
- Scaling effects can mask or amplify the influence of specific components

The intricate interplay between these factors creates a multidimensional optimization problem where improvements in one area can have cascading effects on others. For instance, Nakkiran et al. [2021] and Belkin et al. [2019a] observe a "double descent" phenomenon where increasing model size beyond the point of interpolation can lead to improved generalization, contrary to classical statistical learning theory. This non-monotonic relationship between model complexity and performance further complicates attribution efforts. Moreover, the impact of these factors can vary significantly across different tasks and domains. What works well for image classification might not be optimal for natural language processing or reinforcement learning. This task-dependence adds another layer of complexity to performance attribution, necessitating careful, context-specific analysis rather than one-size-fits-all explanations.

2.2.4.2 Impact of Training Methodologies

Training methodologies are central in model performance but are often overlooked in evaluation. The choice of training approach can significantly influence not only the final performance of a model but also its convergence speed, generalization ability, and robustness.

The choice of optimizer (e.g., SGD, Adam) can significantly affect model performance [Kingma and Ba, 2015]. While Adam has become popular due to its adaptive learning rates and momentum, Wilson et al. [2017] show that carefully tuned SGD can often outperform Adam in terms of generalization. Furthermore, Loshchilov and Hutter [2019] introduce AdamW, demonstrating that the interaction between weight decay and adaptive learning rates in Adam can lead to suboptimal performance, highlighting the nuanced impact of optimizer choice on model behavior.

Different **learning rate strategies** can lead to varied convergence and final performance [Smith, 2017]. Beyond simple decay schedules, techniques like learning rate warmup [Goyal et al., 2017] and cyclic learning rates [Smith, 2017] have shown to improve both convergence speed and final performance. Loshchilov and Hutter [2017] introduce cosine annealing with warm restarts, demonstrating how complex learning rate schedules can help models escape local optima and achieve better generalization.

Augmentation techniques like mixup or cutout can improve generalization but complicate performance attribution [Zhang et al., 2018]. Cubuk et al. [2019] show that automated augmentation strategies can significantly boost performance across various tasks. However, the effectiveness of augmentation can vary greatly depending on the dataset and model architecture. Hernandez-Garcia and König [2020] demonstrate that the impact of data augmentation can be more pronounced in smaller datasets, potentially masking the true capabilities of the underlying model architecture.

The **impact of pre-training** on different datasets or with different objectives (e.g., supervised vs. self-supervised) can be substantial but hard to quantify [He et al., 2020a]. Self-supervised pre-training methods like SimCLR [Chen et al., 2020a] and BERT [Devlin et al., 2019] have shown remarkable success in learning transferable representations. However, Neyshabur et al. [2020] argue that the benefits of pre-training may be more about optimization than learned representations, complicating our understanding of why pre-training works.

Pre-training Impact

A model pre-trained on a large dataset like ImageNet may perform well on a downstream task with limited data. However, it is challenging to determine how much of the performance is due to the pre-training versus the model architecture or fine-tuning strategy. For instance, Kornblith et al. [2018] show that better ImageNet performance doesn't always translate to better transfer learning performance, suggesting that the relationship between pre-training and downstream performance is not straightforward.

The complexity of training methodologies extends beyond these individual components. Their interactions can lead to surprising effects. For example, Smith [2018] demonstrate that the relationship between batch size and learning rate can be leveraged to dramatically reduce training time without loss of accuracy. Similarly, You et al. [2020] show how careful tuning of optimization algorithms and learning rate schedules can enable training on extremely large batch sizes, allowing for better utilization of distributed computing resources.

Recent work has also highlighted the importance of considering the entire training pipeline holistically. Bello et al. [2021] show that many advances attributed to architectural innovations might be explained by improved training techniques, emphasizing the need for careful ablation studies and controlled comparisons.

2.2.4.3 Difficulty in Isolating Contributory Factors

The interdependence of various factors in deep learning models makes it challenging to isolate the contribution of individual components. This complexity stems from the intricate interplay between model architecture, dataset characteristics, training methodologies, and even hardware configurations.

Ablation studies: While useful, ablation studies may not capture complex interactions between components [Meyes et al., 2019]. Traditional ablation studies involve removing or replacing individual components to assess their impact. However, Hooker et al. [2019] demonstrate that this approach can be misleading in deep learning, as the importance of a feature may depend on the presence or absence of other features. They propose a more nuanced approach called ROAR (Remove and Retrain) to better capture these interdependencies.

Hyperparameter sensitivity: Performance can be highly sensitive to hyperparameters, making it difficult to compare models fairly [Yang and Shami, 2020]. The high-dimensional nature of hyperparameter spaces in deep learning models means that seemingly minor changes can lead to significant performance differences. Li et al. [2020] demonstrate that many reported improvements in neural architecture search may be due to differences in hyperparameter optimization rather than architectural innovations. This sensitivity highlights the need for rigorous hyperparameter tuning protocols and fair comparison methodologies.

Random seeds: The impact of random initialization and data shuffling can be significant, especially for smaller datasets [Mishchenko et al., 2020]. Dodge et al. [2020] show that the choice of random seed can sometimes have a larger impact on performance than architectural changes, particularly for smaller datasets or models. This variability complicates the reproducibility of results and the assessment of genuine improvements. Bouthillier et al. [2021] propose methods for quantifying this variability and suggest reporting practices to improve the reliability of deep learning research.

To address these challenges, researchers are developing more rigorous evaluation frameworks:
Standardized benchmarks: Initiatives like MLPerf aim to provide consistent evaluation across different hardware and software stacks [Reddi et al., 2020]. These benchmarks define specific tasks, datasets, and evaluation metrics to enable fair comparisons. However, Dehghani et al. [2021] argue that while useful, fixed benchmarks can lead to overfitting to specific datasets and metrics, potentially hindering innovation. They propose dynamic benchmarks that evolve over time to mitigate this issue.

Controlled studies: Carefully designed experiments that vary only one factor at a time while keeping others constant [Bello et al., 2021]. This approach, inspired by scientific experimental design, aims to isolate the impact of individual components. Lipton [2017] emphasize the importance of such controlled experiments in machine learning, arguing for a more rigorous empirical methodology in the field.

Meta-analysis: Aggregating results across multiple studies to identify consistent trends and effects. Henderson et al. [2018] demonstrate the value of this approach in reinforcement learning, showing how meta-analysis can reveal robust trends that may not be apparent in individual studies. Similarly, Hutson [2018] argue for the importance of meta-analyses in AI research to improve reproducibility and identify genuine advances.

This complexity in performance attribution highlights the need for more nuanced evaluation approaches in deep learning. Researchers must consider the interplay of various factors and employ rigorous methodologies to gain meaningful insights into model performance and improvements. Amershi et al. [2019] propose a set of guidelines for human-AI interaction that emphasize the importance of comprehensive evaluation across different contexts and user groups. Recent work has started to explore causal inference techniques to better understand the factors contributing to model performance. Scholkopf et al. [2021] argue for the importance of causal models in machine learning, suggesting that causal reasoning could help disentangle the complex relationships between different components of deep learning systems.

Overall, accountability and rigorous evaluation are fundamental to advancing deep learning research and its applications, ensuring that models are not only effective, but also transparent and trustworthy. We will now focus on some specialized open research topics, presenting their overall setup and highlighting their specific evaluation challenges.

2.3 Evaluation in EFCIL

2.3.1 Class-Incremental Learning (CIL)

Parisi et al. [2019] define continual learning as "*adaptive algorithm capable of learning from a continuous stream of information, with such information becoming progressively available over time and where the number of tasks to be learned [...] are not predefined*". Class-incremental learning is a subset of continual learning that focuses specifically on the classification task.

Definition: Class-Incremental Learning

Class-Incremental Learning (CIL) is a machine learning paradigm that enables classification models to continuously incorporate new classes over time without full retraining on all data.

The primary purpose of CIL is to enable machine learning systems to adapt to evolving environments where new classes emerge sequentially [Parisi et al., 2019; Rebuffi et al., 2017; Lange et al., 2019]. In a CIL scenario, a model is initially trained on a set of classes (or pre-trained with external data) and then presented with new classes in subsequent stages. The goal is to learn these new classes while retaining knowledge of previously learned ones [Masana et al., 2021]. This approach is particularly relevant in real-world applications where data arrives in streams, and full access to past data may be limited or impossible [Hayes and Kanan, 2022; Van de Ven and Tolias, 2019].

2.3.1.1 Challenges

The main challenge in CIL is the stability-plasticity dilemma, which refers to the fundamental trade-off between a model's ability to acquire new knowledge (plasticity) and its capacity to retain previously learned information (stability) [Mermillod et al., 2013]. This dilemma is at the core of the catastrophic forgetting problem in neural networks [Mccloskey and Cohen, 1989; French, 1999]. When a model is too plastic, it rapidly adapts to new information but risks overwriting or disrupting existing knowledge, leading to catastrophic forgetting [Kirkpatrick et al., 2017]. Conversely, if a model is too stable, it maintains its existing knowledge well but struggles to incorporate new information effectively [Zenke et al., 2017]. Striking the right balance between stability and plasticity is crucial for successful CIL systems [Parisi et al., 2019; Schwarz et al., 2018].

Challenges in CIL

Key challenges in CIL include :

- The stability-plasticity dilemma
- Limited or no memory for storing old data: Many CIL scenarios restrict or prohibit the storage of old data, making it difficult to rehearse past examples [Hayes and Kanan, 2020]. As new classes are added, the model may become biased towards newer classes due to the recency of their training data [Wu et al., 2019].
- Developing appropriate metrics to assess both the model's ability to learn new classes and retain knowledge of old ones is crucial for evaluating CIL systems [Masana et al., 2021].

Addressing these challenges is essential for developing robust CIL systems that can adapt to new information while maintaining performance on previously learned tasks. Various approaches have been proposed, including rehearsal methods, parameter regularization, and architectural strategies, each with its own trade-offs in terms of performance, memory requirements, and computational cost [Lange et al., 2019; Wu et al., 2021a; Wang et al., 2024].

2.3.2 Exemplar-Free Class-Incremental Learning (EFCIL)

2.3.2.1 Presentation

Definition: Exemplar-Free Class-Incremental Learning

Exemplar-Free Class-Incremental Learning (EFCIL) is a variant of CIL where the model cannot store or revisit any examples from previously learned classes.

This constraint introduces additional challenges beyond those faced in standard CIL [Hayes and Kanan, 2020; Petit et al., 2023]. Unlike rehearsal-based CIL methods, EFCIL algorithms cannot leverage stored examples to mitigate catastrophic forgetting [Rebuffi et al., 2017; Belouadah and Popescu, 2020]. Without access to old data, the feature space learned by the model may gradually shift, leading to increased forgetting and decreased performance on earlier classes [Yu et al., 2020]. Traditional knowledge distillation techniques used in CIL become challenging without exemplars, requiring alternative approaches to preserve past knowledge [Li and Hoiem, 2016; Hou et al., 2019]. Additionally, maintaining clear decision boundaries between old and new classes becomes more difficult without exemplars to refine these boundaries [Zhao et al., 2020].

Recent advancements in EFCIL have focused on developing more sophisticated methods to mitigate catastrophic forgetting without relying on stored examples. These approaches often leverage innovative techniques in representation learning and knowledge preservation. For instance, Wu et al. [2021a] and Zhu et al. [2021b] explored the use of self-supervised learning to improve the generalizability of feature representations across incremental steps. Others, like Petit et al. [2023], have proposed novel ways to simulate or synthesize information about past classes using only statistical summaries. Simulating past and future classes has also gained popularity in recent years by leveraging the knowledge of class names and new conditional generation methods [Jodelet et al., 2023; Feillet et al., 2024]. Approaches using pre-trained models, particularly large language models or vision transformers, have also gained traction, demonstrating improved performance in EFCIL scenarios [Wang et al., 2022a; Smith et al., 2023]. Additionally, some recent works have focused on developing more robust distance-based classifiers that can better

handle the challenges of class separation in fixed feature spaces [Hayes and Kanan, 2020; Goswami et al., 2024]. These diverse strategies reflect the ongoing efforts to push the boundaries of what's achievable in EFCIL, making it increasingly viable for real-world applications where data storage is constrained or prohibited [Hayes and Kanan, 2022; Ravaglia et al., 2021].

2.3.2.2 Importance in Real-World Applications

EFCIL addresses significant challenges in scenarios where data storage is constrained, privacy is a concern, or rapid adaptation to new information is necessary. Several factors underscore its importance. As Hayes and Kanan [2022] point out, in applications involving sensitive data such as medical records or financial transactions, storing old examples may contravene privacy regulations. EFCIL aligns with regulations like the GDPR [GDP] and the European AI Act [UE, 2024], which encourage algorithms that don't require personal data storage [Verma et al., 2023]. Ravaglia et al. [2021] and Pellegrini et al. [2021] note that for edge computing or mobile devices with limited storage, maintaining a growing set of exemplars is often impractical. Belouadah et al. [2021] and Hayes and Kanan [2022] highlight that EFCIL methods offer relatively low update costs in terms of memory and execution time. Furthermore, EFCIL enables models to incorporate new information dynamically without full retraining, which is computationally expensive and time-consuming, especially in rapidly evolving environments.

Examples of EFCIL Applications

- Autonomous driving: on-board systems must learn new types of objects and road conditions in real-time, with limited storage capacity for past scenarios.
- Medical diagnostics: models need to adapt to new variants of diseases without retaining sensitive patient data.
- Manufacturing: systems must incorporate new products or materials efficiently without storing extensive historical data.
- Social media content analysis: platforms process thousands of posts per second, making it impractical to store all historical data for retraining.
- Fraud detection in finance: institutions need to adapt to new fraud patterns quickly without retaining sensitive transaction data.

In dynamic environments where past data quickly becomes obsolete, EFCIL facilitates efficient model adaptation without reliance on potentially outdated examples, as noted by Aljundi et al. [2019]. Zhu et al. [2022] and Hayes and Kanan [2022] suggest that by developing effective EFCIL methods, researchers aim to create more versatile and deployable incremental learning systems capable of operating under strict data retention constraints. This makes machine learning more applicable in sensitive and resource-constrained environments. Additionally, EFCIL aligns with the use of off-the-shelf pre-trained models whose training data is not always publicly available, further expanding its practical applications [Oquab et al., 2023].

Summary of Applications Scenarios of EFCIL

EFCIL is valuable in privacy-sensitive domains, resource-constrained environments, scenarios with rapid data streams, and applications subject to strict legal and ethical considerations. It enables continuous learning without storing old data, addressing both practical and privacy concerns in various fields including healthcare, finance, manufacturing, and autonomous systems.

2.3.3 Existing Evaluation Approaches for EFCIL

2.3.3.1 Common Experimental Setups

Evaluation of EFCIL algorithms typically involves specific experimental setups designed to simulate real-world incremental learning scenarios. These setups often include dataset splitting, where full datasets are divided into smaller subsets, each representing a batch of new classes [Rebuffi et al., 2017; Hou et al., 2019].

Common Datasets for EFCIL Evaluation

Some common datasets in the evaluation of EFCIL include:

- CIFAR-100 [Krizhevsky, 2009]
- ImageNet [Deng et al., 2009]
- CUB-200 [Belouadah et al., 2021]

The learning process is typically divided into multiple incremental steps, with each step introducing a new set of classes [Castro et al., 2018; Wu et al., 2019]. In EFCIL, no exemplars are allowed, in contrast to standard CIL where a limited memory budget is often set to limit the number of stored exemplars [Hayes and Kanan, 2020; Rebuffi et al., 2017].

Many setups use a larger initial set of base classes, followed by smaller increments of new classes [Hou et al., 2019; Belouadah and Popescu, 2020]. To assess generalization across different domains, Masana et al. [2021] and Petit et al. [2023] suggest evaluating on multiple datasets.

2.3.3.2 Standard Metrics

Several metrics are commonly used to evaluate the performance of CIL and EFCIL algorithms. **Average Incremental Accuracy** measures the overall classification accuracy after each incremental step, averaged over all steps [Rebuffi et al., 2017; Hou et al., 2019]. It is defined as:

$$\text{Average Incremental Accuracy} = \frac{1}{T}\sum_{i=1}^{T} A_i \tag{2.1}$$

where T is the total number of incremental steps and A_i is the accuracy on all seen classes after the i-th step.

Chaudhry et al. [2018] introduce the complementary measure **Forgetting**, which measures the decrease in performance on previously learned classes. It can be calculated as::

$$\text{Forgetting} = \frac{1}{T-1}\sum_{i=1}^{T-1} \max_{j\in\{1,\dots,i\}} (A_{j,j} - A_{T,j}) \tag{2.2}$$

where $A_{i,j}$ is the accuracy on the j-th task after learning the i-th task.

Additionally, a few other metrics can be found in the literature. **Learning Accuracy** evaluates the model's ability to learn new classes at each step [Rebuffi et al., 2017]. **Final Model Accuracy** is the overall accuracy of the final model on all classes [Wu et al., 2019]. Lopez-Paz and Ranzato [2022] propose **Backward Transfer**, which measures how learning new tasks affects the performance on previous tasks, and **Forward Transfer**, which assesses how learning previous tasks influences the performance on new tasks.

These metrics provide a comprehensive view of a model's performance in incremental learning scenarios, capturing both its ability to learn new information and retain old knowledge. However, it is essential to note that the relative importance of these metrics may vary depending

on the specific application and requirements of the incremental learning system [Masana et al., 2021; Belouadah et al., 2021].

2.3.4 Shortcomings of Current Evaluation Methods

While existing evaluation approaches have provided valuable insights into CIL and EFCIL algorithms, several limitations have been identified in recent literature:

2.3.4.1 Lack of Comprehensive Analysis Across Diverse Datasets

Many studies focus on a small set of popular datasets (e.g., CIFAR-100, ImageNet) [Belouadah et al., 2021; Masana et al., 2021]. This narrow focus can lead to overfitting on specific dataset characteristics and may not adequately represent the diversity of real-world data. Wu et al. [2019] argue for more extensive evaluations on large-scale datasets to better simulate real-world scenarios. Delange et al. [2021] emphasize the need for more realistic and challenging datasets that better reflect the complexities of continual learning in practical settings. Van de Ven and Tolias [2019] highlight the importance of considering different types of CIL setups, introducing three continual learning scenarios.

Data-Induced Shortcomings in EFCIL

Current EFCIL evaluations often suffer from limited dataset variety, domain specificity, and scale limitations, potentially overlooking real-world challenges.

2.3.4.2 Limited Consideration of Initial Training Strategies

The influence of pre-training strategies on EFCIL performance is a decisive yet often neglected aspect of evaluation. Different pre-training approaches, such as supervised learning on large datasets or self-supervised methods, can significantly affect the model's ability to adapt to new classes and retain knowledge of old ones [Wu et al., 2022]. For instance, Neyshabur et al. [2021] suggest that models pre-trained on diverse datasets might exhibit better transfer learning capabilities, potentially improving their performance in EFCIL scenarios.

Model architecture choices play a crucial role in the success of EFCIL algorithms, but their impact is not consistently analyzed across studies [Masana et al., 2021]. The trade-offs between different architectures, such as convolutional neural networks (CNNs) and transformers, in terms of their ability to learn new classes incrementally while maintaining performance on old classes, remain understudied. This lack of comprehensive analysis makes it challenging to determine which architectures are best suited for specific EFCIL tasks or domains.

Furthermore, the initial data distribution and quantity can have far-reaching effects on the entire incremental learning process. The number of classes and examples in the initial training set, as well as their diversity and representativeness, can significantly influence the model's ability to generalize and adapt to new classes [Belouadah et al., 2021]. However, many studies do not systematically vary these factors or analyze their impact, potentially missing important insights into the robustness and scalability of EFCIL algorithms in different data regimes.

Addressing these limitations in future research could provide valuable insights into the design of more effective and generalizable EFCIL algorithms, better equipped to handle the complexities of real-world incremental learning scenarios.

Initial Training-Induced Shortcomings in EFCIL

Current EFCIL evaluations often overlook:

- The impact of various pre-training strategies on performance
- The influence of model architecture choices on incremental learning capabilities
- The effects of initial data distribution and quantity on subsequent learning steps

These limitations can lead to an incomplete understanding of EFCIL algorithm behavior and limit real-world applicability.

2.3.4.3 Need for More Rigorous Statistical Analysis

The field of EFCIL research currently faces several challenges related to statistical rigor. Many studies report performance improvements without conducting rigorous statistical validation, potentially overstating the significance of their results. This lack of statistical testing makes it difficult to determine whether observed differences between algorithms are truly meaningful or simply due to chance.

Detailed ablation studies, essential for isolating the effects of different components in CIL/EFCIL algorithms, are often missing from published works, as in many other fields of ML [Mousavi et al., 2020]. These studies are vital for understanding which aspects of an algorithm contribute most significantly to its performance and for guiding future improvements.

Some Consequences of Inconsistent Reporting

Variations in reporting metrics and experimental setups across studies can lead to:

- Difficulty in directly comparing results
- Potential misinterpretation of algorithm effectiveness
- Challenges in reproducing reported findings

The impact of random initializations and data ordering on performance is not consistently reported, potentially masking the stability of algorithms [Prabhu et al., 2020]. This oversight can lead to overestimating the robustness of certain approaches, as their performance may be highly dependent on specific initializations or data presentations.

Statistical Shortcomings in EFCIL Research

Current EFCIL evaluations present:

- A lack of statistical significance testing
- Insufficient ablation studies
- Inconsistent reporting of metrics and setups
- Limited analysis of performance variances

Several approaches have been proposed in the literature to address these statistical shortcomings. Jiménez-Guarneros and Alejo-Eleuterio [2022] emphasize the importance of using statistical significance tests when comparing EFCIL algorithms, suggesting the use of paired t-tests or Wilcoxon signed-rank tests to assess whether performance differences are statistically significant. To address the issue of inconsistent reporting, Rodríguez et al. [2018] suggest adopting standardized reporting practices, including the use of consistent evaluation metrics and clear descriptions of experimental setups. This approach can facilitate more meaningful

comparisons across different studies.

By addressing these shortcomings, researchers can develop more reliable and generalizable insights into the performance and applicability of CIL and EFCIL algorithms in real-world scenarios [Feillet et al., 2023; Petit et al., 2023]. This improved rigor will not only enhance the quality of individual studies but also contribute to the overall advancement of the field by enabling more meaningful comparisons and insights across different EFCIL approaches.

2.4 Biases in Face Recognition

2.4.1 Face verification

Face recognition technology has become an increasingly important topic in computer vision and biometrics over the past few decades [Zhao et al., 2003]. This broad field encompasses various tasks related to the automatic processing and analysis of human faces in digital images or video streams. These tasks include face detection, face alignment, face identification, and face verification, each serving distinct purposes within the broader context of facial analysis [Jain and Li, 2011].

Among these tasks, face verification has gained particular prominence due to its wide-ranging applications in security, authentication, and identity management systems [Jain et al., 2004]. As a specific subset of face recognition, face verification focuses on a unique challenge within the field [Zhao et al., 2003].

2.4.1.1 Definition and Purpose

Face Verification

Face verification is a biometric technology that determines whether two face images belong to the same individual by comparing their representations [Jain and Li, 2011].

The primary purpose of face verification is to determine whether two given facial images represent the same person without necessarily knowing *a priori* who that person is. This process typically involves:

1. Receiving two facial images for comparison.
2. Extracting facial features or representations from both images.
3. Computing the similarity or distance between the extracted features.
4. Deciding whether the two images represent the same person based on the computed similarity or distance.

Face verification systems aim to achieve high accuracy by minimizing false acceptances (incorrectly verifying an impostor) and false rejections (incorrectly rejecting a genuine user) [Phillips et al., 2012].

2.4.1.2 Applications

Face verification technology has found widespread applications across various domains, primarily due to its non-intrusive nature and increasing reliability [Masi et al., 2018].

Example of Applications of Face Verification

Some key areas of application include:

- Security and access control
- Law enforcement and surveillance
- Personal device authentication
- Financial services
- Border control and immigration

Indeed, face verification is increasingly used for secure access to physical spaces such as offices, restricted areas, and smart homes [Masi et al., 2018]. It offers a touchless and efficient method of identity verification, enhancing both security and convenience.

In law enforcement, Purshouse and Campbell [2022] note that face verification aids in identifying suspects, missing persons, and potential security threats. However, its use in this domain has raised significant ethical and privacy concerns.

Example: Airport Security

Many airports now employ face verification systems for passenger identification and streamlined boarding processes, enhancing both security and efficiency in air travel [Labati et al., 2016].

Smartphones and personal computers increasingly use face verification as a secure and convenient user authentication method [Chen et al., 2018]. This application has become particularly prevalent in mobile devices, offering an alternative or complement to traditional PIN codes or fingerprint scans.

Banks and financial institutions are adopting face verification for secure customer authentication in various transactions, including ATM withdrawals and online banking services [Amato et al., 2019].

Face verification systems are deployed at international borders to expedite traveler processing while maintaining high-security standards [Labati et al., 2016]. These systems can quickly verify a traveler's identity against their passport or visa information.

The diverse applications of face verification technology highlight its potential to enhance security, streamline processes, and improve user experiences across various sectors. However, as the technology continues to evolve and its use becomes more widespread, it also raises important questions about privacy, bias, and ethical implementation that need to be carefully addressed [Van Noorden, 2020].

2.4.2 Fairness Challenges in Face Verification

As face verification systems become more prevalent in various applications, concerns about their fairness and potential biases have come to the forefront of research and public discourse. These challenges stem from observed performance disparities across different demographic groups and raise significant ethical and legal concerns.

2.4.2.1 Performance Disparities Across Demographic Groups

As in any image classification problem, a face verification model's performance is heavily influenced by various image characteristics, such as image quality and the pose (i.e., orientation) of the person. However, some of these characteristics are protected attributes, which are demographic or personal features that are legally protected from discrimination.

Definition: Protected Attribute

A protected attribute is a characteristic of an individual that should not be used as a basis for discrimination, as defined by laws and regulations [Barocas et al., 2023].

Key Demographic Factors Affecting Face Verification Performance

Some of the protected demographic attributes known to affect face verification performance include:

- Ethnicity
- Gender
- Age
- Skin tone

Multiple studies have demonstrated that face verification systems can exhibit varying levels of accuracy depending on these protected attributes. Buolamwini and Gebru [2018] and Grother et al. [2019] have shown that certain racial or ethnic groups, particularly individuals with darker skin tones, often experience higher error rates. Gender-based differences in accuracy have also been observed, with Albiero et al. [2020] noting lower performance for women in many systems. Additionally, Terhörst et al. [2021] found age-related variations in performance, with systems generally performing worse for very young or very old individuals.

Example: Gender Shades Study

The "Gender Shades" study by Buolamwini and Gebru [2018] found that commercial gender classification systems had error rates of up to 34.7% for darker-skinned females compared to just 0.8% for lighter-skinned males.

These performance disparities can lead to unfair outcomes in real-world applications, potentially disadvantaging certain demographic groups and reinforcing existing societal biases.

2.4.2.2 Ethical and Legal Concerns

The presence of biases in face verification models is an example of algorithmic bias.

Algorithmic Bias

Algorithmic bias refers to systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one arbitrary group of users over others [Friedman and Nissenbaum, 1996].

The observed fairness issues in face verification systems raise significant ethical and legal concerns. Unequal performance across demographic groups can lead to discriminatory practices, especially when these systems are used in high-stakes applications such as law enforcement or access to services. Purshouse and Campbell [2022] highlight that biased face verification

systems in law enforcement can exacerbate existing racial disparities in policing, potentially leading to wrongful arrests or disproportionate surveillance of minority communities.

The widespread use of face verification technology also raises questions about individual privacy rights and the potential for unauthorized surveillance. Van Noorden [2020] argue that the pervasive use of facial recognition in public spaces can create a chilling effect on free speech and assembly, as individuals may feel constantly monitored. This concern is particularly acute in contexts where face verification is used for social control or political repression [Rezende, 2020].

Another significant issue is the collection and use of facial data for training and operating these systems, which often occurs without explicit consent. This practice raises concerns related to data protection regulations such as the General Data Protection Regulation [GDP] in the European Union. Staunton et al. [2019] point out that the use of facial recognition technologies may conflict with GDPR principles, particularly regarding data minimization and purpose limitation, and leading to the withdrawal of some face recognition training datasets.

The complexity of deep learning models used in face verification poses challenges for explainability and accountability. Taskiran et al. [2020] emphasize that the "black box" nature of these models makes it difficult to explain their decision-making process, leading to concerns about accountability and the right to explanation. This lack of transparency can be particularly problematic when face verification systems are used in legal or administrative decision-making processes.

Lastly, there are growing concerns about the potential misuse of face verification technology for mass surveillance or social control. The integration of facial recognition with other surveillance technologies can create powerful tools for tracking and profiling individuals, potentially infringing on civil liberties and human rights [Smith et al., 2021].

Summary of Ethical and Legal Challenges

The main risks linked to the presence of biases in face verification systems include:

- Potential for discrimination and reinforcement of societal biases
- Infringement on privacy rights
- Issues with data protection and consent
- Lack of transparency and accountability in decision-making
- Risk of misuse for surveillance

Addressing these fairness challenges requires a multifaceted approach involving technical improvements in algorithms, diverse and representative training data, rigorous testing across demographic groups, and the development of clear ethical guidelines and legal frameworks for the deployment of face verification systems [Wang et al., 2019a; Sarridis et al., 2023b]. As Raji et al. [2020] argue, this may also involve rethinking the appropriateness of face verification technology in certain high-risk applications and considering alternative solutions that better protect individual rights and societal values.

2.4.3 Existing Evaluation Approaches for Face Verification Fairness

Identifying and mitigating biases implies having a rigorous experimental and evaluation setup.

2.4.3.1 Common Experimental Setups

Using specially curated datasets to evaluate face verification fairness is standard in the domain. These datasets are designed to represent diverse demographic groups and challenging verification scenarios.

Balanced Face Datasets

A **balanced** dataset is specifically designed to have equal or controlled representation of different demographic groups, enabling a more robust evaluation of algorithmic fairness in face verification.

Some of the most commonly used datasets for fairness evaluation include:

- **RFW (Racial Faces in the Wild)** [Wang et al., 2019a]: A dataset balanced across four racial groups: African, Asian, Caucasian, and Indian.
- **BFW (Balanced Faces in the Wild)** [Robinson et al., 2023]: Provides a balance across gender and ethnicity categories.
- **FAVCI2D (Face Verification with Challenging Imposters and Diversified Demographics)** [Popescu et al., 2022]: Focuses on challenging impostor pairs and diverse demographics. Here, the diversification is based on the gender and geographical origin of the identities.

The RFW Dataset Composition

RFW contains 40,607 images of 11,950 identities:

- Caucasian: 10,196 images
- African: 10,415 images
- Asian: 9,688 images
- Indian: 10,308 images

This nearly balanced composition allows for fair comparison across racial groups.

These datasets enable researchers to assess face verification performance across different demographic groups and identify potential biases.

2.4.3.2 Standard Performance Metrics

Several standard metrics are commonly used to evaluate the overall performance of face verification systems. These metrics provide a comprehensive view of a system's performance but may not fully capture fairness across demographic groups.

Average Accuracy is a fundamental metric that measures the overall correctness of the face verification system. It can be calculated in two ways: *Micro-averaged accuracy* gives equal weight to each sample, regardless of class imbalance. It is calculated as the total number of correct predictions divided by the total number of predictions:

$$\text{Micro-averaged Accuracy} = \frac{\text{Total Correct Predictions}}{\text{Total Predictions}}$$

Macro-averaged accuracy gives equal weight to each attribute $a \in A$, which is particularly useful when dealing with imbalanced datasets. It is calculated as the average of the accuracies for each attribute:

$$\text{Macro-averaged Accuracy} = \frac{1}{|A|} \sum_{a \in A} \text{Accuracy}_a$$

where Accuracy_a is the accuracy for the data with attribute a.

True Match Rate (TMR), also known as True Positive Rate (TPR) or Recall, represents the proportion of correct positive matches among all actual positive pairs. It is calculated as:

$$\text{TMR} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

False Match Rate (FMR), also known as False Positive Rate (FPR), is the proportion of incorrect positive matches among all actual negative pairs. It is calculated as:

$$\text{FMR} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

The FMR is particularly critical in security applications, as it represents the likelihood of the system incorrectly authenticating an impostor [Ho et al., 2020a].

These performance-oriented metrics provide a comprehensive view of a system's overall effectiveness. However, they may not fully capture fairness across demographic groups, as high overall performance can mask significant disparities in accuracy for different subpopulations. Therefore, when evaluating face verification systems, it's important to consider these standard metrics in conjunction with specific fairness metrics to ensure equitable performance across all demographic groups.

2.4.3.3 Fairness Metrics

To specifically address fairness concerns, several metrics have been developed:

Fairness Metrics

Quantitative measures designed to assess the equitability of a face verification system's performance across predefined different demographic groups.

Demographic Parity [Agarwal et al., 2019]: Requires that the probability of a positive outcome is the same for all demographic groups. Mathematically, for a predictor $\hat{Y}$ and a protected attribute A, demographic parity is achieved if:

$$\mathbb{P}(\hat{Y} = 1|A = a) = \mathbb{P}(\hat{Y} = 1|A = b), \quad \forall a, b \tag{2.3}$$

It is often measured using:

- Demographic Parity Difference (DPD), the difference between the largest and the smallest group-level selection rate $\mathbb{P}(\hat{Y} = 1|A = a)$ across values of a.
- Demographic Parity Ratio (DPR), the ratio between the smallest and the largest group-level selection rate $\mathbb{P}(\hat{Y} = 1|A = a)$ across values of a

Equalized Odds [Hardt et al., 2016]: Requires that both TMR and FMR are equal across all demographic groups. It is achieved if:

$$\mathbb{P}(\hat{Y} = 1|A = a, Y = y) = \mathbb{P}(\hat{Y} = 1|A = b, Y = y), \quad \forall a, b, y \tag{2.4}$$

It's typically measured using:

- Equalized Odds Difference (EOD), the biggest value between the biggest difference in TMR across groups, i.e. between $\mathbb{P}(\hat{Y} = 1|A = a, Y = 1)$ across a, and the biggest difference in FMR across groups, i.e. between $\mathbb{P}(\hat{Y} = 1|A = a, Y = 0)$ across a.
- Equalized Odds Ratio (EOR), which is defined similarly as EOD, but with ratios instead of differences.

Degree of Bias (DoB) [Gong et al., 2020]: Measures the standard deviation of accuracy across different subgroups.

These metrics help researchers and practitioners quantify the fairness of face verification systems and identify areas where bias mitigation efforts should be focused. By combining standard performance metrics with these fairness metrics, researchers can better understand a face verification system's behavior across different demographic groups. This holistic approach to evaluation is essential for developing more equitable and reliable face verification technologies.

2.4.4 Shortcomings of Current Evaluation Methods

While existing evaluation approaches provide valuable insights into fairness in face verification systems, several limitations can be identified:

2.4.4.1 Limitations in Dataset Representation

Current face verification fairness evaluations often suffer from significant limitations in dataset representation. While balanced datasets like RFW [Wang et al., 2019a], FAVCI2D [Popescu et al., 2022], and BFW [Robinson et al., 2020] provide a solid foundation for fairness evaluation, many studies still rely on unbalanced datasets that do not adequately represent diverse demographic groups.

The limited demographic diversity in existing datasets is a major concern. Merler et al. [2019] highlight that many popular face recognition datasets are heavily skewed towards certain ethnic groups, particularly Caucasians, leading to potential biases in model training and evaluation. This lack of diversity can result in models that perform well on majority groups but fail to generalize to underrepresented populations.

Another issue is the potential for biases in dataset curation and annotation. Grother et al. [2019] point out that collecting and labeling face images can introduce unintended biases, such as differences in image quality or facial expressions across demographic groups. These biases in the dataset can then propagate to the models trained on them, leading to unfair performance disparities.

Creating truly representative balanced datasets presents significant challenges. Kärkkäinen and Joo [2019] discuss the difficulties in obtaining diverse, high-quality face images while ensuring proper consent and ethical data collection practices. Additionally, Klare et al. [2012] note that achieving balance across multiple demographic dimensions simultaneously (e.g., age, gender, and ethnicity) is particularly challenging and often leads to trade-offs in dataset design.

The lack of comprehensive demographic metadata in many widely used datasets further complicates fairness evaluations. Buolamwini and Gebru [2018] emphasize that without detailed information about the demographic characteristics of individuals in the dataset, conducting thorough analyses across different demographic groups becomes challenging, if not impossible.

These unbalanced datasets, often created without explicit consideration for fairness evaluation, may lead to biased assessments of face verification systems, and evaluations conducted on non-representative datasets can provide an incomplete or misleading picture of a system's real-world performance across diverse populations.

Dataset-Related Shortcomings in Face Verification Fairness Evaluation

Current face verification fairness evaluations often suffer from:

- Limited demographic diversity in existing datasets
- Potential biases in dataset curation and annotation

- Challenges in creating truly representative balanced datasets
- Lack of comprehensive demographic metadata

These issues may lead to an incomplete assessment of real-world fairness challenges.

Addressing these limitations in dataset representation is a first step towards developing more robust and fair face verification systems. It requires concerted efforts from the research community to create more diverse and well-documented datasets, as well as to develop evaluation methodologies that can account for and mitigate these inherent biases.

2.4.4.2 Limitations of Current Fairness Metrics

While metrics like Demographic Parity, Equalized Odds, and Degree of Bias provide valuable insights into the fairness of face verification systems, they have several limitations that may lead to an incomplete assessment of system fairness.

One significant limitation is that these metrics primarily focus on outcome fairness, without providing insights into the underlying causes of bias. This surface-level approach to fairness can mask deeper issues in the model's decision-making process. For instance, a model might achieve demographic parity by making different types of errors for different groups, rather than by genuinely eliminating bias.

Moreover, these metrics often fail to capture the nuanced and multifaceted nature of fairness. Mehrabi et al. [2021] point out that fairness is a complex, context-dependent concept that may not be fully encapsulated by a single metric or even a combination of existing metrics. Different fairness criteria can sometimes be in tension with each other, leading to situations where improving one aspect of fairness can worsen another.

Another limitation is the binary nature of many fairness metrics. Metrics like demographic parity and equalized odds [Agarwal et al., 2018] are typically computed based on binary outcomes (correct or incorrect identification), which may not capture the full spectrum of a model's behavior. This binary approach can obscure important nuances in model performance across different subgroups.

Furthermore, current fairness metrics often fail to account for intersectionality. Buolamwini and Gebru [2018] demonstrate that face recognition systems can exhibit compounded biases at the intersection of multiple protected attributes (e.g., ethnicity and gender), which may not be fully captured by metrics that consider these attributes in isolation.

An additional concern is that these metrics do not provide insights into the structure of the model's learned representations. Zemel et al. [2013] argue that truly fair models should learn representations that are invariant to protected attributes, but current fairness metrics do not directly measure this aspect of model behavior.

Lastly, Kearns et al. [2018] raise concerns about the potential for "fairness gerrymandering," where models are optimized to perform well on specific fairness metrics without addressing underlying biases. This highlights the need for a more holistic approach to fairness evaluation that goes beyond simple metric optimization.

Metric-Related Shortcomings in Face Verification Fairness Evaluation

Current fairness metrics have limitations:

- Focus on outcome fairness without explaining the origin of biases
- May not capture the full complexity and context-dependence of fairness

- Often based on binary outcomes, missing nuances in model behavior
- Fail to account for intersectionality and compounded biases
- Do not provide insights into the fairness of learned representations
- Vulnerable to "fairness gerrymandering" through metric optimization

These limitations may lead to an incomplete assessment of system fairness.

Addressing these limitations requires developing more sophisticated fairness metrics and evaluation frameworks that can provide a more comprehensive and nuanced understanding of bias in face verification systems.

2.4.4.3 Need for More Comprehensive Statistical Analysis

Current fairness evaluations in face verification systems often lack the robust statistical analysis necessary to draw reliable conclusions about fairness improvements. This shortcoming can lead to misinterpretations of results and hinder progress in developing truly fair systems.

One significant issue is the lack of robust statistical significance testing for fairness metrics. Friedler et al. [2019] highlight that many studies report improvements in fairness metrics without conducting rigorous statistical validation. This oversight makes it difficult to determine whether observed differences in fairness are truly meaningful or simply due to chance. They argue for the use of appropriate statistical tests to assess the significance of fairness improvements.

The inconsistent reporting of fairness metrics across studies poses another challenge. Mehrabi et al. [2021] note that the lack of standardization in fairness evaluation makes it difficult to compare results across different studies and draw generalizable conclusions. Standardized reporting practices should facilitate more meaningful comparisons and meta-analyses.

An often overlooked aspect is the understanding of interactions between tested characteristics. Analyzing how different demographic attributes interact to influence system performance is important since univariate analyses of individual protected attributes may miss complex patterns of bias that emerge from the interaction of multiple characteristics.

Finally, Mitchell et al. [2021] argue for the importance of causal analysis in fairness evaluations. They suggest that many current approaches to fairness assessment rely on correlational analyses, which may not capture the true causal relationships underlying observed biases. Developing causal models of fairness could provide deeper insights into the sources of bias and more effective strategies for mitigation.

Statistical Shortcomings in Face Verification Fairness Evaluation

Current fairness evaluations often lack:

- Robust statistical significance testing for fairness metrics
- Comprehensive analysis of performance variations across demographic groups
- Consistent reporting of fairness metrics across studies
- Understanding of the interactions between tested characteristics
- Rigorous testing protocols beyond simple accuracy metrics
- Causal analysis of fairness and bias

These limitations can lead to misinterpretations of results and hinder progress in developing truly fair systems.

Addressing these shortcomings requires a fundamental shift in our analytical approach. They highlight a need to develop more robust statistical methodologies and standardized evaluation

protocols to adequately capture the nuances of fairness in these complex systems.

To address all of these shortcomings related to dataset, fairness metrics, and a lack of statistical considerations, face verification evaluation would benefit from:

- Developing more diverse and representative datasets for fairness evaluation.
- Exploring additional fairness metrics that capture different aspects of bias.
- Not only focus on fairness of the performance but also on understanding how the latent space structure affects these biases.
- Incorporating robust statistical analyses, including significance tests for fairness metrics.
- Standardizing reporting practices for fairness evaluations to facilitate comparisons across studies.
- Studying the interplay between tested characteristics.

By addressing these limitations, researchers can develop more comprehensive and reliable approaches to evaluating fairness in face verification systems, leading to more equitable and robust technologies.

2.5 Evaluation of Recommender Systems

2.5.1 Overview of Recommender Systems

Recommender systems have become an integral part of our digital experience, playing an important role in filtering and personalizing content across various domains, from e-commerce to entertainment and news consumption [Nilashi et al., 2013; Konstan, 2004]. These systems aim to predict users' preferences and provide personalized suggestions, effectively addressing the information overload problem that has become increasingly prevalent in the digital age. The history of recommender systems can be traced back to the mid-1990s, with the emergence of collaborative filtering techniques [Konstan and Riedl, 2012]. As the internet grew and digital platforms proliferated, the need for effective recommendation algorithms became more pressing. The field gained significant attention with the Netflix Prize competition (2006-2009), which spurred innovation in collaborative filtering algorithms [Bennett et al., 2007a].

Definition: Types of Recommender Systems

- **Content-based filtering** recommends items similar to those a user has liked in the past, based on item features [Lops et al., 2019].
- **Collaborative filtering** recommends items based on the preferences of users with similar tastes [Ekstrand et al., 2011; Afoudi et al., 2018].
  - Memory-based: uses user-item interactions directly.
  - Model-based: learns latent factors to make predictions.
- **Hybrid approaches** combine content-based and collaborative filtering techniques [Burke, 2002].
- **Knowledge-based** approaches Use explicit knowledge about user preferences, item features, and recommendation criteria[Burke, 2000].

In recent years, the field has seen significant advancements with the use of deep learning techniques. While some works, like Zhang et al. [2019], claim these new methods lead to more sophisticated and accurate recommendations, others, like Ferrari Dacrema et al. [2019], advance that traditional approaches can outperform deep learning ones if correctly trained. However, as

recommender systems become more pervasive and influential in shaping user behavior, there is an increasing focus on evaluating not just their accuracy but also their impact on user experience, diversity, and fairness [Ping et al., 2024; Diricic et al., 2023].

The importance of recommender systems extends beyond user satisfaction. They play an essential role in driving user engagement, increasing sales in e-commerce, and helping users discover new content in large item catalogs. As such, they have become a key component of many online platforms' business strategies [Pavlidis, 2019].

Examples of Real-world Recommender Systems

- **E-commerce:** Amazon uses item-to-item collaborative filtering, among other techniques, to suggest products [Jannach et al., 2012].
- **Streaming services:** Netflix employs various algorithms, including matrix factorization, to recommend movies and TV shows [Bennett et al., 2007a].
- **Music platforms:** Spotify utilizes a combination of collaborative filtering, natural language processing, and audio analysis for music recommendations [Konstan and Riedl, 2012].
- **Social media:** Platforms like Facebook use hybrid systems to suggest connections and content [Zhang et al., 2019].

2.5.2 Challenges in Characterizing User Behavior

Understanding and characterizing user behavior is central for the development of effective recommender systems. However, this task presents several challenges due to the complex and varied nature of user interactions with digital platforms.

2.5.2.1 Disparate Nature of User Interactions

Users exhibit a wide range of behaviors when interacting with recommender systems, making it difficult to create a one-size-fits-all model of user preferences. This disparity manifests in several ways, each presenting unique challenges for recommender systems.

Firstly, the **variability in user preferences** poses a significant challenge. As Amatriain et al. [2009] point out, users' tastes can vary significantly, not only between individuals but also over time for a single user. This temporal aspect of preference variation adds more complexity to user modeling. For instance, a user's music preferences might shift dramatically based on mood, season, or life events, requiring recommender systems to adapt dynamically to these changes.

Secondly, **inconsistency in rating behavior** introduces considerable noise into the data. Amatriain et al. [2009] highlight that users may not be consistent in how they rate or interact with items. This inconsistency can stem from various factors, such as changes in mood, context, or even the user's understanding of the rating scale. For example, a user might rate a movie highly immediately after watching it due to recency bias, but their opinion might change upon reflection, leading to inconsistent ratings for similar movies.

Lastly, the **diversity in consumption patterns** among users presents another challenge. As noted by Kaminskas and Bridge [2016], some users may have very focused interests, while others exhibit more eclectic tastes. This spectrum of consumption behavior makes it difficult for recommender systems to strike the right balance between specificity and diversity in their recommendations. A system optimized for users with focused interests might perform poorly for those with eclectic tastes, and vice versa.

Example of Diverse User Behavior in Movie Recommendations

Consider two users of a movie recommendation system:

- User A consistently watches and highly rates action movies, showing a focused interest.
- User B watches a mix of genres (comedy, drama, sci-fi) and rates them inconsistently, sometimes giving high ratings to movies they didn't finish watching.

The model must simultaneously cater to User A's focused preferences and User B's diverse, inconsistent behavior without overfitting. User B's inconsistent ratings introduce noise that could skew the model's understanding of their preferences. Balancing these contrasting behaviors requires sophisticated algorithms adapting to individual patterns while maintaining overall system performance.

2.5.2.2 Importance of Understanding User Behavior Disparity

Recognizing and accounting for the disparate nature of user behavior is fundamental for several reasons.

Improved personalization is a key benefit of understanding individual user patterns. Kim et al. [2021] demonstrate that this understanding allows for more accurate and tailored recommendations, enhancing the user experience and increasing the likelihood of positive interactions. By recognizing the unique preferences and behavioral patterns of each user, recommender systems can provide suggestions that are more likely to resonate with the individual.

Awareness of user behavior disparity enables enhanced algorithm selection. Konstan and Riedl [2012] point out that different recommendation algorithms may perform better for different types of users. For instance, collaborative filtering might work well for users with mainstream tastes, while content-based approaches might be more effective for users with niche interests. By understanding these differences, system designers can implement adaptive approaches that select the most appropriate algorithm based on the identified user type.

Understanding user behavior disparity is crucial for fairness and bias mitigation in recommender systems. Diricic et al. [2023] argue that awareness of diverse user behaviors helps in identifying and addressing potential biases. This understanding allows developers to implement measures that ensure equitable treatment across different user groups, preventing the system from inadvertently favoring certain behavioral patterns over others.

Accommodating diverse user behaviors significantly impacts overall user satisfaction and engagement. By recognizing and adapting to various user interaction styles, recommender systems can provide a more inclusive and satisfying experience [Konstan and Riedl, 2012]. This adaptability not only improves individual user satisfaction but also contributes to increased engagement across the entire user base, ultimately enhancing the system's overall effectiveness and value.

2.5.2.3 Challenges Inherent to the Data

In addition to the complexities of user modeling, several characteristics of recommender system data pose significant challenges for user behavior characterization. These data-related challenges are equally important and often intertwined with user behavior complexities.

Data sparsity is a fundamental issue in recommender systems [Batmaz et al., 2019]. Most users interact with only a small fraction of available items, leading to sparse user-item interaction matrices. This sparsity makes it difficult to draw reliable conclusions about user preferences

and can lead to less accurate recommendations. Moreover, it complicates finding similar users or items, which is key for many collaborative filtering approaches.

The cold-start problem presents another significant challenge. As Khusro et al. [2016] explain, new users or items have little or no interaction data, making it difficult to generate accurate recommendations. This issue is particularly acute in dynamic environments where new users and items are frequently added. The cold-start problem not only affects the quality of recommendations for new entities but also impacts the overall system performance and user satisfaction.

Traditional recommender systems often rely on tabular data, primarily user-item interactions. However, Zhang et al. [2019] point out that this data format is not always well-suited to newer deep learning approaches. These limitations of tabular data can hinder the adoption of more advanced machine-learning techniques that have shown promise in other domains. Bridging the gap between traditional data formats and modern algorithmic approaches remains an ongoing challenge.

Temporal dynamics add another layer of complexity to recommender systems [Rabiu et al., 2020]. User preferences and item popularity can change over time, requiring models that can adapt to these shifts. These temporal changes can occur at various scales, from short-term fluctuations due to external events to long-term evolving trends. Developing models that can effectively capture and respond to these temporal dynamics while maintaining stable and accurate recommendations is a significant challenge.

Summary of Challenges in User Behavior Characterization

The main challenges when trying to model user behavior for recommender systems include:

- Capturing the diverse and sometimes inconsistent nature of user preferences: This involves developing models that can handle variability both across users and within individual user behavior over time.
- Addressing data sparsity and the cold-start problem: These issues require innovative approaches to infer preferences from limited data and to effectively integrate new users and items into the system.
- Incorporating temporal dynamics and contextual information: This challenge involves creating adaptive models that can capture evolving preferences and respond to changing contexts.
- Developing metrics that accurately reflect user satisfaction beyond simple accuracy measures: This requires a more holistic approach to evaluation that considers factors such as diversity, novelty, and long-term user engagement.

Techniques such as latent factor models, deep learning approaches, and hybrid methods have been developed to address these issues, but the field continues to evolve as new challenges emerge [Batmaz et al., 2019; Zhang et al., 2019].

Understanding the disparate nature of user behavior and the challenges inherent in recommender system data is decisive for developing more effective, fair, and personalized recommendation algorithms. This understanding motivates the advanced evaluation metrics and analysis techniques discussed in subsequent sections of this thesis, which aim to provide a more comprehensive framework for assessing and improving recommender system performance. By addressing these challenges, researchers and practitioners can work towards creating recommender systems that not only provide accurate suggestions but also enhance user satisfaction and engagement in diverse and dynamic environments.

2.5.3 Existing Evaluation Approaches for Recommender Systems

2.5.3.1 Recommendation Datasets

The choice of dataset is crucial in evaluating recommender systems, as it significantly impacts the assessment of algorithm performance, generalizability, and real-world applicability. Different datasets capture various aspects of user behavior, item characteristics, and interaction patterns, which can influence the effectiveness of recommendation algorithms. Moreover, the dataset's properties, such as sparsity, scale, and temporal aspects, can reveal or mask certain strengths or weaknesses of the algorithms being evaluated. Therefore, selecting appropriate datasets is essential for conducting fair comparisons, identifying domain-specific challenges, and ensuring that the developed algorithms are robust across diverse scenarios.

Several benchmark datasets have become standard in the field, each offering unique characteristics and challenges:

Examples of Recommendation Datasets

- **MovieLens:** A series of movie rating datasets of varying sizes, widely used for benchmarking [Harper and Konstan, 2016]. MovieLens datasets come in different versions (e.g., 100K, 1M, 20M, 25M), allowing researchers to test algorithms on different scales. They contain explicit ratings (1-5 stars) and timestamps, enabling both rating prediction and temporal analysis. The datasets also include movie genres and, in some versions, tag data, facilitating content-based and hybrid recommendation approaches.
- **Netflix Prize:** A large-scale movie rating dataset that spurred significant advancements in collaborative filtering [Bennett et al., 2007a]. This dataset contains over 100 million ratings from about 480,000 users on nearly 18,000 movies. It is known for its sparsity and the presence of a withheld test set, which was used for the famous Netflix Prize competition. The dataset's large scale and the associated competition have made it a benchmark for testing the scalability and accuracy of collaborative filtering algorithms.
- **Amazon product reviews:** Datasets covering various product categories, useful for e-commerce recommendation tasks [Lakkaraju et al., 2013]. These datasets include millions of reviews spanning multiple product categories like books, electronics, and clothing. They contain not only ratings but also review text, product metadata, and in some cases, browsing logs. This rich data allows for the evaluation of diverse recommendation techniques, including content-based, collaborative, and hybrid approaches in an e-commerce context.
- **Last.fm:** Music listening data, valuable for studying temporal patterns in recommendations. The Last.fm dataset contains music listening histories of users, including artist names, track titles, and timestamps. Unlike explicit rating datasets, Last.fm provides implicit feedback data (listening events), making it useful for evaluating algorithms that work with implicit user preferences. The dataset's temporal nature allows for the study of evolving music tastes and the evaluation of time-aware recommendation algorithms.

These datasets vary in size, domain, and type of user-item interactions, allowing researchers to evaluate recommender systems under different conditions. By using a combination of these datasets, researchers can assess the robustness and generalizability of their algorithms

across various domains and data characteristics. This diverse evaluation approach helps identify domain-specific challenges and ensures that the developed recommender systems are effective across various scenarios.

2.5.3.2 Data-Splitting Strategies

The way data is split into training and testing sets can significantly impact evaluation outcomes [Meng et al., 2020]. Common strategies include:

Definition of Data Splitting Strategies in Recommender Systems

- **Random splitting:** Randomly assigning interactions to train and test sets.
- **Leave-one-item:** Holding out one item per user for testing, while using the rest for training.
- **Leave-one-basket:** Holding out the last basket or session of interactions for each user for testing.
- **Global temporal splitting:** Dividing the entire dataset based on a global time threshold, using earlier interactions for training and later ones for testing.

Each strategy has its strengths and weaknesses, and the choice can significantly affect the perceived performance of recommender systems [Ji et al., 2020]. Random splitting provides a baseline evaluation but may not reflect real-world temporal dynamics. The leave-one-item approach allows for evaluating the model's ability to recommend new items to users with known preferences, but it may not reflect scenarios where users often have multiple new items to discover. Leave-one-basket is particularly useful for session-based recommendation tasks, as it mimics the process of predicting a user's next set of interactions based on their history. Global temporal splitting provides a more realistic evaluation setting, especially for systems where time plays an important role, such as news recommendation or trend-sensitive product recommendations.

Moreover, the impact of data-splitting strategies extends beyond just performance metrics. It can affect the assessment of other important aspects of recommender systems, such as their ability to handle cold-start problems, adapt to changing user preferences over time, or provide diverse recommendations. Researchers and practitioners should be aware of these implications and clearly report their chosen splitting strategy along with the rationale behind it to ensure reproducibility and fair comparison of results across different studies.

2.5.3.3 Performance Metrics

Traditional evaluation metrics in recommender systems primarily focus on accuracy, measuring how well a system can predict or rank items that a user has interacted with. These metrics are the main tool for assessing the system's ability to recover hidden user interactions [Silveira et al., 2019]. Common metrics include Recall@K, Precision@K, and NDCG@K.

Recall@K measures the proportion of relevant items found in the top-K recommendations. It is defined as:

$$\text{Recall@K} = \frac{|\text{relevant items} \cap \text{recommended items@K}|}{|\text{relevant items}|} \tag{2.5}$$

This metric is particularly useful for assessing the system's ability to identify a wide range of relevant items, but it doesn't consider their ranking within the top-K recommendations.

Precision@K, on the other hand, focuses on the proportion of relevant items among the top-K recommendations. It is calculated as:

$$\text{Precision@K} = \frac{|\text{relevant items} \cap \text{recommended items@K}|}{K} \tag{2.6}$$

Precision@K is valuable for evaluating the accuracy of the top recommendations, which directly impacts user satisfaction. However, it may not fully capture the system's ability to find all relevant items, especially when the number of relevant items exceeds K.

NDCG@K (Normalized Discounted Cumulative Gain) is a ranking metric that takes into account both the presence and the position of relevant items in the recommendation list. It is computed as:

$$\text{NDCG@K} = \frac{\text{DCG@K}}{\text{IDCG@K}} \tag{2.7}$$

where DCG@K is the Discounted Cumulative Gain:

$$\text{DCG@K} = \sum_{i=1}^{K} \frac{2^{rel_i} - 1}{\log_2(i+1)} \tag{2.8}$$

and IDCG@K is the Ideal DCG@K (the best possible DCG@K). NDCG@K provides a more comprehensive evaluation by considering both the relevance and the ranking of recommended items. It assigns higher importance to relevant items appearing earlier in the recommendation list, aligning with the typical user behavior of focusing on top recommendations.

2.5.3.4 Beyond-Accuracy Metrics

Recent research has emphasized the importance of considering factors beyond simple accuracy to provide a more comprehensive evaluation [Kuanr and Mohapatra, 2021]. These metrics aim to capture aspects of user satisfaction that go beyond mere prediction accuracy [Kaminskas and Bridge, 2016]. Beyond-accuracy metrics include diversity, novelty, serendipity, and coverage, each addressing a different facet of recommendation quality.

Diversity measures the variety of recommended items, ensuring that users receive a broad range of suggestions rather than a narrow, potentially repetitive set. This can be quantified using the Intra-List Distance (ILD):

$$\text{ILD} = \sum_{i \in L_u} \sum_{i \neq j} d(i,j) \tag{2.9}$$

where L_u is the list of recommendations for user u and $d(i,j)$ is a distance between items i and j. A higher ILD indicates greater diversity in the recommendations, which can improve user satisfaction and engagement by exposing users to a wider range of options.

Novelty assesses the system's ability to recommend new or unexpected items, helping users discover content they might not have found on their own. It is often quantified using the mean self-information of recommended items:

$$\text{Novelty} = -\frac{1}{|L_u|} \sum_{i \in L_u} \log_2 p(i) \tag{2.10}$$

where $p(i)$ is the probability of item i being recommended to a random user. Higher novelty scores indicate that the system recommends less common or more surprising items, potentially increasing user engagement and satisfaction.

Serendipity evaluates the "pleasantly surprising" nature of recommendations, capturing the system's ability to make unexpected yet relevant suggestions. It can be measured as the proportion of relevant and novel items:

$$\text{Serendipity} = \frac{|\text{relevant} \cap \text{novel} \cap \text{recommended}|}{|\text{recommended}|} \tag{2.11}$$

Serendipitous recommendations can enhance user experience by introducing users to items they might enjoy but wouldn't have discovered through their typical browsing patterns.

Coverage measures the proportion of items that the system is able to recommend, providing insights into the breadth of the recommender system's capabilities. It is often calculated as the proportion of items that the system recommends to at least one user:

$$\text{Coverage} = \frac{\left| \bigcup_{u \in U} L_u \right|}{|I|} \tag{2.12}$$

where U is the set of users and I is the set of all items. Higher coverage indicates that the system utilizes a more significant portion of the available item catalog, which can be particularly important for long-tail item discovery and overall system utility.

These beyond-accuracy metrics provide a more holistic view of a recommender system's performance, considering not just its ability to predict user preferences accurately, but also its capacity to provide diverse, novel, and serendipitous recommendations while covering a substantial portion of the item catalog. By incorporating these metrics into evaluation frameworks, researchers and practitioners can develop recommender systems that not only accurately predict user preferences but also enhance overall user experience and satisfaction.

2.5.3.5 Evaluation Frameworks

To enhance evaluation rigor and standardization, several comprehensive frameworks have been developed. These include Elliot [Anelli et al., 2021] and ReChorus2.0 [Li et al., 2024], which provide standardized tools for evaluating recommender systems across various metrics and methodologies. While these evaluation setups provide a solid foundation for assessing recommender systems, they also have limitations. While these evaluation setups provide a solid foundation for assessing recommender systems, they also have limitations. Offline evaluation biases, as highlighted by Sun [2023], can arise from mismatches between historical data and live user behavior, potentially skewing performance assessments. Replicability challenges, noted by Dong et al. [2023], stem from variations in data preprocessing, algorithm implementations, and evaluation protocols, making it difficult to reproduce results consistently across studies. Additionally, Dietz et al. [2023] emphasize the need for domain-specific evaluation approaches, as generic frameworks may not capture the nuances of particular recommendation contexts, such as Point-of-Interest recommendation. These issues underscore the complexity of recommender system evaluation and highlight the ongoing need for refined methodologies to provide reliable and meaningful performance assessments.

2.5.4 Shortcomings of Current Evaluation Methods

While existing evaluation approaches have provided valuable insights into recommender systems, several limitations can be identified.

2.5.4.1 Replicability Issues

The lack of standardization in dataset creation, preprocessing, and algorithm implementation poses significant challenges for replicability in recommender system evaluation [Sun et al., 2020]. This issue manifests in several ways, each contributing to the difficulty in reproducing results and comparing studies across the field.

The process of dataset creation and preprocessing often varies widely between studies. Researchers may employ different data cleaning techniques, handle missing values differently, or apply varying feature engineering methods. For instance, one study might remove users with fewer than five interactions, while another might set this threshold at ten. Such seemingly minor differences can lead to significant variations in the final dataset characteristics and, consequently, in the performance of recommender algorithms.

The implementation of algorithms across studies is frequently inconsistent. Even when researchers claim to use the same algorithm, subtle differences in implementation details, such as initialization methods, learning rate schedules, or convergence criteria, can lead to notably different results. This inconsistency is particularly problematic in the case of complex models like deep learning-based recommenders, where numerous hyperparameters and architectural choices can significantly impact performance.

These replicability issues collectively create a significant barrier to progress in the field of recommender systems. As Dong et al. [2023] point out, the inability to consistently reproduce findings across studies makes it difficult to establish reliable benchmarks and to truly understand the state-of-the-art in recommendation performance. Moreover, this lack of replicability can lead to the perpetuation of suboptimal practices or the overlooking of important factors that contribute to recommendation quality.

Replicability-Related Shortcomings in Recommender System Evaluation

Current recommender system evaluations often suffer from:

- Non-standardized processes for dataset creation and preprocessing
- Inconsistent implementation of algorithms across studies
- Lack of detailed reporting on experimental setups

These issues may lead to difficulties in reproducing results and comparing studies.

Addressing these replicability challenges requires a concerted effort from the research community to establish and adhere to standardized practices in dataset preparation, algorithm implementation, and experimental reporting. Some initiatives, such as the development of standardized evaluation frameworks and the promotion of open-source implementations, are steps in the right direction. However, more work is needed to ensure that recommender system research can build upon a foundation of reproducible and comparable results.

2.5.4.2 Limited Cross-Domain Applicability

The evaluation of recommender systems often suffers from limited cross-domain applicability, which can lead to an incomplete understanding of system performance across diverse scenarios. This limitation manifests in several key ways, each contributing to potential gaps in our knowledge of recommender system capabilities and effectiveness.

Many studies in recommender systems research focus on a single domain, such as movie recommendations or e-commerce [Dietz et al., 2023]. While this approach allows for in-depth analysis within a specific context, it significantly limits our understanding of how these systems

perform across different domains. For instance, a recommendation algorithm that performs exceptionally well for movie suggestions might struggle in a music recommendation context due to differences in user behavior, item characteristics, or interaction patterns. This narrow focus can lead to overly optimistic assessments of a system's general applicability or effectiveness. The lack of cross-domain studies also makes it challenging to assess the generalizability of both models and evaluation metrics. Recommender systems that demonstrate high performance in one domain may not necessarily translate well to others. Similarly, evaluation metrics that effectively capture user satisfaction in one context might be less relevant or even misleading in another. Without comprehensive cross-domain evaluations, it becomes difficult to identify truly robust and versatile recommendation approaches.

Furthermore, domain-specific challenges and performance criteria are often overlooked in generalized evaluation approaches. As Fan et al. [2024] point out, this oversight can lead to either overly optimistic or pessimistic assessments of system performance. For example, in a news recommendation system, the freshness of articles might be a very important factor, while in a movie recommendation system, diversity of genres could be more important. Failing to account for these domain-specific factors can result in evaluations that don't accurately reflect the real-world utility of the recommender system. Without a clear understanding of how systems perform across different domains, researchers and practitioners may struggle to design algorithms that can effectively operate in multiple contexts or easily adapt to new domains.

Cross-Domain Shortcomings in Recommender System Evaluation

Current evaluation practices often lack:

- Studies spanning multiple domains or application areas
- Evaluation of model transferability across different contexts
- Consideration of domain-specific challenges and metrics

These limitations may lead to an incomplete understanding of recommender system performance across diverse scenarios.

Addressing these limitations requires a more holistic approach to recommender system evaluation. This could involve conducting systematic cross-domain studies, developing evaluation frameworks that can flexibly incorporate domain-specific metrics, and exploring the transferability of models and metrics across different application areas.

2.5.4.3 Need for User Modeling

Current evaluation methods for recommender systems often fall short in adequately modeling user behavior and preferences, leading to potential misrepresentations of real-world performance. This limitation stems from an overreliance on aggregate metrics that fail to capture the nuances of individual user experiences.

Many evaluation approaches focus on broad performance measures without considering the intricacies of individual user behavior [Kleinberg et al., 2022]. This approach can overlook crucial factors that influence user interactions with recommender systems, such as contextual information, current mood, or evolving interests. For instance, a user's preference for movie recommendations might vary significantly depending on whether they're watching alone or with family, or whether it's a weekday evening or a weekend afternoon. By failing to account for these situational factors, evaluations may present an incomplete picture of a system's effectiveness.

Another significant oversight in many current evaluation approaches is the lack of consideration for long-term user satisfaction and engagement metrics. As Sun [2024] emphasize,

understanding how recommender systems impact user behavior and satisfaction over extended periods is key for assessing their true value. Short-term metrics might indicate high performance, but they may not reflect whether the system is fostering user loyalty, encouraging exploration of diverse content, or contributing to overall platform engagement in the long run.

User Modeling Shortcomings in Recommender System Evaluation

Current evaluation approaches often neglect:

- Comprehensive modeling of user behavior and preferences
- Consideration of user context and situational factors
- Evaluation of long-term user satisfaction and engagement

These omissions may result in evaluations that do not accurately reflect real-world user experiences.

Addressing these shortcomings requires a shift towards more comprehensive user modeling in recommender system evaluations. This could involve incorporating contextual factors into evaluation frameworks, developing metrics that capture the evolution of user preferences over time, and implementing long-term studies to assess sustained user engagement and satisfaction.

2.5.4.4 Need for Statistical Tests

Many studies report improvements in recommender system performance without conducting rigorous statistical validation [Ji et al., 2020]. This lack of statistical testing makes it difficult to determine whether observed differences in performance are truly meaningful or due to chance. Additionally, the focus on statistical significance without considering effect sizes can lead to overemphasis on small, practically insignificant improvements [Shevchenko et al., 2024]. The problem is further compounded in large-scale evaluations where multiple comparisons are made without proper statistical corrections.

To address these shortcomings, recommender system evaluation would benefit from:

- Developing standardized protocols for dataset creation, preprocessing, and algorithm implementation [Anelli et al., 2021]
- Conducting more cross-domain studies to assess model generalizability [Dietz et al., 2023]
- Incorporating comprehensive user modeling in evaluation frameworks [Kleinberg et al., 2022]
- Implementing robust statistical analyses, including significance tests and effect size measurements [Shevchenko et al., 2024]
- Standardizing reporting practices to facilitate comparisons across studies and domains [Li et al., 2024]

By addressing these limitations, researchers can develop more comprehensive and reliable approaches to evaluating recommender systems, leading to more accurate assessments of performance and better understanding of system behavior across diverse scenarios and user groups [Sun, 2024].

2.6 Conclusion

This Chapter has provided a comprehensive overview of the current state of evaluation practices in deep learning, focusing on three key domains: Class Incremental Learning, Face Recognition, and Recommender Systems. Through our examination, we have highlighted several common themes and challenges that persist across these areas.

First, we have shown that traditional evaluation metrics, while useful, often fall short in capturing the full complexity of deep learning model performance [Doshi-Velez and Kim, 2017]. In Class Incremental Learning, standard accuracy measures may fail to adequately reflect a model's ability to retain knowledge of previously learned classes [Masana et al., 2021]. Similarly, in Face Recognition, conventional metrics might not account for biases across different demographic groups [Grother et al., 2019]. For Recommender Systems, beyond-accuracy metrics have emerged as crucial complements to traditional accuracy-based evaluations [Kaminskas and Bridge, 2016].

Second, we have observed a growing awareness of the need for more diverse and representative datasets in evaluation [Torralba and Efros, 2011]. This is particularly evident in Face Recognition, where the importance of inclusive datasets that span various ethnicities, ages, and genders has been emphasized [Buolamwini and Gebru, 2018]. In Recommender Systems, the challenge of capturing the disparate nature of user interactions has highlighted the need for more comprehensive data collection and evaluation strategies [Harper and Konstan, 2016].

Third, the importance of rigorous statistical analysis has been a recurring theme. Across all domains, there is an increasing call for more robust testing, proper handling of multiple comparisons, and quantifying importance alongside statistical significance [Bouthillier et al., 2021]. This trend reflects a broader move towards more reliable and reproducible evaluation practices in deep learning [Pineau et al., 2021].

Fourth, we have noted the emergence of domain-specific evaluation frameworks and metrics [Anelli et al., 2021]. These tailored approaches aim to address the unique challenges posed by each application area, moving beyond one-size-fits-all evaluation strategies.

Future Research for Better Evaluation Methodologies

Moving forward, the field of evaluation presents several promising avenues for further investigation and refinement of methodologies:

1. Development of standardized evaluation protocols that facilitate fair comparisons across different models and studies [Bouthillier et al., 2021].
2. Integration of causal inference methods to better understand the factors driving model performance [Xu et al., 2021].
3. Exploration of long-term evaluation strategies, particularly relevant for Class Incremental Learning and Recommender Systems [Hayes and Kanan, 2022].
4. Investigation of multi-metric evaluation frameworks that can provide a more holistic view of model performance [Reddi et al., 2020].
5. Advancement of interpretability and explainability methods to complement quantitative evaluations [Lipton, 2017].

By addressing these challenges and pursuing these research directions, the field can move towards more rigorous, comprehensive, and insightful evaluation practices. This progress is essential not only for advancing the state-of-the-art in deep learning but also for ensuring the responsible and effective deployment of these powerful technologies in real-world applications [Doshi-Velez and Kim, 2017].

As we proceed to the subsequent chapters of this thesis, we will build upon these insights, proposing novel and solid evaluation methodologies that aim to address the limitations identified in current practices. Through this work, we aspire to contribute to the development of more robust, fair, and informative evaluation frameworks for deep learning models [Hendrycks et al., 2021].

It is easy to lie with statistics.
It is hard to tell the truth without it.

— Andrejs Dunkels

3

Statistical Tools for a Better Analysis of Machine Learning

Contents

3.1 Introduction

Building upon the challenges in model evaluation outlined in Chapters 1 and 2, we present the statistical methods and causal inference tools that form the foundation of our analytical framework. Our approach draws primarily from established econometric techniques, adapting them specifically for machine learning analysis. Unless noted otherwise, all our statistical elements are drawn from "Mostly Harmless Econometrics" by Angrist and Pischke [2009],

"Econométrie: méthodes et applications" by Crepon and Jacquemet [2010], and "Econometrics" by Wooldridge [2013].

The methods presented here move beyond traditional performance metrics to provide a rigorous statistical framework for understanding model behavior. This includes techniques for hypothesis testing, confidence interval estimation, analysis of variance (ANOVA), and causal inference methods. These tools enable us to quantify uncertainty in our evaluations, isolate the effects of individual factors, and test hypotheses about the relative importance of different model components. The application of these statistical approaches is particularly crucial when evaluating complex machine learning systems in real-world scenarios. They provide the tools needed to understand how models perform across different contexts and conditions, enabling more reliable and interpretable research outcomes.

3.2 Foundations of Econometrics: Ordinary Least Squares (OLS)

Ordinary Least Squares (OLS) is a fundamental method in econometrics and serves as the foundation for many advanced statistical techniques. It provides a straightforward yet powerful approach to modeling relationships between variables, and its understanding is fundamental to being able to draw meaningful conclusions from analytic experiments.

3.2.1 The Linear Regression Model

The linear regression model is a statistical method used to model the relationship between a dependent variable and one or more independent variables.

Definition: Endogenous and Exogenous Variables

In econometrics, variables are classified as:

- **Endogenous variable**: The dependent variable (Y) that the model aims to explain or predict. It is determined within the model and is influenced by the exogenous variables.
- **Exogenous variables**: The independent variables (X) that are used to explain or predict the endogenous variable. They are determined outside the model and are assumed to influence the endogenous variable.

Let Y and X be random variables modeling an endogenous and exogenous variable. Then, the simple linear model is:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{3.1}$$

Where:

- $\beta_0 \in \mathbb{R}$ is the called ***intercept term***
- $\beta_1 \in \mathbb{R}$ is the ***slope coefficient***
- ε is the ***error*** random variable

In practice, however, we often want to consider multiple explanatory factors, leading to the full linear model:

Definition: The Linear Regression Model

Let Y be an endogenous variable and $X_1, ..., X_K$ be exogenous variable. The full Linear Regression Model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_K X_K + \varepsilon \tag{3.2}$$

where $\beta_1, \beta_2, ..., \beta_k$ are their regression coefficients and ε is models the error

The model in equation 3.2 should be familiar to anyone with a Machine Learning background, since linear regression is the building block of many ML algorithms, including neural networks. However, whereas in applied machine learning, using this model as a predictive tool is common, in econometrics, the focus is put on finding the right assumptions that allow us to interpret the coefficients β_k as causal effects and draw explanatory conclusions.

3.2.2 Interpreting the Coefficients

Several key assumptions must hold for the β_k coefficients in the linear regression model to be interpreted as causal effects. These assumptions are fundamental to the Gauss-Markov theorem and are often referred to as the Classical Linear Regression Model (CLRM) assumptions.

Classical Linear Regression Model (CLRM) assumptions

- **Correct Specification**: the regression equation contains all of the relevant predictors, including any necessary transformations. That is, the model has no missing, redundant, or extraneous predictors.
- **Linearity in Parameters**: The relationship between Y and X_k is linear in the parameters β_k. This doesn't mean X_k and Y must have a linear relationship, but rather that the parameters enter the equation linearly.
- **Exogeneity**: $\mathbb{E}[\varepsilon|X_k] = 0]$ for all k. In other words, there are no omitted variables that are correlated with both the dependent variable and the independent variables. This is needed for causal interpretation as it ensures that unobserved factors do not bias the estimated effects.
- **Normality of the Error Term**: $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. While not strictly necessary for unbiasedness, the assumption that the error term is normally distributed allows for valid inference (hypothesis testing and confidence intervals).
- **Homoscedasticity**: $\mathbb{V}(\varepsilon|X_k) = \sigma^2$. This assumption states that the variance of the error term is constant across all levels of the independent variables. Homoscedasticity ensures that the precision of the β_k estimates is consistent across the range of X_k values.
- **No Perfect Multicollinearity**: There should be no exact linear relationships among the independent variables. This ensures that we can uniquely estimate the effect of each variable.

When all the criteria are met, we can interpret the coefficients as causal effects:

Linear Regression Model Interpretation

When all CLRM assumptions are met, then each coefficient β_k represents the **marginal effect** of the corresponding exogenous variable X_k on the endogenous variable Y, holding all other variables constant. Mathematically, this can be expressed as:

$$\beta_k = \frac{\partial Y}{\partial X_k} \tag{3.3}$$

The "holding all other variables constant" condition is known as the *ceteris paribus* assumption. It's essential for isolating the effect of a single variable and is a key concept in causal interpretation [Angrist and Pischke, 2009].

Regression over many variables X_k simultaneously allows us to disentangle the effect of each variable and, under the right assumptions, draw meaningful conclusions on the relationship between X_k and Y. In particular, under the exogeneity and correct specification assumptions, the β_k coefficient can be interpreted as the **causal effect** of X_k on Y.

3.2.3 Fitting the Model

From now, let us denote by X the row vector $(1, X_1, ..., X_K)$, and β the vector of all β_k. Then, under the CLRM assumptions, we have a closed formula for the coefficients:

Formula for the Coefficients

Under the exogeneity and no perfect colinearity assumptions, the expression

$$\beta = \mathbb{E}\left[X^T X\right]^{-1} \mathbb{E}\left[X^T Y\right] \tag{3.4}$$

is a solution to the equation 3.2.

Proof. We have :

$$\begin{aligned} Y &= X\beta + \varepsilon \\ X^T Y &= X^T X\beta + X^T \varepsilon \\ \mathbb{E}[X^T Y] &= \mathbb{E}[X^T X]\beta + \mathbb{E}[X^T \varepsilon] \\ \mathbb{E}[X^T Y] &= \mathbb{E}[X^T X]\beta \quad \text{by exogeneity} \\ \beta &= \mathbb{E}[X^T X]^{-1}\mathbb{E}[X^T Y] \end{aligned}$$

The inversion of $\mathbb{E}[X^T X]^{-1}$ is possible under the no perfect colinearity assumption. □

Of course, in practice, we only have access to realizations of (X, Y). We thus define the empirical counterparts of our quantities, by considering a set of n realizations $(x_i, y_i) \sim (X, Y)$, and the equation:

$$y_i = x_i\beta + \epsilon_i \tag{3.5}$$

where ϵ_i is called the ***residual*** the model for the observation i. Under the hypothesis that the (x_i, y_i) are i.i.d., we should have $\epsilon_i \overset{i.i.d.}{\sim} \varepsilon$, in other terms, that the empirical residuals should be realizations of the noise variable. We can then define our estimator:

The OLS Estimator

Under the CLRM assumptions

$$\widehat{\beta}^{\text{OLS}} := \left(\frac{1}{n}\sum_{i=1}^{n} x_i^T x_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^T y_i\right) \tag{3.6}$$

is an estimator of β, and is called the **Ordinary Least Squares Estimator (OLS)**.

Proof. Direct by the continuous mapping theorem. □

Through convex optimization, it it easy to show that $X\beta$ is the best linear approximation of $\mathbb{E}[Y|X]$ with respect to the square norm L^2, i.e. $\beta = \arg\min_b \mathbb{E}[(\mathbb{E}[Y|X] - Xb)^2]$, which justifies the name of the OLS estimator, which minimizes the empirical mean square error.

Let us denote by $\boldsymbol{X}$ the matrix of the observations x_i, and $\boldsymbol{y}$ the vector of observations of y_i. Then, the OLS estimator can be easily rewritten as

$$\widehat{\beta}^{\text{OLS}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{3.7}$$

and can be viewed as the pseudo-inverse of $\boldsymbol{X}$ acting on $\boldsymbol{y}$. Plugging back $\widehat{\beta}^{\text{OLS}}$ in the equation 3.2, we get the predicted values vector $\hat{\boldsymbol{y}}$ defined by :

$$\hat{\boldsymbol{y}} := \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \overset{not}{=} \hat{\boldsymbol{H}}\boldsymbol{y} \tag{3.8}$$

where $\hat{\boldsymbol{H}}$ is the called the **hat matrix**, and is the projection matrix on the linear space generated by the rows of $\boldsymbol{X}$. Informally, the predicted values are thus can thus be seen as an orthogonal projection of $\boldsymbol{y}$ on $\boldsymbol{X}$.

3.2.4 Statistical Significance and Hypothesis Testing

3.2.4.1 Individual Variables

The OLS estimator thereby denoted simply $\hat{\beta}$, estimates the theoretical β with a certain precision since it is a function of the realizations (x_i, y_i). We can know its precision, both in a finite sample setup and in an asymptotic setup:

Properties of the OLS estimator

(Finite sample) Under the CLRM assumptions and i.i.d sampling,

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right) \tag{3.9}$$

(Asymptotic distribution) Under the CLRM assumptions, i.i.d. sampling, and CLT moment conditions, $\hat{\beta}$ is asymptotically normal:

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2\mathbb{E}[X^TX]) \tag{3.10}$$

Proof. (Finite sample): We can write $\hat{\beta}$ as :

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^T x_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^T(x_i\beta + \epsilon_i)\right)$$

which simplifies to

$$\hat{\beta} = \beta + \left(\sum_{i=1}^{n} x_i^T x_i\right)^{-1} \left(\sum_{i=1}^{n} x_i^T \epsilon_i\right)$$

Since $\epsilon_i \overset{i.i.d.}{\sim} \varepsilon \sim \mathcal{N}(0, \sigma^2)$, $\hat{\beta}$ is normal, with tractable expected value and covariance matrix. (Asymptotic distribution): Direct by applying the CLT and calculating the variance of $\hat{\beta}$. Homoscedascity reduces the expression since $\mathbb{E}[\varepsilon^2 X^T X] = \sigma^2 \mathbb{E}[X^T X]$. □

This formulation of the variance supposes homoscedasticity of the errors, but other formulations exist for cases where this assumption is not valid.

In practice, we have to estimate $\sigma^2 \mathbb{E}[X^T X]$ from the data, leading to the empirical counterpart $s^2 \hat{\mathbb{E}}[X^T X]$, where s^2 is the unbiased variance estimator. However, the estimation of the variance with an empirical counterpart changes the law which describes our error from a Normal distribution to a Student distribution:

Definition: t-statistic

Let $s^2 = \frac{\boldsymbol{\epsilon}^T \boldsymbol{\epsilon}}{n-K}$ be the unbiased estimator of σ^2, where $\boldsymbol{\epsilon}$ is the vector of the residuals. Let $se(\hat{\beta}_k) = \sqrt{s^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}_{kk}}$. Then:

$$t_k := \frac{\hat{\beta}_k - \beta}{se(\hat{\beta}_k)} \sim \mathcal{T}(n-K) \tag{3.11}$$

where $\mathcal{T}(n-K)$ is the Student distribution with $n-K$ degrees of freedom. The quantity t_k is called the t-statistic of the coefficient β_k.

Proof. s^2 is the classical unbiased variance estimator. In particular, since the ϵ_i follow normal distributions, then we can show that $\frac{n-K}{\sigma^2} s^2 \sim \chi^2_{n-K}$ via Cochran's theorem. Rewriting the expression:

$$t_k = \frac{\hat{\beta}_k - \beta}{\sqrt{\sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}_{kk}}} \cdot \sqrt{\frac{\sigma^2}{s^2}}$$

The first factor follows a Gaussian distribution, and the second factor can be rewritten as $\frac{1}{\sqrt{U/(n-K)}}$ where U follows a χ^2 distribution with $n-K$ degrees of freedom. By definition, t_k follows a Student distribution with $n-K$ degrees of freedom. □

The t-statstic is the object that allows us to get confidence intervals for β_k, by looking at the likelihood of t_k with respect to its theoretical Student distribution In particular, this is useful to verify if $\beta_k \neq 0$, i.e. if there is a significant effect of X_k on Y, and if yes, its sign and strength.

Confidence Intervals

Let $t_{n-K,1-\alpha/2}$ be the $1-\alpha/2$-th quantile for the Student distribution with $n-K$ degrees of freedom. Then, with probability $1-\alpha$,

$$\beta_k \in \left[\hat{\beta}_k \pm t_{n-K,1-\alpha/2} \times se(\hat{\beta}_k)\right] \tag{3.12}$$

A common threshold for α is 5%.

If 0 is contained inside the confidence interval, it means we do not have enough data points to draw conclusions. This does not mean that X_k has *no* effect on Y, but that the noise in the observed data is too strong to even say if the effect is positive or negative. On the other hand, we can have a very narrow confidence interval, but around a small $\hat{\beta}$ value, corresponding to a statistically significant but small effect. Therefore, assessing both the **value** and **significance** of the effects β_k is important.

3.2.4.2 Global Significance

The main measure of the overall fit of a linear regression model is the coefficient of determination, commonly known as R^2. This statistic quantifies the proportion of variance in the dependent variable that is predictable from the independent variable(s). R^2 ranges from 0 to 1, where 0 indicates that the model explains none of the variability of the data around its mean, and 1 indicates perfect prediction. Mathematically, R^2 can be defined in terms of variance and covariance:

Definition: R^2

$$R^2 = \frac{\widehat{Cov}(y, \hat{y})^2}{\widehat{Var}(y)\widehat{Var}(\hat{y})} = \frac{\widehat{Var}(\hat{y})}{\widehat{Var}(y)} = 1 - \frac{\widehat{Var}(\epsilon)}{\widehat{Var}(y)} \tag{3.13}$$

Formulation 3.13 highlights that R^2 represents the squared correlation between the observed and predicted values, or equivalently, the ratio of the variance of the predicted values to the variance of the observed values. While R^2 provides a useful measure of model fit, it should be interpreted cautiously, especially when comparing models with different numbers of predictors or when working with small sample sizes. In such cases, the adjusted R^2 or other information criteria like AIC may provide more reliable measures of model quality, and will be described following sections.

3.2.5 Performing Regression in Practice

3.2.5.1 Notation

When performing any OLS regression of a variable Y on the variables $X_1, ..., X_K$, we will denote the regression using the R-style formula:

R-style Notation

The Linear Regression Model equation 3.2 is denoted by:

$$\mathtt{Y} \sim \mathtt{X}_1 + ... + \mathtt{X}_K \tag{3.14}$$

This notation omits the intercept, the coefficients, and the residuals, to clarify what are endogenous ($\mathtt{Y}$) and exogenous ($\mathtt{X}_k$) variables.

3.2.5.2 Categorical Variables

In some cases, a variable X can be categorical instead of being continuous. In this case, if X take C discrete unordered values $D_1, ..., D_C$, we employ a ***dummy coding***:

1. Take a value, for example D_1, as a reference value,
2. Encode X as a one-hot encoding vector in $\{0, 1\}^{C-1}$, with a 1 in the position j corresponding to the observation of the value D_j,

3. Get $C-1$ coefficients $\beta_2, ..., \beta_C$, corresponding to the **marginal effect of X taking the value D_j instead of the reference value** D_1, i.e.

$$\beta_k = \mathbb{1}(Y|X = D_k) - \mathbb{1}(Y|X = D_1) \tag{3.15}$$

3.2.5.3 Product Variables

In a regression of the form $\texttt{Y} \sim \texttt{X}_1 + \texttt{X}_2$, it is possible to model the case where the value of one variable can impact the marginal effect of the other one. For example, we might want to consider that the effect of training time on model performance depends on the model's complexity. This can be achieved by including an interaction term:

Definition: Product variables

To model interaction between variables, the equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2) + \varepsilon \tag{3.16}$$

is denoted equivalently by the formulas

$$\texttt{Y} \sim \texttt{X}_1 + \texttt{X}_2 + \texttt{X}_1 : \texttt{X}_2 \iff \texttt{Y} \sim \texttt{X}_1 \times \texttt{X}_2 \tag{3.17}$$

where $\texttt{X}_1 : \texttt{X}_2$ represents the interaction between $\texttt{X}_1$ and $\texttt{X}_2$.

To get sound interpretations, it is best to first center the exogenous variables, so $X_1 \cdot X_2$ is less correlated with X_1 and X_2. In this case:

- β_1 represents the effect of X_1 when X_2 is at its mean
- β_2 represents the effect of X_2 when X_1 is at its mean
- β_3 represents how the effect of X_1 changes for each unit increase in X_2 (and vice versa)

This allows us to model quite complicated dependencies between explanatory variables and the target variable.

Plain OLS regressions can be performed and interpreted as long as Y is a continuous variable and that the CLRM hypothesis holds. The difficulty in a rigorous evaluation protocol does not come from complicated evaluation models but from a rigorous choice of explanatory variables, hypothesis verification, and careful interpretation of the results.

3.3 Analysis of Variance (ANOVA)

OLS regressions allow us to get marginal effects associated with each variable, which can be quite informative when the explanatory variables are continuous. For categorical variables, while dummy coding allows us to still get marginal effects from one group to another, it does not allow us to easily know what categorical variable is the most important, i.e. for which categorical variable the variance in the category is the greatest. For example, if we evaluate models by varying both model types and training types (both categorical variables), we need a way to say that one aspect has a greater influence on performance than the other.

3.3.1 ANOVA as an Extension of OLS

Analysis of Variance (ANOVA) is a powerful statistical technique used to analyze the differences among group means in a sample. While originally developed for experimental design, ANOVA has found wide applications in various fields, including machine learning evaluation. ANOVA can be viewed as a special case of the Ordinary Least Squares (OLS) regression we discussed earlier. In fact, ANOVA and linear regression are two faces of the same coin, both falling under the General Linear Model framework.

Consider a model with two categorical variables X_1 and X_2:

$$Y = \beta_0 + \sum_i \beta_{1i} D_{1i} + \sum_j \beta_{2j} D_{2j} + \varepsilon \tag{3.18}$$

Where D_{1i} and D_{2j} are dummy variables for X_1 and X_2, respectively. This OLS formulation is equivalent to the ANOVA model:

$$Y_{ij} = \mu + A_i + B_j + \varepsilon \tag{3.19}$$

where Y_{ij} is the response variable when $X_1 = D_{1i}$ and $X_2 = D_{2j}$, μ is the overall mean effect, A_i is the effect of D_{1i}, and B_j is the effect of D_{2j}. We have $\mu + \alpha_1 + \beta_1 = \beta_0$ (for the reference categories), $A_i = \beta_{1j}$ for $i = 2, ..., J$

The ANOVA formulation can be advantageous since it naturally partitions the total variance into components associated with each factor and residual error. It allows for easier interpretation of main effects in the presence of multiple categorical variables, and provides a framework for analyzing complex experimental designs, including nested and crossed factors.

3.3.2 Effect Sizes: Decomposing Model Variance

While OLS focuses on estimating coefficients, ANOVA emphasizes decomposing the total variance in Y. This decomposition allows us to calculate effect sizes, particularly partial eta-squared η_p^2, which quantifies the proportion of variance explained by each factor.

Definition: Partial η^2

For a regression of the observed y_i on the exogenous variables $x_{1i}, ..., x_{Ki}$, let $\hat{y}_i$ be the predicted outcome of the full model for the observation i, and $\hat{y}_{ki}$ the predicted outcome based only on x_{ki}. Then define the partial and error sum of squares by:

$$SS_{effect} := \sum_i (\hat{y}_{ki} - \hat{y}_i)^2$$
$$SS_{error} := \sum_i (y_i - \hat{y}_i)^2$$

Then the partial η^2, also denoted for X_k is defined by:

$$\eta_p^2(X_k) = \frac{SS_{effect}}{SS_{effect} + SS_{error}} \tag{3.20}$$

Interpretation of η_p^2

η_p^2 represents the proportion of variance in Y explained by a factor, after accounting for other factors in the model. It's directly related to the increase in R^2 when adding the factor to a model that already includes other factors.

The ANOVA framework is particularly useful in machine learning evaluation when:

- Comparing the effects of multiple categorical factors (e.g., model architecture, dataset type)
- Quantifying the relative importance of different factors using η_p^2
- Analyzing interaction effects between factors (in multi-way ANOVA)

While OLS provides coefficient estimates, ANOVA's focus on variance decomposition often provides a more intuitive understanding of each factor's importance in explaining variability in model performance.

3.4 Binary Outcome Models: The Logit Model

The general linear model is not adapted in the case where the endogenous variable Y is binary. If $Y \in \{0, 1\}$, then $\mathbb{E}[Y|X] \in [0, 1]$. In the linear model, we assume a solution of the form $\mathbb{E}[Y|X] = X\beta$, but nothing guarantees that $X\beta$ is in $[0, 1]$, which raises serious specification questions.

Additionally, the usual interpretation of β_k as a marginal effect becomes less clear. If Y is binary, interpreting a β_k value of, for example, 0.6, is not easy. This is because the marginal effect on a binary outcome cannot be directly interpreted in the same way as it would for a continuous outcome, since the effect on the probability is nonlinear.

3.4.1 Definition of the Logit Model

A solution is to restrict the predictions to a function F that maps the linear predictor to the $[0, 1]$ interval, ensuring that the predicted values are valid probabilities. One such function is the logistic function, commonly used in the Logit model.

Definition: Generalized Linear Model (GLM)

If F is a strictly increasing bijective function from $\mathbb{R}$ to $]0, 1[$, it is a cumulative distribution function, and a **Generalized Linear Model** is specified as:

$$\mathbb{E}[Y|X] = F(X\beta) \tag{3.21}$$

In particular, this allows us to model non-linear situations. In the binary outcome case, we can consider that there exists a latent variable Y^*, such that $Y^* = X\beta + \varepsilon$, where $-\varepsilon$ has a c.d.f F, and $Y = \mathbb{1}(Y^* \geq 0)$. Then

$$\mathbb{E}[Y|X] = \mathbb{P}(Y = 1|X) = \mathbb{P}(Y = 1|X) = \mathbb{P}(-\varepsilon \leq X\beta|X) = F(X\beta) \tag{3.22}$$

A common choice of function F is the sigmoid function, which gives rise to the Logit model:

Definition: The Logit Model

Let σ be the sigmoid function, i.e.

$$\forall x \in \mathbb{R}, \sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.23}$$

Then the logit model is defined by :

$$\mathbb{E}[Y|X] = \sigma(X\beta) \tag{3.24}$$

In R-style notation, it will be denoted:

$$\texttt{Y} \sim \sigma(\texttt{X}_1 + ... + \texttt{X}_K) \tag{3.25}$$

3.4.2 Fitting the Model

Sadly, there is no closed-form solution for the β coefficients in the equation 3.24 as in the OLS formulation. Thus, in practice, the model is fitted using the Maximum Likelyhood Estimation (MLE). The likelihood function for n independent observations is:

$$L(\beta) = \prod_{i=1}^{n} [\sigma(x_i\beta)]^{y_i} [1 - \sigma(x_i\beta)]^{1-y_i} \tag{3.26}$$

We maximize the log-likelihood:

$$\ell(\beta) = \sum_{i=1}^{n} y_i \log(\sigma(x_i\beta)) + (1 - y_i) \log(1 - \sigma(x_i\beta)) \tag{3.27}$$

This is typically done using numerical optimization methods like Newton-Raphson or Fisher scoring.

To obtain confidence intervals for β, we use the fact that the MLE estimator $\hat{\beta}$ is asymptotically normal. The variance-covariance matrix of $\hat{\beta}$ can be estimated using the inverse of the observed Fisher information matrix:

$$\widehat{Var}(\hat{\beta}) = I(\hat{\beta})^{-1} = \left(-\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^T}\Big|_{\beta=\hat{\beta}} \right)^{-1} \tag{3.28}$$

This allows us to construct confidence intervals:

Confidence Intervals for the Logit Model

For a given coefficient β_k, let $z_{1-\alpha/2}$ be the $1-\alpha/2$-th quantile for the Gaussian distribution. Then, with probability $1 - \alpha$:

$$\beta_k \in \left[\widehat{\beta}_k \pm z_{1-\alpha/2} \sqrt{\frac{\widehat{Var}(\hat{\beta})_{k,k}}{n}} \right] \tag{3.29}$$

where $\widehat{Var}(\hat{\beta})_{k,k}$ is the k-th diagonal element of $\widehat{Var}(\hat{\beta})$.

Proof. The MLE estimator $\hat{\beta}$ is asymptotically normal with distribution $\mathcal{N}(\beta, \frac{1}{n} I(\beta)^{-1})$ where $I(\beta)$ is the Fisher information matrix evaluated at β. Evaluating the Fisher information matrix at the empirical counterpart $\hat{\beta}$ yields the desired confidence interval. □

3.4.3 Interpreting the Model

3.4.3.1 Marginal Effects

Reversing the equation 3.24, we can the equivalent model:

$$\ln \frac{\mathbb{P}(Y = 1|X)}{\mathbb{P}(Y = 0|X)} = X^T\beta \tag{3.30}$$

Therefore, the logit regression can be viewed as a linear regression in the log-odds ratio scale. This makes it quite difficult to directly interpret the values β_k as concrete effects, since we would want to have an interpretation on the probability $\mathbb{P}(Y|X)$. For the logit model, we get :

$$\frac{\partial \mathbb{P}(Y=1|X)}{\partial X_k} = \sigma(X\beta)(1-\sigma(X\beta))\beta_k \tag{3.31}$$

Contrary to the OLS, the marginal effect of X_k on the probability of outcome Y depends on all the other variables X_{-k}. A common quantity to consider is then the Average Marginal Effect:

Definition: Average Marginal Effect

In the logit model, the **Average Marginal Effect (AME)** of a variable X_k on the probability of outcome Y is:

$$AME(X_k) = \mathbb{E}\left[\frac{\partial \mathbb{P}(Y=1|X)}{\partial X_k}\right] \tag{3.32}$$

It represents the effect of a unit change in X_k, on average, on the probability of outcome Y. Its empirical counterpart is estimated by:

$$\widehat{AME}(X_k) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \widehat{\mathbb{P}}(Y=1|X)}{\partial X_k}\bigg|_{X=x_i} = \frac{1}{n}\sum_{i=1}^{n} \sigma(x_i\hat{\beta})(1-\sigma(x_i\hat{\beta}))\hat{\beta}_k \tag{3.33}$$

The confidence intervals for $AME(X_k)$ are computed using the delta method. If $g_k(\beta)$ is the function that computes the marginal effector X_k, then:

$$\sqrt{n}(g_k(\hat{\beta}) - g_k(\beta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \nabla g_k(\beta)^T I(\beta)^{-1} \nabla g_k(\beta)) \tag{3.34}$$

The variance of the marginal effect estimate is then approximated as:

$$\widehat{Var}(g_k(\hat{\beta})) = \nabla g_k(\hat{\beta})^T \hat{I}(\hat{\beta})^{-1} \nabla g_k(\hat{\beta}) \tag{3.35}$$

This allows us to construct confidence intervals for the marginal effects:

Confidence Intervals of the AME

Let $z_{1-\alpha/2}$ be the $1-\alpha/2$-th quantile for the Gaussian distribution. Then, with probability $1-\alpha$:

$$AME(X_k) \in \left[\widehat{AME}(X_k) \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{Var}(g_k(\hat{\beta}))}{n}}\right] \tag{3.36}$$

3.4.3.2 Goodness of Fit

In logistic regression, the standard R^2 used in linear regression is not applicable due to the non-linear nature of the model. McFadden's R^2, also known as the likelihood ratio index, is one of several pseudo-R^2 measures developed for logistic regression and other models estimated by maximum likelihood.

Definition: McFadden's R^2

For a logit model, McFadden's R^2 is defined as:

$$R^2_{\text{McFadden}} = 1 - \frac{\ln(L_{\text{full}})}{\ln(L_{\text{null}})} \tag{3.37}$$

Where:

- L_{full} is the likelihood of the full model
- L_{null} is the likelihood of the null model (intercept-only model)

Unlike the R^2 in linear regression, McFadden's R^2 does not have an upper bound of 1, and values between 0.2 and 0.4 are considered to represent an excellent fit. However, like other pseudo-R^2 measures, McFadden's R^2 should be used in conjunction with other diagnostic tools and substantive interpretation of the model coefficients. It's particularly useful when the goal is to understand the factors influencing binary outcomes in machine learning models, rather than purely predictive tasks.

3.5 Model Selection and Validation

In the context of evaluation and explainability using regressions, selecting the most appropriate statistical model is required for drawing valid conclusions. This section discusses key tools for model selection and validation, focusing on information criteria and residual analysis.

3.5.1 Information Criteria

While the R^2 of a regression gives us an idea of the model explainability power, it suffers from problems, particularly overfitting. To select the best model, the **Akaike Information Criterion (AIC)** is a widely used tool for model selection, balancing model fit against complexity to avoid overfitting.

Definition: Akaike Information Criterion (AIC)

For a model with K parameters and maximum likelihood L, the AIC is defined as:

$$AIC = 2K - 2\ln(L) \tag{3.38}$$

Where:

- K is the number of estimated parameters in the model
- L is the maximum value of the likelihood function for the model

The AIC is founded on information theory and provides a relative estimate of the information lost when a given model is used to represent the process that generates the data. When comparing models, the one with the lower AIC is generally preferred. The AIC has some interesting properties:

- **Trade-off**: AIC rewards goodness of fit (as assessed by the likelihood function) but includes a penalty that increases with the number of estimated parameters. This penalty discourages overfitting.
- **Relative measure**: AIC values are only meaningful when compared between models. The absolute value of AIC for a single model is not interpretable on its own.

- **Model comparison**: When comparing models, a difference in AIC of 2 is often considered the threshold for meaningful difference.

3.5.2 Residual Analysis

To ensure the validity of our statistical inferences, we need to verify the regression assumptions. Since the theoretical model cannot be accessed, the key OLS assumptions are checked through various residual plots and statistical tests, called **regression diagnostics**. Recall that the residuals are defined by:

$$\epsilon_i = y_i - \hat{y}_i \tag{3.39}$$

The key assumptions are verified using different methods:

- **Linearity**: Plot the ϵ_i vs. the fitted values $\hat{y}_i$, and verify it is centered around 0. The idea is that to check if y_i is linear in x_i and that all the linear effects have been accounted for, we can check if $\mathbb{E}[\epsilon_i|x_i] = 0$. In particular, any pattern in this plot, such as a U-shaped curve, indicates missing non-linear effects
- **Normality** of residuals: Plot the Q-Q plot of the standardized ϵ_i against the Gaussian distribution. Tests like the Kolmogorov-Smirnov (KS) can be used to check if the residuals are indeed Gaussian, but should be used with caution: for a sample large enough, a small deviation from the Gaussian distribution, even if not too problematic, can be detected by the KS test. A graphical inspection can be more insightful in these cases.
- **No multicollinearity** among predictors: Compute each predictor's Variance Inflation Factors (VIF), which indicates how much a predictor is linear in the others. A VIF value greater than 5 or 10 indicates problematic multicollinearity.
- **Homoscedasticity** (constant variance of residuals): plot ϵ_i vs $\hat{y}_i$ or $\sqrt{\tilde{\epsilon}_i}$ vs $\hat{y}_i$, where $\tilde{\epsilon}_i$ are the re-scaled residuals. The distribution should be the same for all values of $\hat{y}_i$. If the homoscedasticity assumption does not hold, the value of the β_k coefficients is still valid, but the confidence intervals become imprecise. This can be solved by using more robust and more conservative confidence intervals.
- **No influential outliers**: The influence of an observation i on the model can be measured by its **leverage** $\hat{\boldsymbol{H}}_{ii}$ i.e. the i-th diagonal element of the hat matrix. Plotting the leverage vs. the residual value highlights aœny outlier point. Additional measures, like Cook's distance, can be used to quantify the effect of each observation.

3.5.2.1 Residuals Analysis for the Logit Model

While the overall principles of model diagnostics remain important for logit regressions, the methods used for linear regression cannot be directly applied due to the non-linear nature of the logit model. Indeed, in linear regression, residuals are expected to be normally distributed with constant variance. However, in logistic regression:

- **Non-normality**: The residuals in logistic regression are not normally distributed. Instead, they follow a binomial distribution.
- **Heteroscedasticity**: The variance of residuals is not constant but depends on the predicted probability.
- **Bounded nature**: Residuals in logistic regression are bounded, unlike in linear regression where they can take any value.

These characteristics make traditional residual plots (like residuals vs. fitted values) less informative and potentially misleading for logistic regression.

Several methods have been proposed to find counterparts to the linear diagnostics for the logit model, like deviance residuals or Pearson residuals [Pierce and Schafer, 1986; McCullagh and Nelder, 1989]. However, in practice, they often fall short of providing a clear understanding are the practical problems in a logit model.

DHARMa (Diagnostics for HierArchical Regression Models) [Hartig, 2018] is a method designed to create readily interpretable residuals for generalized linear mixed models (GLMMs), including logistic regression. The key ideas of DHARMA are:

- **Simulation-based approach**: DHARMA simulates new data from the fitted model multiple times.
- **Quantile residuals**: It calculates the quantile of the observed data within the simulated data distribution.
- **Uniformity expectation**: If the model is correctly specified, these quantile residuals should follow a uniform distribution.

The generated DHARMa residuals can be used to verify all the classical diagnostics plots presented in the previous subsection. This allows us to overcome the limitations of traditional residual analysis in logistic regression and gain insights into model fit and potential issues. This approach provides a robust method for assessing the validity of our logistic regression models used in binary classification tasks or when analyzing binary outcomes in ML experiments.

3.6 Model Correction

When testing and validating an explanatory model, some of the CLRM assumption violations can be detected through the diagnostics plots described in the previous section. Changing and adapting the model to satisfy these assumptions is usually a trial-and-error process, but some general guidelines can still be identified.

3.6.1 Model Misspecification

When the model is misspecified, the relationship between some exogenous variables and the endogenous variable is not linear. This type of error is arguably the most problematic one, since it means we're estimating the wrong relationship. In this case, the estimators are biased and inconsistent, leading to incorrect conclusions about the relationships we want to explain. Misspecification can be seen through patterns in the residuals vs. fitted values plot, which can indicate the true relationship between X and Y.

The most common way of correcting misspecification is to gain modelization insights about how what kind of non-linearity is at play. For example, it is common for variables such as `age`, the age of an individual to have a quadratic effect on some outcome Y. **Adding non-linear transformations of the variables** to the model can thus be a straightforward way to solve misspecification, with transformations such as polynomial transformations or log transformations.

Another common cause for misspecification is the **lack of product variables**. For example, if a deep model's size *and* the choice of a model's architecture interact in a non-linear way to produce the final performance, not considering the product variable can result in a non-linear dependence.

Misspecification can also be caused by an **omitted variable**, a situation where we did not take into account an important explanatory variable in the model. For example, if

a facial recognition system performance is heavily influenced by the gender of the person, but the regression is performed only on their ethnicity, the residuals will tend to group in two distinct clusters, one for each gender[1].

3.6.2 Endogeneity

Endogeneity occurs when an exogenous variable X is correlated with the error term ε. This typically happens when there exists an unobserved variable Z that affects both X and Y, creating a confounding effect. As with model misspecification, endogeneity makes the OLS estimators both biased and inconsistent, leading to incorrect inferences about causal relationships.

To understand endogeneity, causal graphs (Directed Acyclic Graphs, or DAGs), heavily advocated by Pearl and Mackenzie [2018], provide a useful visualization framework. Consider a situation in face recognition where we want to measure the effect of the brightness of the image X (measured as the mean pixel values in the image) on model performance Y. If the ethnicity Z affects both the mean brightness and the performance (because the model is undertrained on minorities), but is unobserved, we have:

$$Z \to X \to Y \quad \text{and} \quad Z \to Y \tag{3.40}$$

This graph indicates that Z impacts Y directly but also impacts X, which in turn impacts Y, creating a new indirect effect. For instance, Caucasian people are often the most represented group in face recognition datasets, which leads the model to have better performance on this subgroup. At the same time, images depicting Caucasian people are correlated with higher mean pixel values. The presence of this **backdoor path** through ethnicity Z makes it impossible to identify the true causal effect of image brightness X on model performance Y through simple regression.

This is where background illumination can serve as an instrumental variable: it is associated with a variation in overall brightness that is independent of ethnicity. This last point is a modelization hypothesis that can be assessed by verifying the data collection process.

If the confounding variable Z is observed, the solution is straightforward: including it as a control variable in the regression blocks the backdoor path. However, when Z is unobserved (in our example, ethnicity is a protected attribute), we need to use **instrumental variables** (IV). An instrumental variable W must satisfy three conditions:

- **Relevance**: W must be correlated with X
- **Exclusion**: W must affect Y only through X
- **Exogeneity**: W must be independent of unobserved confounders

In our example, background illumination is a valid instrument: it is correlated with the overall image brightness, independent of ethnicity, and if the model has seen various background illumination situations, should only affect model performance through the overall image brightness.

When a valid instrument is found, the relationship is estimated using **Two-Stage Least Squares** (2SLS). First, X is regressed on W to obtain predicted values $\hat{X}$ that are free from the influence of Z. Then, Y is regressed on these predicted values to obtain unbiased estimates of the causal effect. The key challenge in practice is finding instruments that satisfy all three conditions, as violation of any condition can lead to estimates that are more biased than simple OLS.

[1]Here, we consider for simplicity a modelization with exactly 2 genders

3.6.3 Heteroscedasticity

When the variance of the error term varies with the exogenous variables or the predictions, we say there is heteroscedasticity. Unlike misspecification or endogeneity, heteroscedasticity does not bias the OLS estimators: they remain consistent and centered on the true values. However, it affects their efficiency, making the standard errors and confidence intervals unreliable. In machine learning evaluation, heteroscedasticity is common: for example, when studying model performance, the variance of the performance often increases with model size, as larger models tend to be more sensitive to initialization and training conditions.

Heteroscedasticity can be detected through patterns in the residuals vs. fitted values plot. If the spread of residuals consistently increases or decreases with fitted values, this suggests heteroscedasticity. For example, a fan-shaped pattern in the residuals plot, where the spread of residuals increases with the fitted values, is a classic sign of heteroscedasticity.

The most common solution is to use **heteroscedasticity-robust standard errors**, also known as White standard errors. These adapt the standard error computation to account for varying error variance, providing valid confidence intervals even under heteroscedasticity. Another approach is to use **weighted least squares (WLS)**, where observations are weighted inversely to their error variance. However, this requires knowing or estimating the error variance structure.

Sometimes, heteroscedasticity can be reduced through variable transformations. For example, if the variance increases with the mean, a log transformation of the dependent variable might help stabilize the variance. However, such transformations should be used cautiously as they change the interpretation of the coefficients.

3.7 Conclusion

We have presented a comprehensive overview of statistical tools that can significantly enhance the analysis and evaluation of machine learning models. By adapting econometric methods to the context of machine learning, we have demonstrated how researchers can move beyond simple performance metrics to gain deeper insights into model behavior and performance.

The ordinary least squares (OLS) regression provides a fundamental framework for understanding the relationships between various factors and model performance. Its extension to Analysis of Variance (ANOVA) offers a powerful tool for assessing the relative importance of different categorical variables in explaining model behavior. For scenarios involving binary outcomes, often encountered in classification tasks, the logistic regression model offers a robust approach to analysis.

These statistical techniques offer several key advantages in the context of machine learning evaluation. Under appropriate assumptions, these methods allow us to draw causal conclusions about the factors influencing model performance, going beyond mere correlation. Moreover, the confidence intervals and hypothesis tests associated with these methods provide a rigorous framework for assessing the reliability of our findings.

However, it's important to note that these methods are not without limitations. The assumptions underlying these techniques (such as linearity, homoscedasticity, and normality of residuals) may not always hold in the context of complex machine learning models. Despite these challenges, the rigorous application of these statistical tools can significantly enhance our understanding of machine learning models. They provide a complementary perspective to traditional machine learning evaluation metrics, offering insights into not just how well a model performs, but why it performs as it does.

In a time of drastic change, it is the learners who inherit the future. The learned usually find themselves equipped to live in a world that no longer exists.

— Eric Hoffer, *Reflections on the Human Condition*

4

An Analysis of Initial Training Strategies for Exemplar-Free CIL

Contents

4.1 Introduction

Real-world applications of machine learning often involve training models from data streams characterized by distributional changes and limited access to past data [Hayes and Kanan, 2022; Van de Ven and Tolias, 2019]. This scenario presents a challenge for standard ML algorithms, as they assume that all training data is available at once. Continual learning addresses this challenge

[*] Equal contribution

by building models designed to incorporate new data while preserving previous knowledge [Ring, 1997]. This highlights the main challenge in continual learning: finding a balance between keeping old knowledge (stability) and learning new things (plasticity) [Mermillod et al., 2013].

Within continual learning, Class-Incremental Learning (CIL) deals with situations where the data stream is composed of batches of classes. CIL algorithms need to learn these new classes while still performing well on old ones. This task becomes even harder in the exemplar-free setting (EFCIL), when storing examples of previous classes is impossible due to memory or confidentiality constraints [Hayes and Kanan, 2020; Zhu et al., 2022]. EFCIL applies to real-world situations, like healthcare apps where patient privacy is important, or small devices with limited storage. These practical needs have led to different EFCIL methods, mainly falling into two groups:

1. **Fine-tuning with Distillation:** These methods, used by many popular EFCIL approaches [Jodelet et al., 2021; Li and Hoiem, 2016; Rebuffi et al., 2017; Zhu et al., 2021a,b, 2022; Madaan et al., 2023], updates the whole model at each step. It uses supervised fine-tuning with a distillation loss to avoid forgetting. While it works well, it often focuses more on learning new things than keeping old knowledge.
2. **Feature Extraction with Classifier Update:** Another group of methods [Hayes and Kanan, 2020; Petit et al., 2023] keeps the initial feature extractor the same and only updates the classifier. This has become more popular with the rise of pre-trained models, often created through self-supervised learning on big datasets [He et al., 2020a; Oquab et al., 2023].

While pre-trained models offer more transferable features, they're not optimal for every task [Abnar et al., 2022]. We still don't fully understand how pre-training strategies, model structures, and task features work together in EFCIL.

This work aims to fill this gap by providing a thorough analysis framework to untangle the many factors that affect EFCIL performance. We focus on ways to get the starting model for the incremental learning process. We carefully look at:

- How much does the choice of neural network structure matter ?
- How different EFCIL training methods affect performance ?
- Do pre-trained model perform better than fine-tuned ones ?
- What is the importance of the specific datasets used as benchmarks ?
- How different types of supervision in pre-training change things ?

To ensure our findings are solid and widely applicable, we test these initial training strategies using three EFCIL algorithms, representative for the state of the art. We use 16 different datasets and two challenging CIL scenarios. This comprehensive approach enables the understanding of the complex workings of class-incremental learning without keeping old examples.

The results of our analysis, shown in Table 4.1, offer useful insights for researchers and practitioners in continual learning. By explaining the key factors that drive EFCIL performance, this work aims to help develop better and more flexible machine learning systems that can keep learning in real-world situations with limited data.

Main Findings

We find that:

1. No combination of an EFCIL algorithm and an initial training strategy is best in all cases, as shown in Table 4.1, echoing the results of previous studies such as Belouadah et al. [2021]; Feillet et al. [2023].
2. The main factor influencing Average Incremental Accuracy is the pre-taining type,

Initial training strategy					CIL Algorithms					
					BSIL [2021]		DSLDA [2020]		FeTrIL [2023]	
Arch	Method	FT	Ext	Sup	$\mu_{\overline{Acc}}$	#Best	$\mu_{\overline{Acc}}$	#Best	$\mu_{\overline{Acc}}$	#Best
RN50	CE	✓	×	SL	44.9	0	53.7	4	51.0	0
RN50	CE	×	✓	SL	39.9	0	61.4	0	60.6	0
RN50	CE	✓	✓	SL	62.9	1	65.3	0	68.4	1
RN50	BYOL	✓	×	SSL	11.2	0	42.2	0	34.4	0
RN50	BYOL	×	✓	SSL	35.3	0	63.3	0	62.0	0
RN50	BYOL	✓	✓	SSL	60.2	0	70.0	2	70.2	0
RN50	MoCoV3	✓	×	SSL	14.9	0	49.6	0	41.1	0
RN50	MoCoV3	×	✓	SSL	36.3	0	67.9	1	65.3	0
RN50	MoCoV3	✓	✓	SSL	64.7	2	71.8	2	72.0	0
ViT-S	DeiT	×	✓	SL	35.0	0	58.7	0	56.3	0
ViT-S	DeiT	✓	✓	SL	11.2	0	37.4	0	27.4	0
ViT-S	DINOv2	×	✓	SSL	70.4	4	75.7	9	72.4	6
ViT-S	DINOv2	✓	✓	SSL	24.0	0	45.9	0	39.2	0

Table 4.1: Performance of three EFCIL algorithms with different training strategies for the initial model, averaged over 16 target datasets and two EFCIL scenarios. BSIL [Jodelet et al., 2021] is a recent EFCIL algorithm which is representative of fine-tuning-based CIL works. DSLDA [Hayes and Kanan, 2020] and FetrIL [Petit et al., 2023] adapt linear probing [Kumar et al., 2022] for EFCIL. We present the averaged incremental accuracy ($\mu_{\overline{Acc}}$) and the number of cases (W) in which a combination of algorithm and initial training strategy performs best for a combination of target dataset and EFCIL scenario (see Section 4.4). Initial training strategies are defined by: Arch- deep architecture used (ResNet50 (RN50)[He et al., 2016a] or vision transformer (ViT-S)[Dosovitskiy et al., 2021]); Method - initial training method; FT - fine-tuning on initial classes of the target dataset; Ext- use of an external dataset, such as ILSVRC [Russakovsky et al., 2015]; Sup - type of supervision for the initial model: self-supervised (SSL) or supervised (SL).

though the importance of the first state. Other metrics, such as Forgetting, are more impacted by the EFCIL algorithm.

3. Pre-training with external data improves accuracy, particularly when the domain gap is reasonable.
4. Self-supervision in the initial step boosts incremental learning, particularly when the pre-trained model is fine-tuned on the initial classes
5. EFCIL algorithms based on transfer learning have better performance than their fine-tuning-based counterparts

These conclusions are drawn from our rigorous statistical analysis detailed in Section 4.5, which is used to formulate EFCIL-related recommendations in Section 4.6.

The proposed framework can improve the evaluation and analysis of EFCIL methods. Continual learning practitioners can use the results of this study to better design their incremental learning systems.

4.2 Related work

4.2.1 Approaches to CIL/EFCIL

Class-Incremental Learning (CIL) and its variant, Exemplar-Free Class-Incremental Learning (EFCIL), have seen significant developments in recent years. These approaches can be broadly categorized into two main types: fine-tuning based methods and transfer learning based methods.

Fine-tuning based methods in CIL/EFCIL typically involve updating all or most of the model parameters at each incremental step. These methods often employ various techniques

to mitigate catastrophic forgetting.

LUCIR (Learning a Unified Classifier Incrementally via Rebalancing) [Hou et al., 2019] proposes a cosine normalization strategy to address the classifier bias towards newly added classes. It also introduces a less-forget constraint and an inter-class separation loss to maintain discrimination between old and new classes.

BSIL (Balanced Softmax for Incremental Learning) [Jodelet et al., 2021] addresses the class imbalance problem in CIL by introducing a balanced softmax cross-entropy loss. This approach aims to improve the model's ability to learn from imbalanced data distributions that naturally occur in incremental learning scenarios.

Advantages of Fine-tuning Based Methods

- High plasticity, allowing quick adaptation to new classes
- Potential for better performance on new tasks
- Flexibility in modifying the entire model architecture

However, fine-tuning-based methods also face several limitations. First, they are more prone to catastrophic forgetting, especially without careful regularization. The optimal weights at a given incremental step may completely fail to correctly represent the classes of the previous steps. This is particularly the case when the incremental steps contain less data than the initial step, where overfitting prevails on generalization. Tackling this problem often require complex strategies to balance new learning and old knowledge retention, sometimes at the cost of the fixed memory constraint if one decides to modify the entire architecture to accommodate for the new classes.

In contrast to fine-tuning approaches, **transfer learning-based methods** in CIL/EFCIL typically involve freezing most of the pre-trained model and only updating a small part of it, usually the classification layer.

DSLDA (Deep Streaming Linear Discriminant Analysis) [Hayes and Kanan, 2020] freezes the feature extractor of a pre-trained model and updates only an LDA classifier. This approach allows for efficient incremental learning with minimal computational overhead.

FeTrIL (Feature Translation for Incremental Learning) [Petit et al., 2023] introduces a feature translation mechanism to adapt the frozen feature space of a pre-trained model to new classes. This method aims to bridge the domain gap between the pre-training dataset and the target incremental learning task.

Advantages of Transfer Learning Based Methods

- Strong resistance to catastrophic forgetting
- Computational efficiency, especially for large-scale problems
- Potential for better performance in low-data regimes

However, transfer learning-based methods also have limitations. In particular, they may struggle with tasks that are significantly different from the pre-training domain. If the incremental data interpolates the data used in pre-training, transfer learning is able to model the new data features, but as the domain gap grows, there are no garenties of a good representation if the data in the latent space. Transfer-learning-based methods also have limited plasticity, potentially leading to suboptimal performance on new tasks, since their representation power

entirely depends on the pre-training task. The choice of a pre-training dataset aligned with the expected incremental task becomes crucial for these methods.

Both fine-tuning and transfer learning-based approaches offer unique advantages and face distinct challenges in CIL/EFCIL scenarios. Recent work [Petit et al., 2023] has shown that the choice between fine-tuning and freezing can have a significant impact on CIL performance. Freezing pre-trained models and using techniques like feature translation can lead to competitive performance while maintaining high computational efficiency. The choice between these approaches often depends on the specific requirements of the task, the available computational resources, and the characteristics of the data stream.

4.2.2 Pre-training Strategies in CIL

Pre-training has emerged as an important factor in the performance of CIL systems. The choice of pre-training strategy can significantly impact the model's ability to adapt to new classes while retaining knowledge of previously learned ones. More and more EFCIl methods chose to used a pre-trainined model to bootstrap the incremental performance [Hayes and Kanan, 2020; Hayes et al., 2020; Wang et al., 2022c; Goswami et al., 2024]. Recent works such as Panos et al. [2023]; Lee et al. [2023] highlight the question of fairly evaluating CIL methods relying on various pre-training strategies.

Supervised pre-training involves training a model on a large labeled dataset before adapting it to the CIL task. This approach has been widely used in CIL research due to its straightforward nature and the availability of large labeled datasets.

Common Supervised Pre-training Datasets

- ImageNet [Deng et al., 2009]: A large-scale dataset with over 1 million images across 1000 classes.
- Places [Zhou et al., 2017]: A scene-centric dataset with over 10 million images across 400+ unique scene categories.

Models pre-trained on these large-scale datasets often exhibit strong transfer learning capabilities, providing a good starting point for CIL tasks. However, the success of supervised pre-training heavily depends on the similarity between the pre-training dataset and the target CIL task.

Self-supervised pre-training has gained significant attention in recent years due to its ability to learn useful representations from unlabeled data. This approach has shown promising results in CIL, often matching or surpassing supervised pre-training in certain scenarios [Gallardo et al., 2020; Fini et al., 2022; Ahmad et al., 2022; Zhu et al., 2022].

Popular Self-supervised Pre-training Methods

- BYOL (Bootstrap Your Own Latent) [Grill et al., 2020]
- MoCo v3 (Momentum Contrast) [Chen et al., 2021]
- DINOv2 (Self-Distillation with Noisy Students) [Oquab et al., 2023]

BYOL [Grill et al., 2020] uses two neural networks that learn to predict each other's output, creating a form of self-supervision. MoCo v3 [Chen et al., 2021] employs a contrastive learning approach with a momentum encoder to learn visual representations. DINOv2 [Oquab et al., 2023] extends the DINO framework with improved self-distillation techniques and scaled-up

training. These self-supervised methods have shown remarkable performance in CIL tasks, often providing more transferable features compared to supervised pre-training, especially when the target task differs significantly from the pre-training dataset.

The optimal pre-training strategy and the decision to fine-tune or freeze depend on various factors, including the nature of the CIL task, the available computational resources, and the characteristics of the data stream. As the field progresses, developing adaptive strategies that can automatically select the best approach based on these factors remains an important area of research.

4.3 Problem statement

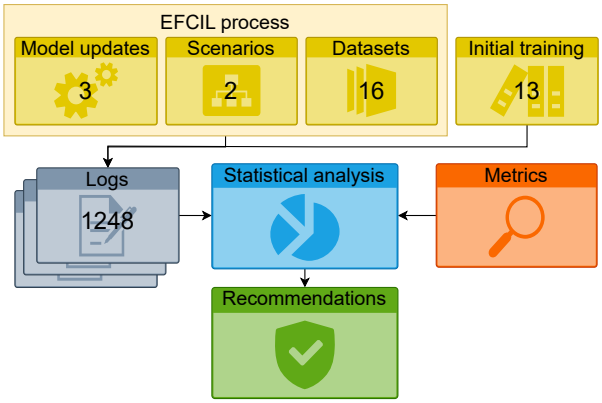


Figure 4.1: Overview of the proposed analysis framework of initial training strategies for EFCIL.

We summarize our proposed analysis framework in Figure 4.1. It combines a comprehensive modeling of the EFCIL process and initial training strategies as inputs for a statistical analysis that uses different EFCIL metrics. Recommendations for the design of EFCIL approaches are made based on the conclusions of the statistical analysis.

4.3.1 EFCIL process

Let us consider a dataset $\mathcal{D}$ split over K subsets, $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \cdots \cup \mathcal{D}_K$, and an EFCIL algorithm $Incr$. A CIL process consists in learning a classification model sequentially over K non-overlapping steps using $Incr$. At each step $k \in [\![1, K]\!]$, the model is updated using $Incr$ and the data subset $\mathcal{D}_k$, whose associated set of classes is denoted by $\mathcal{C}_k$. The data subsets $\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_K$ composing the complete dataset $\mathcal{D}$ satisfy the following constraint: for $k, k' \in \{1, 2, \ldots, K\}$ with $k \neq k'$, $\mathcal{C}_k \cap \mathcal{C}_{k'} = \emptyset$, i.e. each class is only present in a single data subset. The use of an exemplar-free algorithm $Incr$ implies that when the training is performed at the k^{th} step, no example from any of the data subsets of the previous steps can be accessed. Although this is a more difficult setting, it is also more realistic in practice [Hayes and Kanan, 2022; Belouadah et al., 2021].

Incremental model updates. The initial model $\mathcal{M}_1$ is obtained following one of the training strategies presented in 4.3.2. At the k^{th} step of the CIL process, $k \in [\![2, K]\!]$, the classification model $\mathcal{M}_k$ recovers the weights of the model $\mathcal{M}_{k-1}$ obtained in step $k-1$ and is updated using the data subset $\mathcal{D}_k$ and the algorithm $Incr$. Many EFCIL algorithms [Jodelet et al., 2021] fine-tune all network weights at each incremental step, thus favoring plasticity. Alternatively, algorithms such as [Hayes and Kanan, 2020; Petit et al., 2023] only retrain the

classifier, thus favoring stability. As a compromise, it is also possible to freeze a part of the model and to update only the last layers. We cover these three cases in our experiments.

Scenario. A CIL scenario is characterized by the distribution of classes among the steps of the CIL process. We denote by b the proportion of the classes available in the initial step: $b = Card(\mathcal{C}_1)/Card(\mathcal{C})$. There are two commonly used scenarios [Belouadah et al., 2021]:

i. equal splitting of classes across the steps
ii. half of the classes in the first step and the rest of the classes are divided equally between subsequent steps

4.3.2 Training strategies for the initial model

In the following, we describe the main characteristics of the training strategies used in our experimental study to obtain the initial model of the incremental learning process. Further experimental settings are reported in Section 4.4.

Network architecture. So far, most CIL methods have been proposed in combination with a convolutional neural network, but visual transformer (ViT) networks have recently gained popularity in CIL [Douillard et al., 2022]. In order to provide a fair comparison between the two types of architecture, we use a ResNet50 [He et al., 2016a], and a ViT-Small [Dosovitskiy et al., 2021] network, which have a close number of parameters (23.5M and 22.1M parameters respectively).

Model initialization. At the first step of the CIL process, the weights of the model may either be randomly initialized or transferred from a pre-trained model. In the second case, depending on the choice of the user, the dataset $\mathcal{D}^*$ used for pre-training may either be an auxiliary dataset (e.g. ILSVRC [Russakovsky et al., 2015]), referred to as *source* dataset, or the first data subset $\mathcal{D}_1$ of the incremental process.

Label availability. We consider that all examples from the target dataset $\mathcal{D}$ are labeled, and we experiment with both supervised learning and self-supervised learning to obtain the initial model using D_1. Labels may not be available for the external dataset $\mathcal{D}^*$. In this case, the training initialization is performed using a self-supervised pre-training algorithm (e.g., DINOv2 [Oquab et al., 2023]).

4.4 Experimental setting

We describe the experimental parameters and the metrics we use to evaluate EFCIL models. The combination of parameters results in 1,248 experiments in total (Figure 4.1).

4.4.1 Initial training strategies

We compare different strategies for training an initial model, as summarized in Table 4.1. We use Resnet50 [He et al., 2016a] and ViT-S [Dosovitskiy et al., 2021] networks, which are representative of CNNs and transformers and have similar sizes. The training is done either using a self-supervised method (BYOL [Grill et al., 2020], DINOv2 [He et al., 2020a], MoCov3 [Chen et al., 2021]) or a supervised one (DeiT, cross-entropy (CE)). We present results for pre-training with external data (i.e. ILSVRC [Russakovsky et al., 2015] for BYOL, DeiT and CE; a 150M-images dataset + ILSVRC for DINOv2) and training on the first batch. We compare the effect of :

i. freezing the weights of the pre-trained model
ii. further optimizing the last layers of the model (e.g. the last convolutional block in ResNet50) on the initial data subset $\mathcal{D}_1$

The first type of experiment is denoted by the suffix "-*t*" (transfer), the second by the suffix "-*ft*" (fine-tuning). In the case where the pre-training algorithm is applied to $\mathcal{D}_1$ and not to $\mathcal{D}^\star$, there is no suffix.

4.4.2 Target datasets

For a comprehensive evaluation and to account for the diversity of visual tasks, we evaluate the training strategies on 16 target datasets, sampled from publicly available datasets. They cover different domains (plants, animals, landmarks, food, faces, traffic signs etc.), and different types of images (natural, drawings, paintings). IMN100_1 and IMN100_2 consist of 100 classes randomly selected from ImageNet-21k [Deng et al., 2009]. Flora is a thematic subset of ImageNet consisting of 100 classes belonging to the "flora" concept. IMN100_1, IMN100_2 and Flora have no mutual overlap and no overlap with ILSVRC [Deng et al., 2009; Russakovsky et al., 2015]. Amph100 and Fungi100, sampled from iNaturalist [Van Horn et al., 2018], respectively contain 100 classes of amphibians and fungi, selected so as to avoid overlap with animal and fungi classes from ILSVRC. We also sample 100-class subsets from other popular datasets: WikiArt100 [Saleh and Elgammal, 2015], Casia100 [Yi et al., 2014], Food100 [Bossard et al., 2014], Air100 [Maji et al., 2013], MTSD100 [Madani and Yusof, 2016], Land100 [Weyand et al., 2020b], Logo100 [Wang et al., 2020a] and Qdraw100 [Ha and Eck, 2017]. Finally, we consider three 1000-class subsets: Casia1k [Yi et al., 2014], Land1k [Noh et al., 2017], and iNat1k [Van Horn et al., 2018]. The number of training images per dataset varies from 60 to 750. More details on the datasets are provided in the appendix.

4.4.3 Evaluation Metrics

The performance of EFCIL models can be evaluated in several ways [Masana et al., 2021]. However, two particular measures became predominant.

- **Average incremental accuracy** $\overline{Acc}$: In EFCIL, a model trained over a K-step incremental process is commonly evaluated using the average incremental accuracy [Zhu et al., 2021b, 2022, 2021a; Jodelet et al., 2021]. We denote it by $\overline{Acc}$ and compute it by:

$$\overline{Acc} \;=\; \frac{1}{K-1}\sum_{k=2}^{K} acc(\mathcal{M}_k, \bigcup_{i=1}^{k}\mathcal{D}_i) \tag{4.1}$$

  where $acc(\mathcal{M}, D)$ is the accuracy of the model $\mathcal{M}$ on the dataset D. Following common practice in CIL [Castro et al., 2018; Petit et al., 2023; Zhu et al., 2022], $\overline{Acc}$ does not take the accuracy of the initial model into account.
- **Average forgetting** F**.** Average forgetting, denoted here by F, is computed by:

$$F \;= b \times f(\mathcal{D}_1) + \frac{1-b}{K-1}\sum_{k=2}^{K} f(\mathcal{D}_k) \tag{4.2}$$

  where $f(\mathcal{D}_k) = \max_{k' \in [\![k,K]\!]} acc(\mathcal{M}_{k'}, \mathcal{D}_k) - acc(\mathcal{M}_K, \mathcal{D}_k))$ is the difference between the best performance achieved on the data subset $\mathcal{D}_k$ during the EFCIL process and the final performance of the model on this data subset [Mirzadeh et al., 2022].
- **Initial accuracy** $\overline{Acc}_1 = acc(\mathcal{M}_1, \mathcal{D}_1)$: Since the average incremental accuracy is highly driven by the accuracy of the first model on the first data subset, is is important to consider this quantity as well.
- **Final accuracy** $\overline{Acc}_K = acc(\mathcal{M}_K, \mathcal{D})$: This depicts the performance of the final model in practice, on real data sampled from $\mathcal{D}$.

The most commonly used metric is the average incremental accuracy $\overline{Acc}$, but it has some caveats: it gives more weight on early classes, since the model is evaluated on all past classes at each step k. Thus, a high average incremental accuracy is not predictive of the accuracy on the last classes. This is why considering an additional metric, such as F or $\overline{Acc}_K$, is necessary to control the stability of the performance over time.

4.4.4 Incremental learning

EFCIL scenario *b***.** We experiment on two widely used CIL scenarios [Hou et al., 2019; Belouadah et al., 2021]. In the first scenario, the classes are equally distributed over 10 steps, e.g. 10 classes per step for a 100-class dataset. In the second scenario, half of the classes are learned in the initial step, and the other half is equally distributed over 10 incremental steps, e.g. $50 + 10 \cdot 5$ classes for a 100-class dataset.

CIL algorithm *Incr***.** We experiment with one fine-tuning based algorithm, namely BSIL[Jodelet et al., 2021], which adds a balanced softmax without exemplars to LUCIR [Hou et al., 2019]. We also experiment with two fixed-representation-based algorithms, namely DSLDA [Hayes and Kanan, 2020] and FeTrIL [Petit et al., 2023].

4.5 Analysis of results

We present a statistical analysis of the results from Table 4.1, which highlights the effects of pre-training strategies and of EFCIL algorithms on EFCIL performance. The statistical model and associated findings are presented below.

4.5.1 Modeling causal effects

Our objective is to identify the primary factors that influence the performance of EFCIL algorithms. To interpret causal effects, we employ multiple linear regressions using the Ordinary Least Squares (OLS) method, following the statistical and econometric practices described in Chapter 3 For a given experiment, we denote by `Y` the target metric accuracy (endogenous), `Data` the evaluation dataset (exogenous), `Train` the initial training strategy (exogenous), and `Incr` the incremental algorithm (exogenous). We also consider the initial accuracy $\texttt{Acc}_1$ as an endogenous variable that may influence performance and can be controlled in our regressions. Other parameters, such as the total number of classes or the dataset, are examined as potential predictors of a metric.

We fit regressions of the form

$$\texttt{Y} \sim \texttt{Train} + \texttt{Incr} + \texttt{Data} \tag{4.3}$$

Since `Train`, `Incr`, and `Data` are categorical, we encode them as one-hot vectors. Thus, β_1, β_2, and β_3 are vectors of the same size as the number of possible categories for each variable. Other variables were considered, presented bellow:

Summary of variables

To explain the variable $\overline{\texttt{Acc}}$ and `F`, we consider the variables

- $\texttt{Acc}_1$: the accuracy of the first state,
- `Data`: dummy variable for the type of target dataset,
- `Train`: dummy variable for the initial training strategy,

- `Incr`: dummy variable for the incremental method used,
- $\texttt{n}_{mean}$: the mean number of images per class in the experiment,
- `Small`: binary variable encoding if the training images are so small that they have to be up-scaled,
- `Width`: mean width of the images used for the experiment,
- `B`: binary variable encoding for the 2 possible CIL scenarios (i.e. either 10% or 50% of the total number of classes learned in the initial step of the process),
- `N`: the total number of classes,
- $\texttt{N}_1$: the number of images in the first state.

The statistical significance of these effects is assessed by examining the *p-value* of the associated Student t-test for each coefficient [Gareth et al., 2013]. Following established statistical practices [Gareth et al., 2013], we set the significance value at .05. The significance, sign, magnitude, and interpretation of each estimated coefficient depend on the regression model. In particular, introducing more exogenous variables can cause instability in the regression. Therefore, for each metric Y, we adopt the following methodology to select only the most influential factors:

1. We use multiple regression models to represent the evaluation metric Y as a linear combination of different variables, or of the product of these variables. We ensure that the chosen regressions exhibit no collinearity or numerical issues[1].

2. Subsequently, we select a regression model using the Akaike Information Criterion (AIC) [Akaike, 1998], which regularizes the likelihood of the model based on its degrees of freedom.

3. We interpret the regression coefficients, the coefficient of determination R^2, and examine the Q-Q plot of the residuals ϵ_i to verify their normality.

4. Next, we conduct an Analysis of Variance (ANOVA) [Gareth et al., 2013] on the regression to obtain aggregated statistics on the categorical variables.

5. Finally, we interpret the partial η^2 derived from the ANOVA as a measure of the importance of each variable.

A regression on a categorical variable requires the setting of a reference value for it. Therefore, the coefficient(s) associated with this categorical variable represent the causal effects of this variable *with respect to the reference level*. However, we want to compare all initial training strategies with each other to derive practical recommendations. Therefore, we use the following protocol to generate pairwise significant differences:

1. Perform the same regression multiple times using a different reference category
2. Sum-up the pairwise comparisons in a double-entry matrix
3. Since we are performing multiple tests, we need to adjust the significance threshold of each test using Bonferroni correction [Gareth et al., 2013], which consists of dividing the *p-value* threshold by the number of tests
4. Plot a heatmap of the pairwise comparisons between the choice of a parameter.

[1]We assess this by examining the smallest eigenvalue of the Gram matrix of the data $X^T X$. Although Ridge or Lasso regression could address these concerns, their coefficients are less interpretable than those of OLS.

4.5.2 Metrics and confounding Factors

In Figure 4.2, we examine the relationship between the evaluation metrics defined in Subsection 4.4.3.

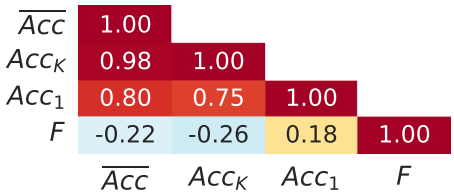


Figure 4.2: Correlation between the endogenous variables.

We observe a strong positive correlation between $\overline{Acc}$ and Acc_K. There is a weak negative correlation between average incremental accuracy and forgetting, which is expected due to the inherent trade-off between stability (i.e. low forgetting) and plasticity in CIL (i.e. high performance on new classes). We note a significant correlation between average incremental accuracy and accuracy in the initial state. This correlation is expected since half of our experiments are done with half of the classes in the initial step. Additionally, the average incremental accuracy (Eq. 4.1) evaluates each model on each class, from the first occurrence of the class to the end of the incremental process, thus giving greater influence to earlier classes. Conversely, there is a weak correlation between forgetting and initial accuracy. This implies that the performance on the initial batch of classes does not significantly impact the model's stability throughout the incremental steps.

Based on these observations, we choose the average incremental accuracy $\overline{Acc}$ and the average forgetting F as the metrics of interest for our study, and include the effect of the initial accuracy in their models. Controlling the initial accuracy in a regression model is important to draw accurate conclusions: if pure accuracy is sought, then it can be left out of the model. However, the goal of CIL algorithms is not solely to be accurate on average, but rather to be accurate while preventing forgetting. Hence, to analyse the actual incremental contribution of each method, initial accuracy should be included in the regression.

4.5.3 Variable selection

Tables 4.2 and 4.3 present the individual R^2 of each exogenous variables when regression Accuracy and Forgetting. As we see, the 4 variables that best explain both metrics are the same, and are the ones we presented in Section 4.5.1. What is striking is that for Accuracy, the most impactful variable seems to be $\texttt{Acc}_1$, while for Forgetting, `Incr` has a similar impact. For the accuracy, this is not very surprising, since the mean accuracy calculation is calculated linearly in $\texttt{Acc}_1$. The next factors seem to play a big role for explaining Accuracy, while for Forgetting, the next most influential variables have a much lower R^2.

Now that we have selected our variables, we can perform regressions using a combination of the variables to disentangle the effects of each one.

4.5.4 Factors influencing incremental performance

This subsection presents the aggregated influence of the considered parameters. The models and findings presented in Table 4.4 are obtained with the methodology presented in Subsection 4.5.1.

Variable	*p-value*	R^2
$\texttt{Acc}_1$	$2.96 \cdot 10^{-240}$	0.63
Train	$1.17 \cdot 10^{-87}$	0.33
Data	$2.25 \cdot 10^{-55}$	0.23
Incr	$7.52 \cdot 10^{-29}$	0.11
$\texttt{n}_{mean}$	$8.16 \cdot 10^{-20}$	0.07
Small	$1.84 \cdot 10^{-05}$	0.02
Width	$9.78 \cdot 10^{-03}$	0.01
B	$1.05 \cdot 10^{-01}$	0.00
N	$2.41 \cdot 10^{-01}$	0.00
$\texttt{N}_1$	$2.87 \cdot 10^{-01}$	0.00

Table 4.2: Variables predicting accuracy, sorted by decreasing importance

Variable	*p-value*	R^2
Incr	$2.20 \cdot 10^{-222}$	0.62
Train	$6.46 \cdot 10^{-15}$	0.08
$\texttt{Acc}_1$	$7.71 \cdot 10^{-10}$	0.03
Data	$2.66 \cdot 10^{-03}$	0.02
N	$7.50 \cdot 10^{-04}$	0.01
B	$3.43 \cdot 10^{-02}$	0.00
$\texttt{N}_1$	$4.13 \cdot 10^{-02}$	0.00
$\texttt{n}_{mean}$	$1.07 \cdot 10^{-01}$	0.00
Small	$6.88 \cdot 10^{-01}$	0.00
Width	$7.17 \cdot 10^{-01}$	0.00

Table 4.3: Variables predicting forgetting, sorted by decreasing importance

4.5.4.1 Main influences

In Table 4.4, the most significant factor affecting average incremental accuracy is the choice of initial training strategy.

Model	R^2	variable	η^2
$\overline{\texttt{Acc}} \sim \texttt{Incr} + \texttt{Train} + \texttt{Data}$	0.69	Train	0.32
		Data	0.24
		Incr	0.11
$\overline{\texttt{Acc}} \sim \texttt{Acc}_1 + \texttt{Incr} + \texttt{Train} + \texttt{Data}$	0.81	$\texttt{Acc}_1$	0.25
		Incr	0.22
		Train	0.10
		Data	0.06
$\texttt{F} \sim \texttt{Incr} + \texttt{Train} + \texttt{Data}$	0.71	Incr	0.61
		Train	0.06
		Data	0.03

Table 4.4: ANOVA results for each considered regression. Variables are significant at $p < 0.05$ and ordered by decreasing importance.

However, upon controlling the impact of initial accuracy, the selected incremental algorithm has a greater importance. This distinction is primarily attributed to BSIL, which exhibits an incremental accuracy 16 points below that of FeTrIL and DSLDA on average.

Regarding forgetting, the incremental algorithm is the most influential parameter. Here, this effect is not driven by any specific outlier method. Further analysis shows that initial accuracy also plays a significant role in predicting the level of forgetting. The associated regression coefficient is .16 (±.02), indicating that a 1-point increase in initial accuracy results in a 16-point increase of forgetting.

Given that accuracy ranges between 0 and 1, a lower initial accuracy decreases the likelihood of experiencing high levels of forgetting. Hence, a trade-off arises concerning the initial accuracy: while its enhancement greatly improves the average incremental accuracy, it also appears to amplify forgetting. This should be taken into account when comparing CIL algorithms. From a research perspective, the incremental algorithm remains influential in the metrics, particularly when controlling for initial accuracy or focusing on forgetting. However, in practical applications

of CIL, the final accuracy may be more important. Given its strong correlation with average incremental accuracy, increasing the initial accuracy becomes more advantageous in this case.

4.5.4.2 Model diagnostics

We validate the regression in Table 4.4 by plotting the diagnostics plots of the regression, as explained in the last section of Chapter 3.

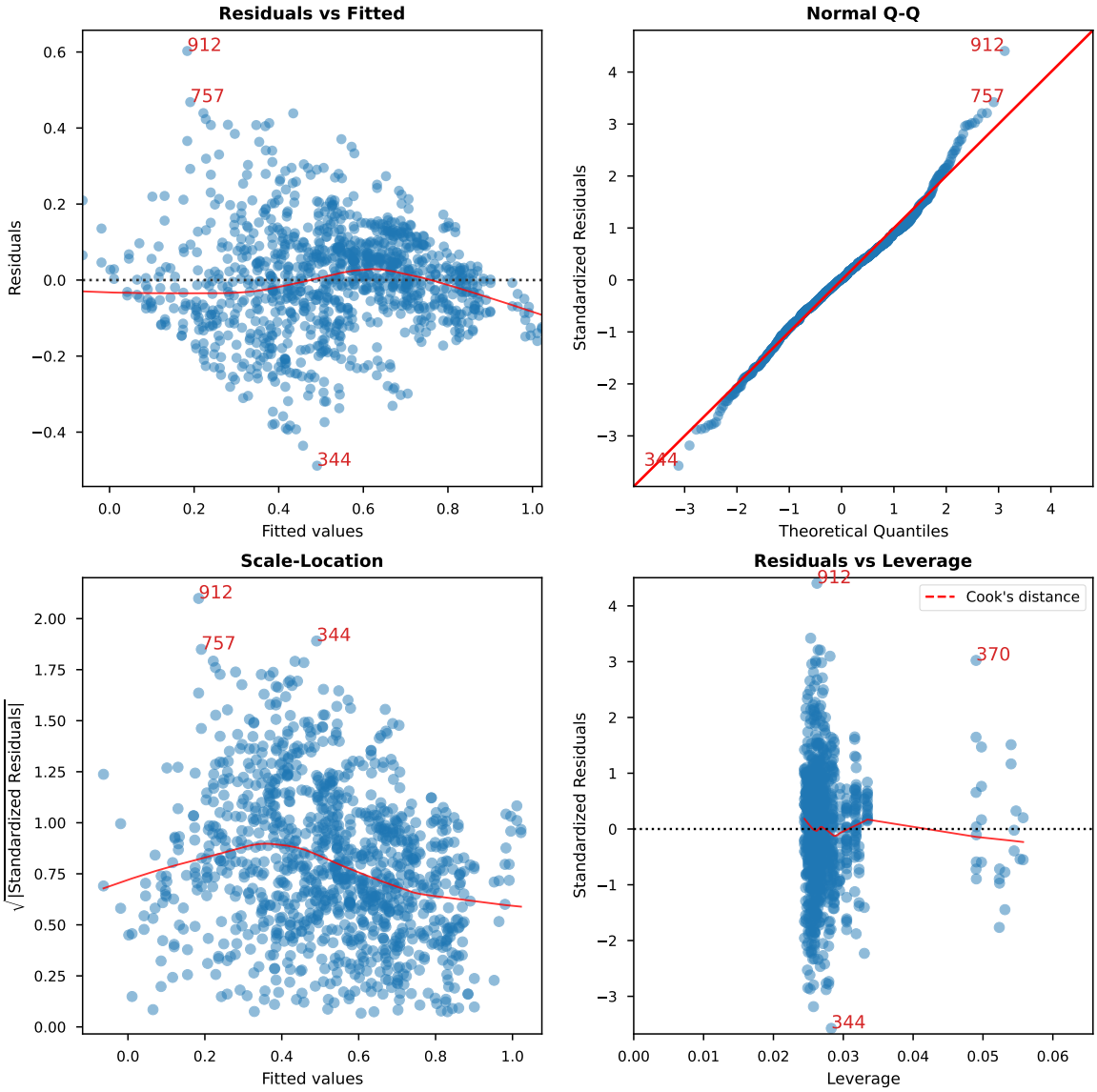


Figure 4.3: Diagnostics plots for the regression of the Average Accuracy on the incremental method, pre-training type, and training dataset.

All plots look valid. The Q-Q plot shows a small deviation from normality for small and high values, and small heteroscedasticity can be observed in the Residuals vs. Fitted plot. This is expected since the accuracy is contained between 0 and 1, so the "true" model is most likely nonlinear, creating problems at the bounds. Nevertheless, for average values of $\overline{\texttt{Acc}}$, the hypotheses of the OLS regression are not violated.

4.5.5 Comparison of initial training strategies

In Figure 4.4, we observe notable variations in accuracy among different initial training strategies, thus prompting the identification of three regimes.

Identifed Dynamics

Three pre-training regimes can be found in EFCIL:

1. **Strategies that surpass supervised learning without transfer:** MoCoV3-ft, DINOv2-t, BYOL-ft, SL(ResNet)-ft, MoCov3-t. These approaches exhibit superior performance by generating a robust latent space, whose features are transferable. MoCoV3-ft enhances its latent space by fine-tuning, enabling better generalization compared to other methods. DINOv2-t follows, leveraging its extensive self-supervised training on a very large amount of data. BYOL-ft and SL(ResNet)-ft closely follow, highlighting the advantage gained from additional adaptation steps on the target dataset following pre-training. MoCov3-t is fifth, showing that features generated through an adapted self-supervised method have a generalization capability that can be leveraged in CIL.
2. **Strategies that exhibit no significant improvement over supervised learning without transfer:** SL(ResNet)-ft, BYOL-t, SL(DeiT)-t. Our analysis underlines the capability of well-designed self-supervised methods to outperform supervised pre-training approaches.
3. **Strategies that underperform compared to supervised learning without transfer:** MoCoV3, BYOL, DINOv2-ft, SL(DeiT)-ft. The inferior performance of self-supervised methods can be attributed to the limited initial data. Furthermore, the challenging nature of fine-tuning for transformer models contributes to the underwhelming outcomes observed in these models.

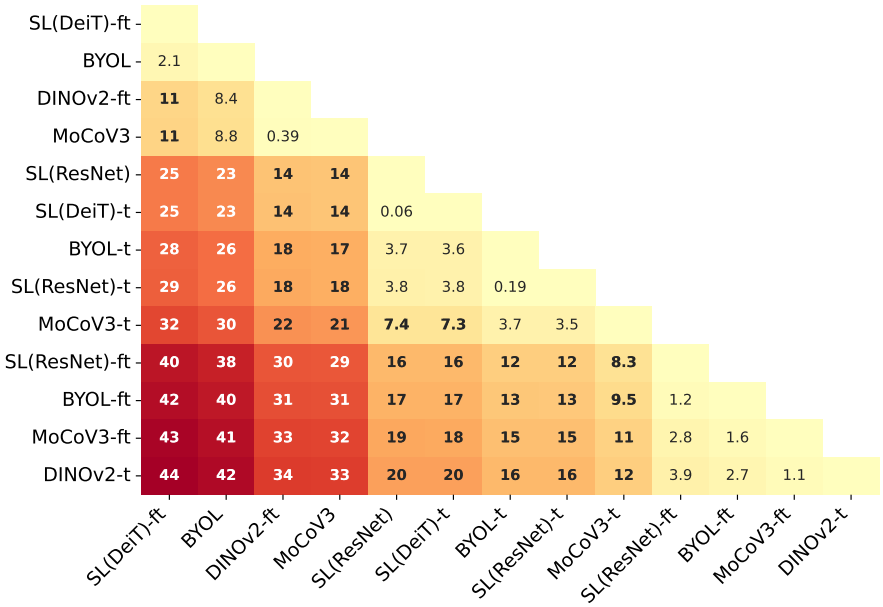


Figure 4.4: Accuracy gain by using strategy in row i over strategy in column j , e.g. "The accuracy of BYOL-ft is 17pts higher than SL(ResNet)". Only results in **bold** are statistically different.

The analysis of the average forgetting, illustrated in Figure 4.5, indicates that the distinctions

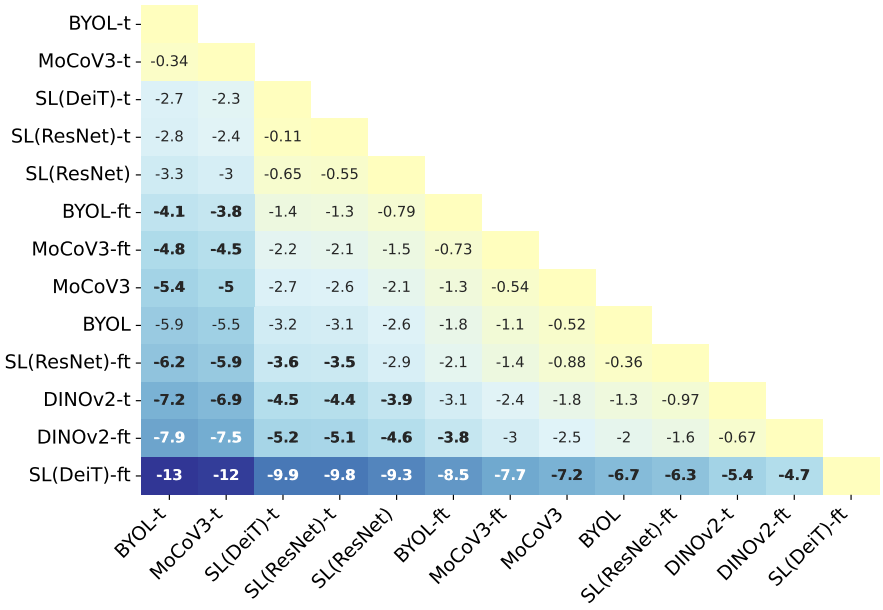


Figure 4.5: Overall pairwise comparisons on Forgetting

between many pretraining strategies are much less significant. However, DINOv2-t exhibits lower forgetting compared to other strategies, including SL (ResNet). This is particularly remarkable considering that DINOv2-t has the highest initial accuracy. Conversely, fine-tuned transfer models (DINOv2-ft, SL(DeiT)-ft) also display a lower forgetting, albeit primarily attributed to their inherently low initial accuracy, which leaves little room for further decline in their accuracy.

4.5.6 Further analysis of initial training strategies

We now inquire whether the preceding general analysis can be nuanced in specific scenarios. To this end, we perform the same analysis as in the previous section by performing the regression on subsets of the data. All complementary graphs that justify the following statements can be found in the appendix.

Influence of the dataset. Regarding target datasets that are furthest from the pre-training dataset, the benefit of pre-training with or without fine-tuning is lower due to the domain gap. We note that specialized datasets, such as Qdraw100 and Casia100, also contain smaller images than those of ILSVRC. Whether the difference in performance is caused by a semantic gap or an image-size gap is unclear.

Influence of the incremental scenario. Regarding accuracy, we find that most differences among methods come from the scenarios with 50 initial classes or less. With 10 initial classes, all strategies that were previously not significantly better than SL(ResNet) start to outperform it. In scenarios with 50 initial classes, it becomes more difficult to precisely rank the top initial training strategies. In scenarios with 100 initial classes, no strategy is significantly better than any other one (which can come from the lower number of experiments with these scenarios).

Influence of Incremental method. We find that FeTrIL and DSLDA exhibit a similar pattern for $\overline{Acc}$ and F, contrary to BSIL. For FeTrIL and DSLDA, the differences between the best initial training strategies are less clear, but the general trend previously described still holds, in particular for the accuracy. The choice of the training strategy does not clearly impact the forgetting. On the other hand, BSIL is much more sensitive to the initial training strategy. Fine-tuned methods clearly outperform classical learning and plain transfer (except

for DINOv2-t), whether it concerns the accuracy or the forgetting. Moreover, SL(ResNet) is a stronger baseline for BSIL than for the other methods when considering incremental accuracy.

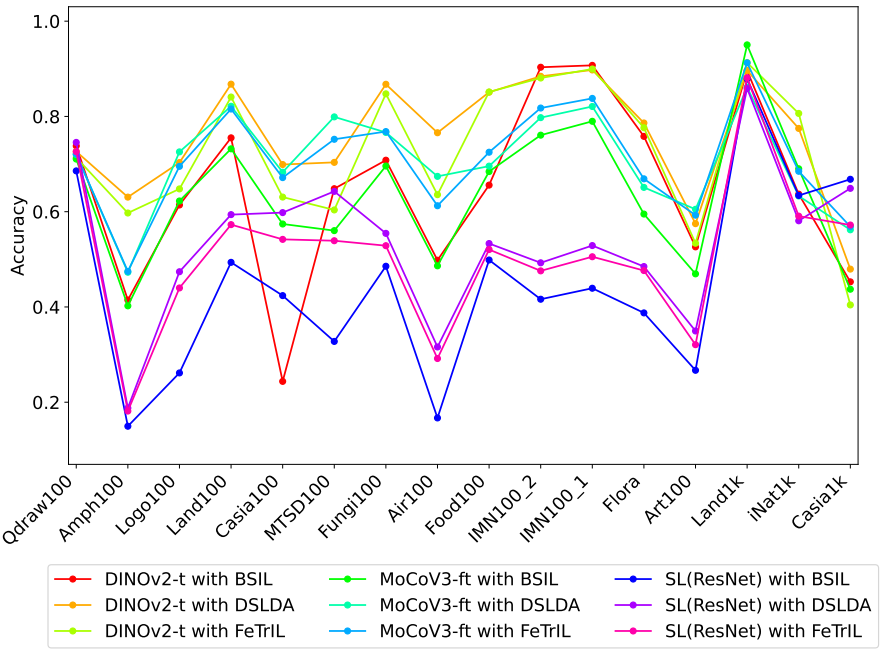


Figure 4.6: Interaction plot of the best strategies for different transfer types and for the 3 CIL algorithms. Similar slopes indicate similar behaviors. A change in slope indicates a change in behavior.

4.6 Discussion

We summarize our findings and propose recommendations for the design of EFCIL approaches.

- **Does the use of a model pre-trained on an external dataset $\mathcal{D}^\star$ always improve performance on the target dataset $\mathcal{D}$?**
  Figure 4.6 highlights that no single initial training strategy outperforms the others on all datasets. As illustrated in Table 4.1, pre-training is clearly better on average, but there are exceptions. Intuitively, the use of a pre-trained model without fine-tuning (DINOv2-t in Figure 4.6), is clearly preferable for datasets such as IMN100_1 and Flora which are closely related to the dataset used for pre-training. Inversely, the supervised training method SL(ResNet) is better when the gap between the source and the target datasets is important, such is the case for Casia1k. MoCov3-ft is a good compromise since it leverages pre-training, but adapts the representation via partial fine-tuning. The initial training strategy should be selected by considering characteristics of the dataset such as: number of classes, number of samples per class, domain gap with pre-training, and size of the initial batch of classes.
- **In the absence of an external dataset, is it better to train the initial model in a supervised way or with a self-supervised learning method ?**
  As shown in Figure 4.4, supervised learning on the initial data is better on average. However, self-supervised learning is better when the amount of data available initially is limited, making it difficult to train a supervised model effectively.

- **Should the pre-trained model be fine-tuned on the first batch of data, or frozen ?**
  Existing EFCIL works that use pre-trained transformers keep their weights fixed [Janson et al., 2022; Pelosin, 2022; Wang et al., 2022b]. This might be explained by the fact that fine-tuning these models might be detrimental in transfer learning [Kumar et al., 2022]. Inversely, the performance of CNN-based training strategies, such as BYOL or MoCov3, increases after partial fine-tuning. This is explained by the fact that the layers of CNNs are reusable across tasks, while fine-tuning the last layers with initial target data improves transferability in subsequent EFCIL steps.
- **How does the performance of EFCIL algorithms vary with initial training strategies ?**
  Table 4.1 and Figure 4.6 show that the performance of BSIL varies much more than that of DSLDA and FeTrIL. This is particularly clear for transformer models, where BSIL performance is strongly degraded when fine-tuning of pre-trained models is used. In contrast, the variation of performance for DSLDA and FeTrIL is much lower when testing partial fine-tuning and transfer strategies on top of pre-trained models. This suggests that both initial training strategies are usable in practice for transfer-learning based EFCIL algorithms.
- **What is the impact of using transformers versus convolutional neural networks ?**
  The averaged results presented in Table 4.1 and the detailed ones from Figure 4.6 show that the difference between the best training strategies based on transformers and on CNNs is small. This is particularly the case when CNNs are pre-trained in a self-supervised manner and then partially fine-tuned on the initial batch of target data. Our finding echoes those reported in recent comparative studies of the two types of neural architectures which conclude that there is no absolute winner [Pinto et al., 2022; Wang et al., 2023]. The implication for EFCIL is that the use of both types of architecture should be explored in future works.

4.7 Conclusion

We perform an analysis of EFCIL in an evaluation setting that includes numerous and diverse classification tasks. We confirm the findings of existing comparative studies which have shown that no CIL algorithm is the best in all cases [Belouadah et al., 2021; Masana et al., 2021; Feillet et al., 2023] and that algorithms based on transfer learning provide accuracy and stability for EFCIL [Hayes and Kanan, 2020; Janson et al., 2022]. Our main finding is that the initial training strategy is the dominant factor influencing the average incremental accuracy, but that the choice of CIL algorithm is more important in preventing forgetting. Beyond the fact that there is no silver bullet approach to dealing with EFCIL, our in-depth statistical study quantifies the effect of different components of EFCIL approaches and thus enables informed decisions when designing new methods or implementing EFCIL in practice.

The doctor said slowly, "All the same, the mask's bound to slip once in a while."
Nurse Hopkins had bustled into the bathroom. Elinor said, raising her delicate eyebrows and looking full at him, "The mask?"
Dr. Lord said, "The human face is, after all, nothing more nor less than a mask."

— Agatha Christie, *Sad Cypress*

5 Mitigating Biases in Face Verification Systems

Contents

5.1 Introduction

Face recognition and verification technologies (FRT and FVT) have seen significant advancements in recent years, with applications ranging from security and surveillance to personal device authentication[Ho et al., 2020a; Selwyn et al., 2023; Van Noorden, 2020]. This widespread adoption of face recognition models has also raised concerns about fairness and potential biases in these systems [Buolamwini and Gebru, 2018; Kärkkäinen and Joo, 2019]. Studies have shown that FRT and FVT can exhibit disparities in performance across different demographic groups, particularly for gender, ethnicity, and age [Sarridis et al., 2023b; Wang et al., 2019a].

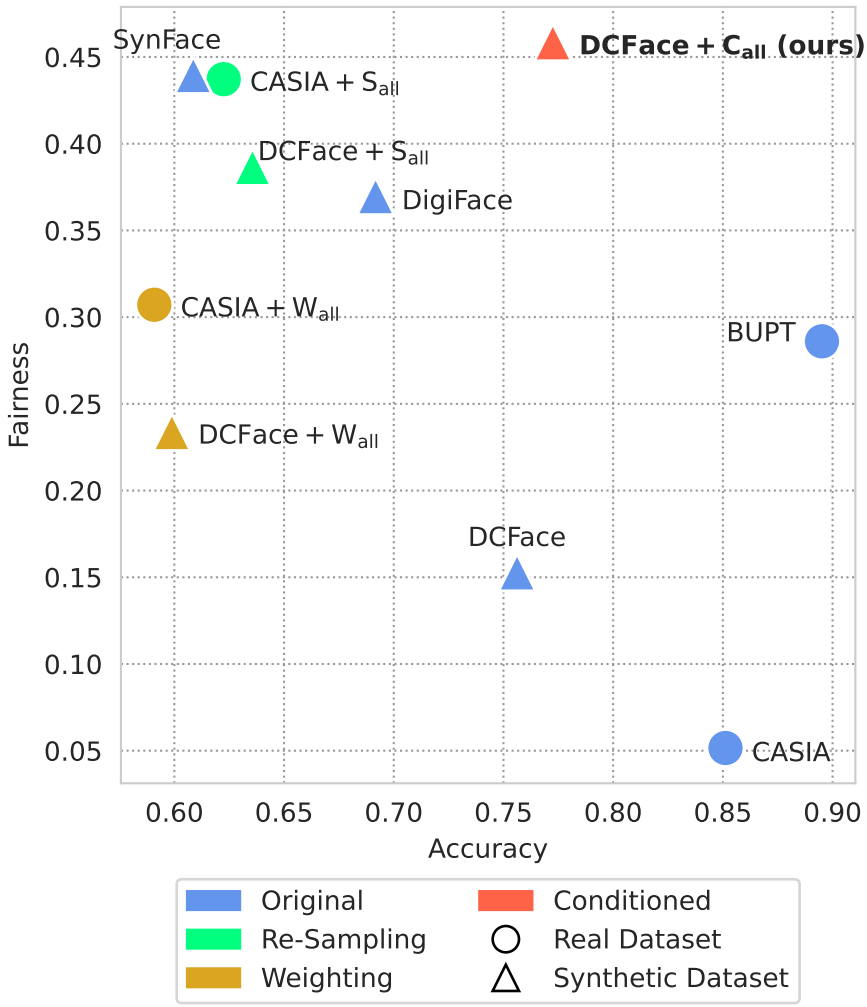


Figure 5.1: Comparison of the face verification fairness (equalized odds ratio) and micro-average accuracy metrics for models trained with real and synthetic images on the RFW dataset [Wang et al., 2019a]. The proposed pipeline improves the generation fairness and accuracy.

To address these fairness challenges, researchers have explored various approaches, including the development of demographically diversified datasets [Grother et al., 2019; Wang et al., 2019a], and debiasing methods [Robinson et al., 2020; Yang et al., 2021]. In parallel, synthetic datasets, generated using computer graphics techniques [Bae et al., 2023; Wood et al., 2021] and generative AI models [Zhao et al., 2017; D'Incà et al., 2024; Bae et al., 2023; Kim et al., 2023], offer the potential to mitigate privacy and copyright issues [Harvey, 2021] associated with real datasets [Cao et al., 2018; Kemelmacher-Shlizerman et al., 2016; Guo et al., 2016].

Nonetheless, the effectiveness of synthetic datasets in improving fairness remains an open question. While existing studies highlighted the potential for generated data to reproduce or even exacerbate the biases present in real datasets [Perera and Patel, 2023], most recent works still do not sufficiently analyze the fairness impact of models trained with their synthetic images [Qiu et al., 2021a; Bae et al., 2023; Kim et al., 2023], despite encouraging initiatives [Deandres-Tame et al., 2024; Neto et al., 2023]. Our first contribution, therefore, introduces a new generation control component based on the existing DCFace pipeline [Kim et al., 2023]. The resulting approach increases the diversity of sensitive attributes such as `gender`, `ethnicity`, and `age`, and also varies the `pose`, resulting in two new synthetic datasets DCFace_{eg} , and

DCFace$_{all}$. We compare models trained on these proposed datasets with models trained using existing generation datasets, with or without bias mitigation techniques applied.

We employ a range of common metrics to measure fairness. Still, we find them insufficient for an in-depth analysis of the origins of the biases since they do not decorrelate the impacts of the considered attributes. We consequently introduce, as a second contribution, a new analysis approach based on logit regression models that unveils the impact of individual attributes. Furthermore, we use an Analysis of Variance (ANOVA) to examine the relation between attributes and distance in the models' latent space.

As highlighted in Figure 5.1, our results demonstrate that the proposed controlled generation approach significantly improves fairness metrics while maintaining accuracy. This is a major advancement compared to more straightforward methods, since usually, decreasing variability in performance comes at the cost of also decreasing the mean performance.

The logit regression and ANOVA analyses draw coherent conclusions and reveal the effectiveness of the proposed controlled generation method in reducing attribute-based biases in both the model predictions and the latent space representations.

Main Findings

We find that:

1. Controlled generation of synthetic datasets (DCFace$_{eg}$ and DCFace$_{all}$) significantly improves fairness in face verification while maintaining competitive accuracy
2. Proposed balanced datasets outperform existing real and synthetic datasets in most fairness metrics across different verification datasets
3. Novel statistical analysis using logit regression and ANOVA effectively quantifies and interprets biases in face recognition outcomes
4. Controlled generation method reduces attribute-based biases in both model predictions and latent space representations more effectively than traditional bias mitigation techniques
5. A significant gap between synthetic and real datasets still persists, both in terms of raw performance and fairness.

However, the analysis reveals a persistent disparity in fairness across all considered approaches, particularly penalizing the `African` and `Indian` subgroups. This highlights the need for continued research and development of more robust bias mitigation strategies in face verification systems.

The code and datasets introduced here will be released to facilitate the adoption of fairness in FRT and FVT at `https://github.com/afm215/TowardFairerFaceRecognitionSets`.

5.2 Related Work

Face verification is a classical yet still open research topic. Following [Robinson et al., 2020], a model is trained to perform face recognition. Then, given a pair of images, the evaluation task is determining whether they belong to the same identity using the trained model as an embedding extractor. A threshold is optimized to separate and predict the positive and negative pairs. Following [Phillips et al., 2012; Popescu et al., 2022; Wang et al., 2019a], we advocate for selecting hard negative images to make verification more realistic and consider datasets including difficult negatives to evaluate the models' performance. We also advocate for more efforts to

integrate fairness in the verification evaluation process. Fairness evaluation can be improved by designing demographically-diversified verification datasets [Grother et al., 2019; Popescu et al., 2022; Wang et al., 2019a] and integrating demographic metadata in them [Sarridis et al., 2023b]. Demographic attributes balance deserves particular attention because it is required for analyzing potentially serious discrimination [Sarridis et al., 2023b; Ho et al., 2020a].

Real training datasets for face recognition are usually created by scraping a large number of images from publicly available sources [Kemelmacher-Shlizerman et al., 2016; Schroff et al., 2015] and then cleaning them [Cao et al., 2018; Guo et al., 2016; Yi et al., 2014] to reduce the number of unrepresentative samples. However, these datasets face several challenges. First, obtaining subjects' consent at scale is impossible, posing a serious legal challenge when collecting sensitive data such as identified faces. Second, most datasets [Cao et al., 2018; Guo et al., 2016; Yi et al., 2014] include copyrighted photos, raising licensing issues. The lawfulness of distributing copyrighted content is a longstanding discussion that applies to other computer vision tasks [Quang, 2021] and was recently revived by the success of foundation models trained with very large datasets [Scao et al., 2022]. Third, existing large datasets exhibit demographic (gender, ethnicity, age) [Popescu et al., 2022; Sarridis et al., 2023b; Wang et al., 2019a], face characteristics (size, make-up, hairstyle) [Albiero et al., 2020, 2021; Terhörst et al., 2021], and visual biases [Zhao et al., 2018], mostly reflecting the sampling bias affecting images datasets [Fabbrizzi et al., 2022]. These biases affect underrepresented segments [Buolamwini and Gebru, 2018; Kärkkäinen and Joo, 2019; Sarridis et al., 2023b] and should be addressed to improve fairness. These problems make the sustainable publication of real datasets very complicated, as proven by the withdrawal of most resources [Cao et al., 2018; Guo et al., 2016; Kemelmacher-Shlizerman et al., 2016] following public pressure [Van Noorden, 2020].

Examples of Real Training Datasets for Face Recognition

1. **VGGFace2** [Cao et al., 2018]
   - Contains over 3.3 million face images
   - Covers a wide range of pose, age, and ethnicity
   - Faces challenges with copyright and consent issues
2. **MS-Celeb-1M** [Guo et al., 2016]
   - Originally contained about 10 million images of 100,000 celebrities
   - Widely used but later withdrawn due to privacy concerns
   - Exhibits demographic biases, particularly in ethnicity and gender representation

Both datasets, while valuable for research, highlight the ethical and legal challenges associated with real training datasets in face recognition.

Synthetic datasets have the potential to reduce or remove privacy, copyright, and unfairness issues compared to real datasets [Kim et al., 2023; Deandres-Tame et al., 2024; Neto et al., 2023]. Computer graphics techniques are used in [Bae et al., 2023; Wood et al., 2021] to render diversified face images, and strong augmentations are added to increase accuracy. Most works rely on generative AI, with [Zhao et al., 2017] being an early example that uses dual-agent GANs to generate photorealistic faces. The authors of [Qiu et al., 2021a] identify the lack of variability of generated images as a central challenge and propose identity and domain mixup to improve synthetic datasets. Diffusion models were used very recently [Kim et al., 2023] to create identities and to diversify their samples based on a style bank. Synthetic datasets have the advantage of including fictitious identities, alleviating privacy and copyright issues associated with real face datasets. However, privacy issues can remain regarding data

replication in GANs [Feng et al., 2021] and diffusion models [Somepalli et al., 2023] but can be controlled and mitigated as shown in [Barattin et al., 2023]. When uncontrolled, synthetic datasets are also likely to reproduce and even exacerbate the biases of real datasets in a constrained evaluation setting [Perera and Patel, 2023].

Examples of Synthetic Datasets for Face Recognition

1. **DigiFace** [Bae et al., 2023]
   - Uses computer graphics techniques to render diversified face images
   - Employs strong augmentations to increase accuracy
   - Aims to reduce privacy and copyright issues associated with real datasets
2. **DCFace** [Kim et al., 2023]
   - Utilizes diffusion models to create identities and diversify samples
   - Based on a style bank for improved variability
   - Generates fictitious identities, addressing privacy concerns

While these synthetic approaches offer advantages in privacy and diversity, they may still face challenges such as potential data replication issues and the risk of reproducing biases present in real datasets if not carefully controlled.

Debiasing methods have been proposed to mitigate biases in face verification. One approach is to adapt the verification process to demographic segments. The authors of [Robinson et al., 2020; Terhörst et al., 2020] propose adaptive threshold-based approaches to improve fairness. Another approach is to address ethnicity-related bias by learning disparate margins per demographic segment in the representation space [Yang et al., 2021; Wang et al., 2021; Wang and Deng, 2020] or by suppressing attribute-related information in the model [Sarridis et al., 2023a]. While technically interesting, these methods are ethically and legally problematic in practice since they assume disparate treatment of human subjects by AI-based systems. We advocate for bias mitigation directly within model training sets, which we show to have a very concrete consequence on model biases.

Comparison of Bias Mitigation Approaches

The multiple bias mitigation approaches have pros and cons:

- **Real Datasets:**
  - \+ High-quality, diverse data
  - \- Privacy and copyright issues
  - \- Inherent demographic biases
- **Synthetic Datasets:**
  - \+ Addresses privacy and copyright concerns
  - \+ Potential for controlled attribute balancing
  - \- May still reproduce biases if not carefully designed
- **Debiasing Methods:**
  - \+ Can improve fairness metrics
  - \- May introduce ethical issues (disparate treatment)
  - \- Often applied post-hoc, not addressing root causes

5.3 Methodology

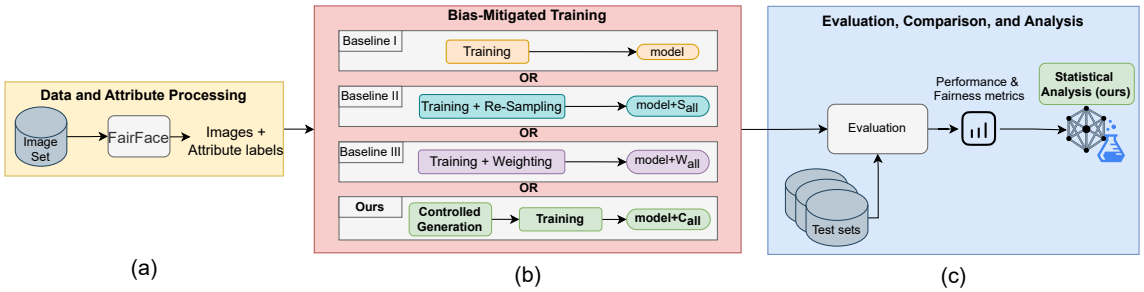


Figure 5.2: Global pipeline overview for training and evaluating models with the baselines and our proposed generative approach. Critical attributes are collected on image sets (a) that enable using bias mitigation techniques before or during model training (b). Models are then evaluated on FRV evaluation sets (c), and their biases are then analyzed using fairness metrics and our proposed statistical analysis. Contributions of this paper are colored green.

The overall training and evaluation pipeline (Figure 5.2) comprises three parts: Part (a) regroups training sets and their attributes. These training sets can be or cannot be combined with bias mitigation techniques to train models (part (b)). These techniques include our proposed controlled data generation (in green). Finally, as explained in section 5.2, these models are used in part (c) to perform FVT using the setup of Robinson et al. [2020]; Huang and Learned-Miller [2014]. The results obtained on FAVCI2D [Popescu et al., 2022], RFW [Conti et al., 2022], and BFW [Robinson et al., 2023] are analyzed in terms of raw performance (accuracy), fairness metrics, and using the statistical approach we introduce in this paper.

Following recent face recognition work [Bae et al., 2023; Kim et al., 2023], we train models using a ResNet50 architecture [He et al., 2016b] with a loss designed specifically for this task [Kim et al., 2022].

We create face recognition models with different training sets. We ensure comparability between these training sets by using the same structure and similar size, compatible with previous studies [Bae et al., 2023; Qiu et al., 2021a; Yi et al., 2014]. They contain 10,000 unique identities and 50 samples per identity.

5.3.1 Considered Biases

We balance the created datasets for four attributes: `ethnicity`, `gender`, `age`, and `pose`. The first three are sensitive attributes contributing directly to demographic fairness and are usually employed in the literature [Sarridis et al., 2023b; Albiero and Bowyer, 2020; Robinson et al., 2023; Yucer et al., 2022]. The fourth ensures face appearance variability and augments model performance. `ethnicity` and `gender` are attributes associated with each identity. When unavailable in the datasets' metadata, these attributes are inferred using FairFace [Kärkkäinen and Joo, 2019]. In this case, `ethnicity` and `gender` are categorical (Asian, Black, Indian, White) and binary variable (female/male). Since they are supposed to be consistent across the images of the same identity, we mitigate the potential inference errors by averaging the FairFace outputs per identity. Age is also inferred at the image level using FairFace.

The `pose` attribute is extracted using the model introduced in Hempel et al. [2022]. We use face rotation around the pitch, the yaw, and the roll axes (i.e., the rotations around the x, y, and z axes) to characterize `pose`.

5.3.2 Proposed Balanced Dataset Generation

Our controlled approach relies on the DCFace [Kim et al., 2023] generation pipeline. It applies the style of a real picture (style image) to a synthetic face picture (Id image) using a dual-conditioned diffusion model DCFace combines a single ID with several style images to produce the samples representing each synthetic identity in the training set. The identity-level attributes (`ethnicity` and `gender`) are, therefore, controlled by the choice of the ID image. The picture-level attributes (`age` and `pose`) are controlled by the choice of the style images.

We thus introduce a joint diversification process on `gender`, `ethnicity`, `age`, and `pose` attributes illustrated in Figure 5.3.

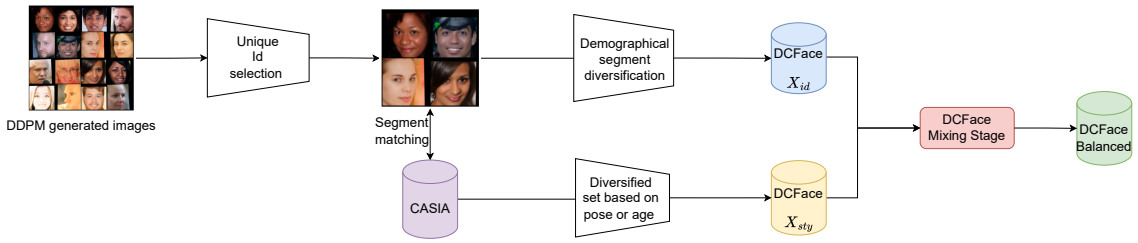


Figure 5.3: Detailed view of our controlled generation method

We select a list of ID images generated with DDPM [Ho et al., 2020b] whose joint `gender` × `ethnicity` distribution is perfectly balanced. We diversify `pose` and `age` by iteratively populating the less-represented age and pose categories of each identity. We also match the demographic segment (`gender`×`ethnicity`) of ID and style images to facilitate the loss convergence process. We implemented this matching following initial tests, which showed that convergence is not guaranteed without anything else. We create two versions of the balanced dataset to assess the influence of identity-level and image-level attributes. **DCFace**$_{eg}$ uses only `gender`×`ethnicity` , **DCFace**$_{all}$ considers all four attributes.

5.3.3 Training Set Baselines

We compare **DCFace**$_{eg}$ and **DCFace**$_{all}$ with a representative set of real and synthetic datasets:

- **CASIA** [Yi et al., 2014], a real dataset representing celebrities from the IMDB dataset.
- **BUPT** [Wang et al., 2021], areal dataset that is balanced for `ethnicity`. Note that the full version includes more than 1M images. We subsample BUPT to match the structure of other baselines.
- **SynFace** [Qiu et al., 2021a], a synthetic dataset created with a GAN architecture using identity and domain mixup to diversify generated faces.
- **DigiFace** [Bae et al., 2023], a synthetic dataset created using rendering technique to obtain diversified representations of faces of each identity.
- **DCFace** [Kim et al., 2023], a synthetic dataset generated using the default uncontrolled pipeline of Kim et al. [2023].

5.3.4 Dataset Biases Analysis

Table 5.1 presents the imbalance degree of each attribute for the tested training sets. We report the attribute diversity a for a dataset $\mathcal{D}$ computed as the normalized entropy applied on the frequency p_{a_i} for the attribute sub-groups $a_{i\in[1,m]}$.

$$Diversity_a(\mathcal{D}) = -\frac{1}{\log(N)} \sum_{i=0}^{N} p_{a_i} \log(p_{a_i}) \tag{5.1}$$

Table 5.1 enables a data-oriented comparison of our datasets and baselines. It highlights the proposed pipeline's effectiveness and the need for joint attribute balancing to avoid unwanted side effects. For instance, only balancing on ethnicity and gender reduces age diversity and does not affect pose, while balancing for all attributes results in a better global trade-off.

Attribute	CASIA	BUPT	DigiFace	SynFace	DCFace	**DCFace$_{eg}$**	**DCFace$_{all}$**
Gender	1.00	0.93	0.93	0.99	0.99	1.00	1.00
Ethnicity	0.47	0.92	0.65	0.40	0.56	0.93	0.90
Age	0.59	0.71	0.42	0.64	0.64	0.61	0.69
Pose	0.61	0.57	0.67	0.58	0.51	0.51	0.58

Table 5.1: Inferred diversity for the compared training datasets. The degree of balance is quantified by the entropy for the considered attributes across the dataset. **Datasets introduced in this paper are shown in bold**.

5.3.5 Baseline Debiasing Methods

We compare the proposed dataset bias mitigation pipeline with two classical baseline methods: resampling and loss weighting.

5.3.5.1 Re-sampling

Data re-sampling balances class distribution within training data by employing strategies other than the default uniform sampling. These strategies can consist of over-sampling the data from the under-represented classes and/or under-sampling majority classes [Tantithamthavorn et al., 2018; Idrissi et al., 2022].

Oversampling [Bennin et al., 2018; Amin et al., 2016; Last et al., 2017; Zheng et al., 2015] increases the number of samples by replicating existing data. However, duplicating data by sampling the several times can lead to over-fitting. On tabular data, interpolating techniques such as SMOTE and its variants [Chawla et al., 2002; Han et al., 2005; Bunkhumpornpat et al., 2009] can be used in order to tackle this overfitting issue. Still, such approaches are not trivial and more costly for non-tabular data such as images.

Undersampling, on the other hand, consists in the reduction of the majority classes so that their representativity matches the underrepresented classes [Liu et al., 2006; Tsai et al., 2019; Lehmann and Ebner, 2022]. The main drawback of such an approach is that it results in unused data, which is not an optimal setup.

Here we use Re-Sampling as a baseline for bias mitigation by combining over-sampling and under-sampling. Specifically, for each attribute a with values a_j, we count n_j, the number of images with value a_j. We then assign a weight $w_j = 1/n_j$ to each image sharing value a_j. For each image x_i, we compute its weight w_i as the product of the weights of all attributes associated with the image. The sampling probability for each image is calculated as $p_i = w_i / \sum_k w_k$. At each

beginning of a training epoch, we sample N images according to the probability distribution $\{p_i\}$, where N is the size of the original dataset.

Note that this approach, coupled with the set of random image augmentations used during training, should mitigate to a certain extent the mentioned limitations of both over-sampling and under-sampling.

5.3.5.2 Loss Weighting

Loss weighting tries to adapt the loss scale depending of the characteristics of the sample. This weighting can be linked to the difficulty of the sample as done implicitly by the Adaface Loss [Kim et al., 2022], which can be induced by the class imbalance or in our use case, by the corresponding attributes representativity. A common way to weight the loss is to use the same weights computed in subsection 5.3.5.1, i.e. using the invert of the frequency/count [Fernando and Tsokos, 2022; Wang et al., 2017b; Huang et al., 2016]. We thus use the same weights w_i for weighting the loss. The weights are normalized batch-wise to ensure the same order of gradient amplitude. The loss of the batch is defined as:

$$\mathcal{L}(x_1, ..., x_K) = \frac{\sum_k w_k \mathcal{L}(x_k)}{\sum_k w_k} \tag{5.2}$$

where $\mathcal{L}(x_k)$ is the sample-wise loss for image x_i.

We add $+\mathrm{S}_{ge}$ and $+\mathrm{W}_{ge}$ to initial dataset names for resampling and loss weighting limited to `gender` and `ethnicity`. We add $+\mathrm{S}_{all}$ and $+\mathrm{W}_{all}$ when all attributes are debiased.

5.4 Toward a Fairer Analysis of FVT evaluation

5.4.1 Evaluation Sets and Protocol

We use RFW [Conti et al., 2022], FAVCI2D [Popescu et al., 2022], and BFW [Robinson et al., 2020] in our fairness analysis. We selected these face verification datasets because they have sufficient identities per demographic segment to enable a rigorous analysis. Similar to training datasets, we extract FairFace attributes whenever they are not provided. For RFW, we use the included `ethnicity` attribute since the dataset is already balanced for it. Figure 5.4 presents a brief description of the pair attributes in the RFW, FAVCI2D, and BFW datasets. While the three datasets have similar balancing on `age` and `pose` attributes, they exhibit different characteristics in terms of gender and ethnicity distributions. FAVCI2D has a relatively balanced gender distribution but a skewed ethnicity distribution, with the `White` ethnicity being the most prevalent. In contrast, RFW has a more balanced representation of ethnicity, with a uniform distribution across `African`, `Indian`, `Asian`, and `Caucasian` ethnicities, but is unbalanced in terms of `gender`. These differences allow for a comprehensive evaluation of face verification models' fairness and performance across diverse demographic groups, assessing how well the models handle variations in gender and ethnicity representation and identifying potential biases arising from imbalanced training data.

5.4.2 Fairness and Performance Metrics

We employ a set of commonly used and complementary metrics to comprehensively evaluate face recognition fairness and performance.

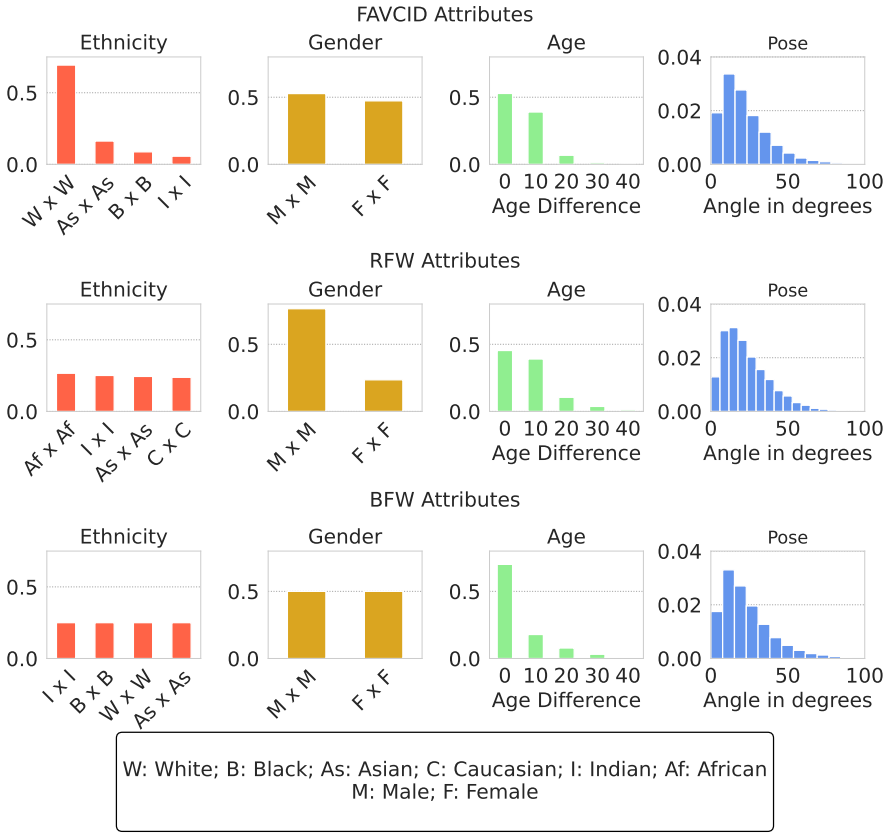


Figure 5.4: Attribute analysis of the evaluation datasets. Attributes are generated using FairFace [Kärkkäinen and Joo, 2019], except for the `ethnicity` of RFW, included with the dataset.

5.4.2.1 Performance Metrics

Micro-Average Accuracy[Raji and Fried, 2021] is a commonly used metric for evaluating the overall performance of a face recognition model. Micro-average accuracy is particularly useful when dealing with unbalanced data, as it gives equal weight to each dataset segment, regardless of the group size. Consequently, the overall accuracy is not biased toward the majority group.

True Match Rate (TMR)[1]**, or TPR**, measures the proportion of actual positive cases that are correctly identified. **False Match Rate (FMR), or FPR**, measures the proportion of negative cases incorrectly identified as positive by the face recognition model. We follow existing face recognition literature [Ho et al., 2020a; Van Noorden, 2020] and consider FMR as a more critical metric compared to TMR.

5.4.2.2 Fairness Metrics

Degree of Bias (DoB) [Gong et al., 2020] is the standard deviation of accuracy across different subgroups, that is higher when the performances varies a lot w.r.t each subgroup. However, datasets with low accuracy tend to inherently have a smaller overall variance. Moreover, DoB does not allow for fine-grained error analysis, which is central to understanding performance variations in our case.

Demographic Parity Difference (DPD) and **Demographic Parity Ratio (DPR)** [Agarwal et al., 2018, 2019] require that the probability for individuals to receive a positive outcome should be the same across all demographic groups. DPD is the absolute difference

[1] is equivalent to 1−FNMR

between the highest and lowest probability across all subgroups, whereas DPR is the ratio between the lowest and highest. The closer the DPD is to zero and the closer the DPR is to one, the fairer the results are.

Equalized Odds Difference (EOD) and **Equalized Odds Ratio (EOR)** [Agarwal et al., 2018] require that the face recognition model's TMR and FMR are independent of the demographic groups, thus ensuring consistent accuracy across groups. EOD is calculated as the maximum absolute difference between the TMRs or FMRs across groups. EOR is the minimum between the ratio of the TMRs and FMRs across groups. The closer the EOD is to zero and the closer the EOR, the fairer the results are.

5.4.3 Proposed Statistical Analysis Approach

We use the logit statistical framework described in Chapter 3. For each pair, we create variables based on the attributes of the 2 identities of the pair: since our goal is to measure the impact in TMR and FMR, we filter the pairs of each dataset to only keep pairs where the ethnicity and gender of both identities are the same. This allows us to assign an `ethnicity` and `gender` variable to the pairs, not only to the identities. We define the `age` variable as the absolute difference in age between the 2 identities. The `pose` variable of a pair is a scalar describing the angle between the 2 pose vectors of the identities.

For each test image pair, we have a binary target variable 0/1 (the pair was either correctly or incorrectly classified). This setup makes it natural to use logit regression [Angrist and Pischke, 2009]. Additionally, since the response variable is function of a distance threshold in the latent space, it seems natural to study the impact of the variables on the pair distances in the latent space. This impact is accessed using Analysis of Variance (ANOVA) [Gareth et al., 2013].

The two statistical methods provide complementary insights into the impact of the studied attributes on the fairness metrics. Section 5.5 presents the detailed application of these methods to the datasets and fairness metrics, along with the interpretation of the results.

Summary of variables

We consider the following variables for each pair:

- `ethnicity` : a categorical variable representing the ethnicity or geographical origin of both individuals of the pair;
- `gender` : a categorical binary variable representing the gender of both individuals of the pair;
- `age` : a scalar variable representing the age difference in the pair;
- `pose` : a scalar variable representing the angle between the pose of the two faces in the pair;
- $\hat{y}$: a binary variable representing the correct and incorrect identification of the pair.

`ethnicity` and `gender` are the sensible studies attributes, while `age` and `pose` are control variables. $\hat{y}$ is the endogenous variable.

5.5 Results and Analysis

We report here performance on multiple test sets and discuss fairness metrics on both FAVCI2D , RFW, and BFW sets. Statistical and ANOVA analysis is performed on RFW and is reported for FAVCI2D in the appendix.

5.5.1 Raw performance on test sets

Verif. dataset	Real dataset		Synthetic datasets				
	CASIA	BUPT	SynFace	DigiFace	DCFace	**DCFace$_{eg}$**	**DCFace$_{all}$**
LFW	99.46	**99.55**	87.28	94.88	98.13	98.24	**98.25**
CFP-FP	**94.87**	90.03	67.01	**83.4**	80.92	80.03	81.28
CPLFW	90.35	85.98	64.91	76.61	79.94	79.32	**80.17**
AgeDB	94.95	94.3	61.78	78.26	**87.96**	86.77	86.53
CALFW	93.78	94.38	73.53	79.78	90.39	**90.6**	90.03
RFW	86.38	**90.35**	64.3	72.73	76.95	78.51	**79.5**
FAVCI2D	**82.77**	81.81	61.19	67.17	72.84	73.31	**73.73**
BFW	89.3	**92.48**	70.08	77.27	84.47	85.45	**88.53**
AVG	91.48	91.11	68.76	78.76	83.95	84.03	**84.75**

Table 5.2: Raw Accuracy obtained for the different used sets on 8 datasets, including five commonly used datasets in addition to BFW, RFW, and FAVCI2D .

In addition to FAVCI2D , BFW, and RFW, we report in Table 5.2 the raw accuracy results on 5 common evaluation sets used in prior work on the FR task [Bae et al., 2023; Kim et al., 2022, 2023; Qiu et al., 2021a] :

- Labeled Faces in the Wild (LFW) [Huang and Learned-Miller, 2014], the reference dataset for the task;
- CALFW [Zheng et al., 2017], a version of LFWwith a larger age variability;
- CPLFW [Zheng and Deng, 2018], a version of LFWwith pose variability;
- AgeDB [Moschoglou et al., 2017], a dataset designed for maximizing age variability;
- CFP-FP [Sengupta et al., 2016], that is designed for pose variability.

Raw accuracy differs from the micro accuracy reported on the paper. Micro accuracy gives the same importance to each demographic segment, whereas raw accuracy performs a simple mean across all images, without any distinction.

Table 5.2 confirms the performance gain of DCFace$_{all}$ over the original generation pipeline: The generation pipeline slightly improves accuracy for four of these datasets (+0.12, +0.36, +0.23, and +0.89 for LFW, CFP-FP, CPLFW, and FAVCI2D) and slightly degrades performance for the other two (-1.43 and -0.36 points for Age-DB and CALFW). On the balanced sets, (i.e. RFW and BFW) the pipeline induces important gains in accuracy (+2.55 for RFW and +4.06 for BFW).

5.5.2 Performance & Fairness Comparison

Table 5.3 presents the fairness metrics and micro-average accuracy for all training approaches on RFW, FAVCI2D, and BFW. These metrics are reported for real and synthetic datasets separately.

Among the real datasets, the model trained on BUPT achieves the highest accuracy on both RFW (0.895) and FAVCI2D (0.818) compared to models trained on CASIA. As expected, BUPT also gets the best fairness metrics on FAVCI2D, but surprisingly, on RFW it shows a mitigated behavior, being first in terms of EOD only. Overall on RFW, CASIA+S$_{ge}$ shows the best behavior in terms of fairness (DPR, DPD, EOD), at the cost of 5.2 points of accuracy compared to the original CASIA set. This surprising behavior is not noticed with our in-depth analysis (especially in Figure 5.5), which draws other conclusions for BUPT model sensitivity, advocating for the usefulness of our analysis approach. Regarding accuracy though, the difference between

	RFW [Wang et al., 2019a]						FAVCI2D [Popescu et al., 2022]						BFW [Robinson et al., 2020]					
	DoB↓	DPD↓	EOD↓	DPR↑	EOR↑	Acc↑	DoB↓	DPD↓	EOD↓	DPR↑	EOR↑	Acc↑	DoB↓	DPD↓	EOD↓	DPR↑	EOR↑	Acc↑
BUPT	**30.3**	23.6	**11.9**	68.4	28.6	**89.5**	**38.4**	**13.0**	**14.4**	**75.2**	19.3	**81.8**	**25.7**	**5.5**	**12.5**	**88.8**	**26.1**	**92.6**
CASIA	35.3	19.0	22.0	71.1	5.2	85.1	39.0	21.2	28.5	66.3	16.5	81.1	29.1	9.2	14.9	82.2	1.3	90.3
CASIA + S_{eg}	39.4	**11.5**	18.8	**80.4**	29.4	79.9	43.1	15.5	18.7	70.6	31.6	75.1	31.8	10.0	18.9	80.4	11.3	88.0
CASIA + S_{all}	48.2	17.8	24.0	68.8	**43.7**	62.3	48.6	22.0	24.1	60.3	**43.8**	61.8	43.5	19.3	33.6	67.8	23.1	74.0
CASIA + W_{eg}	43.5	17.3	22.8	70.2	23.8	74.0	45.3	22.3	21.0	59.2	32.7	71.1	35.4	12.5	18.5	77.0	18.5	84.9
CASIA + W_{all}	49.1	26.7	36.2	54.7	30.7	59.1	49.1	28.1	31.2	47.1	31.0	59.4	46.4	29.6	38.5	52.4	23.5	68.2
SynFace	48.6	13.8	24.9	73.6	44.0	60.9	48.5	22.7	26.4	57.3	37.4	62.0	45.4	20.4	23.2	63.3	36.4	70.7
DigiFace	45.9	15.5	25.6	73.6	37.0	69.2	47.3	21.0	22.2	62.1	40.4	66.0	45.7	16.0	21.1	70.1	**44.8**	70.0
DCFace	42.7	17.2	32.7	71.4	15.3	75.6	45.1	20.0	18.9	62.8	32.1	71.6	35.4	14.2	21.5	74.4	11.7	85.0
DCFace + S_{eg}	44.0	13.7	36.7	76.5	18.2	72.3	45.9	15.5	21.2	68.4	31.5	69.5	37.2	18.6	29.7	68.3	10.1	82.9
DCFace + S_{all}	48.0	16.7	23.8	69.5	38.7	63.6	48.1	22.1	23.0	58.2	43.8	63.4	42.9	16.8	25.8	68.7	21.8	75.3
DCFace + W_{eg}	44.2	16.7	33.4	70.7	18.9	72.7	46.0	19.2	20.9	62.1	29.9	69.4	36.9	14.6	20.1	72.3	12.5	83.5
DCFace + W_{all}	49.0	19.4	31.6	59.9	23.4	59.9	48.5	23.7	26.0	54.9	36.5	61.8	44.4	24.0	24.3	56.5	27.3	72.8
DCFace + C_{eg}	42.2	12.7	**13.7**	77.1	41.2	76.4	44.7	14.3	15.6	**71.1**	**66.0**	72.4	34.7	**11.3**	13.8	**77.8**	23.0	85.7
DCFace + C_{all}	**41.6**	**11.2**	14.6	**80.3**	**45.9**	**77.3**	**44.5**	**14.2**	**14.9**	70.9	58.6	**72.7**	**34.2**	11.5	**13.5**	77.5	24.1	**86.1**

Table 5.3: Fairness metrics and Micro-average accuracy scores of tested datasets and bias mitigation techniques. Real and synthetic datasets are separated. Groups are defined as a combination of `gender` and `ethnicity` . DPD: Demographic Parity Difference; EOD: Equalized Odds Difference; DPR: Demographic Parity Ratio; Equalized Odds Ratio; Acc: Micro-average Accuracy. The best results for each dataset type are in **bold**, and the second-to-best results are underlined.

CASIA and BUPT becomes much higher on RFW than on FAVCI2D (4.4 points vs 0.7 points gap). This chaotic behavior on RFW might result from a domain overlap between BUPT and RFW enhancing model performance at the cost of fairness metrics.

Among synthetic datasets, the proposed DCFace$_{eg}$ and DCFace$_{all}$ show the most promising results across the evaluation sets. These balanced variants improve fairness compared to DCFace, the original generation pipeline they build upon. The fairness gains are large for DPD, EOD, DPR, and EOR, and less important for DoB. The differences between DCFace$_{all}$ and DCFace$_{eg}$ are small for most fairness metrics, but DCFace$_{all}$ provides a mild accuracy gain. The obtained results demonstrate that the proposed balancing pipeline, particularly DCFace + C_{all}, substantially improves fairness metrics across different verification datasets. Importantly, a small accuracy gain compared with the original DCFace dataset is also observed, along with fairness improvement. The models trained with balanced datasets probably benefit from a smaller shift between training and verification datasets, reflected in the micro-average accuracy measured during evaluation.

Summary of raw results

Among real datasets:

- BUPT achieves the highest accuracy on RFW and FAVCI2D
- CASIA+S_{ge} shows best fairness on RFW, but at a cost to accuracy

Among synthetic datasets:

- DCFace$_{eg}$ and DCFace$_{all}$ show the most promising results
- DCFace$_{all}$ Improve fairness compared to the original DCFace, especially for DPD, EOD, DPR, and EOR
- DCFace$_{all}$ provides mild accuracy gain over DCFace$_{eg}$

Overall, proposed balancing pipeline (DCFace + C_{all}) substantially improves fairness metrics while maintaining or slightly improving accuracy

5.5.3 Logit Model for Bias Quantification

To quantify the biases in face recognition outcomes more precisely, we employ a logit model that estimates the impact of person attributes on face verification model predictions. Hence,

we examine the relationship between the studied attribute and the face recognition system's performance in terms of FMR and TMR. The two logit regressions are:

$$\text{(TMR)} \quad \hat{y}|y=1 \sim \sigma(\texttt{ethnicity} + \texttt{gender} + \texttt{age} + \texttt{pose}) \tag{5.3}$$

$$\text{(FMR)} \quad \hat{y}|y=0 \sim \sigma(\texttt{ethnicity} + \texttt{gender} + \texttt{age} + \texttt{pose}) \tag{5.4}$$

where $\hat{y}$ is the prediction of the model; y is the ground-truth label of the pair; σ is the sigmoid function; `ethnicity` and `gender` are categorical variables implemented with the dummy variable coding [Hardy, 1993]; `age` and `pose` are handled as continuous variables.
The logit model coefficients β_k represent the change in the log-odds of the binary outcome (e.g., false positive or true positive) for a unit change in the corresponding attribute, holding other attributes constant. The unit change is computed with respect to the unprotected group (`Caucasian` for ethnicity and `Male` for gender) which is the reference level in the dummy coding. Since the β_k are not easily interpretable by themselves, we then compute the mean marginal effects of each attribute, i.e., how much the TMR or FMR changes when we shift from the unprotected value (for instance `Male`) to a protected one (for instance `Female`). Since we control for all other variables at the same time, this effect can be interpreted as an effect with all other attributes kept constant. Therefore, the marginal effect gives an estimation of the effective demographic biases while accounting for co-founding factors.

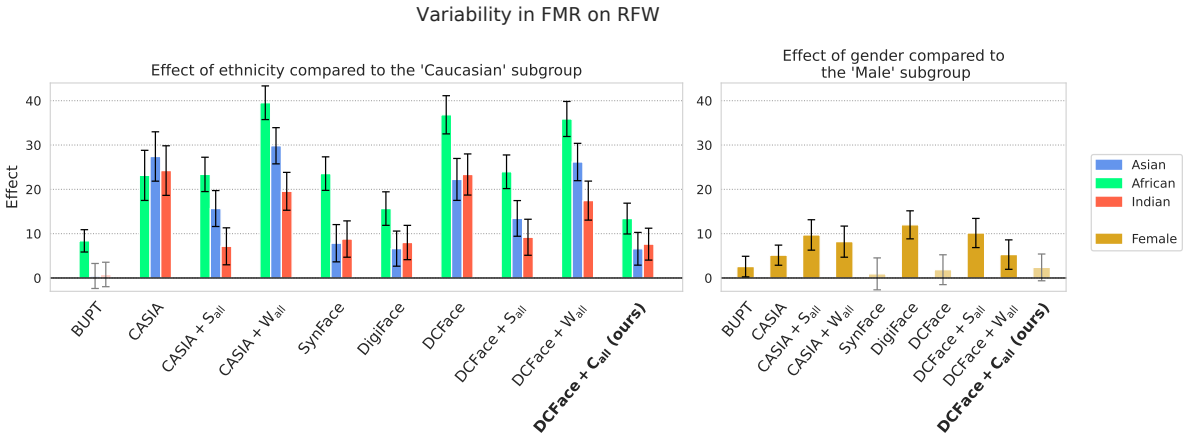


Figure 5.5: Marginal effect on FMR (smaller is better) for each method compared to the unprotected group on RFW. Example: "When using CASIA , on average and other things being equal, two people from the `African` subgroup are 22% more likely to be wrongly misidentified than two people from the `Caucasian` subgroup". Non-significant effects are shown in transparency. Our controlled generation reduces biases of DCFace more effectively than other bias mitigation techniques.

Figure 5.5 presents the logit model results for the `ethnicity` and `gender` attributes on RFW, showing the computed marginal effects on FMR. The marginal effects are calculated relative to each attribute's unprotected reference group. The higher the bar, the higher the bias against the protected subgroup. For example, when using DCFace, our analysis shows that the FMR for the `African` subgroup is 35 points higher than for the `White` subgroup, independently of the other considered attributes. The addition of re-weighting does not affect this bias, while re-sampling reduces it to 22 points. Our proposed controlled generation method further reduces it to 12 points. With regard to gender bias, we observe that despite decreasing the bias for `ethnicity`, re-sampling increases the bias for `gender`. The proposed controlled generation reduces biases for `ethnicity` while keeping the bias in `gender` non-significant.

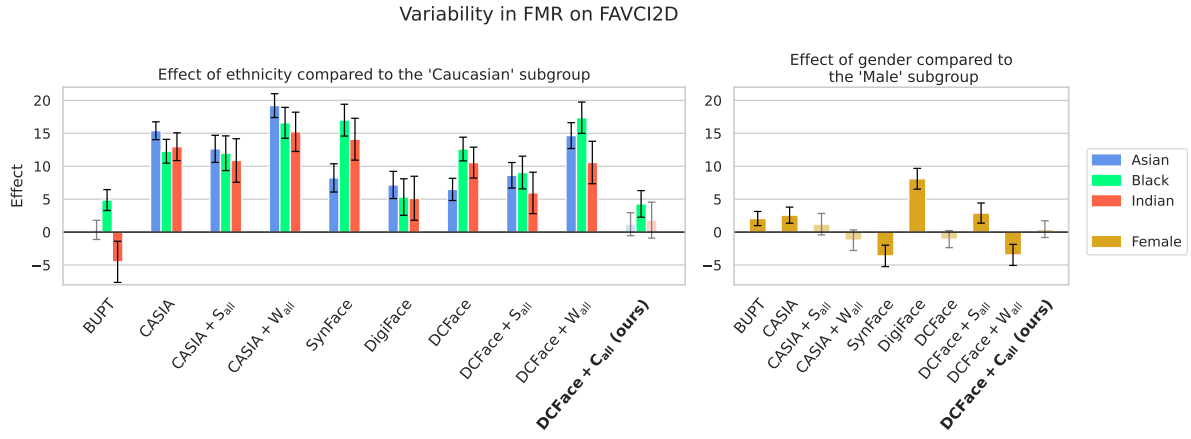


Figure 5.6: Marginal effect on FMR (smaller is better) for each method compared to the unprotected group on FAVCI2D .

Figure 5.6 presents the logit model results for the `ethnicity` and `gender` attributes on FAVCI2D ith the marginal effects on FMR. We get similar marginal effects, with our method being the less sensible to the `ethnicity` attribute, even less than BUPT. The increase of the BUPT-trained model's sensitivity with regard to the inferred labels on FAVCI2D might come from the dataset balancing done on the same labeling system as RFW

The results of the logit model on BFW, and of TMR on FAVCI2D and RFW can be found in the appendix.

The logit model results provide valuable insights into the fairness implications of different face recognition methods and datasets. By comparing the marginal effects across attributes and methods, we can identify the extent and nature of biases present in each approach. Our controlled generation method demonstrates a reduction in biases compared to the original DCFace and other bias mitigation techniques, as evidenced by the significantly smaller marginal effects. The interpretation of the logit model results highlights the disparities in face recognition performance across different attribute subgroups. These findings highlight the importance of considering fairness in the development and evaluation of face recognition systems and the need for effective bias mitigation strategies.

5.5.4 ANOVA on Latent Space

The discrete performance and fairness metrics can be seen as consequences of the variability in the distribution of feature vectors in the model's latent space. Therefore, we utilize ANOVA to investigate the influence of personal attributes on the distances in the models' latent space. In our case, the groups are defined by the person's attributes, such as gender, age, and ethnicity, while the explained variable is the distance between face representations in the latent space.

We use the sum of squares computed during ANOVA to extract the η^2 associated with each attribute. Each η^2 value represents the impact of the variable on the distance variance in the latent space. The η^2 of each attribute sum to the R^2 of the ANOVA, i.e. the total variance explained by the model.

Figure 5.7 shows the result of ANOVA on the distances in the latent space of the RFW dataset, both on the positive and negative pairs. We see that overall, the explained variance on the positive pairs is generally smaller than the explained variance of the negative pairs: this is expected since two images of different people are expected to have more variability than two images of the same person. Moreover, the total R^2 of the ANOVA goes up to

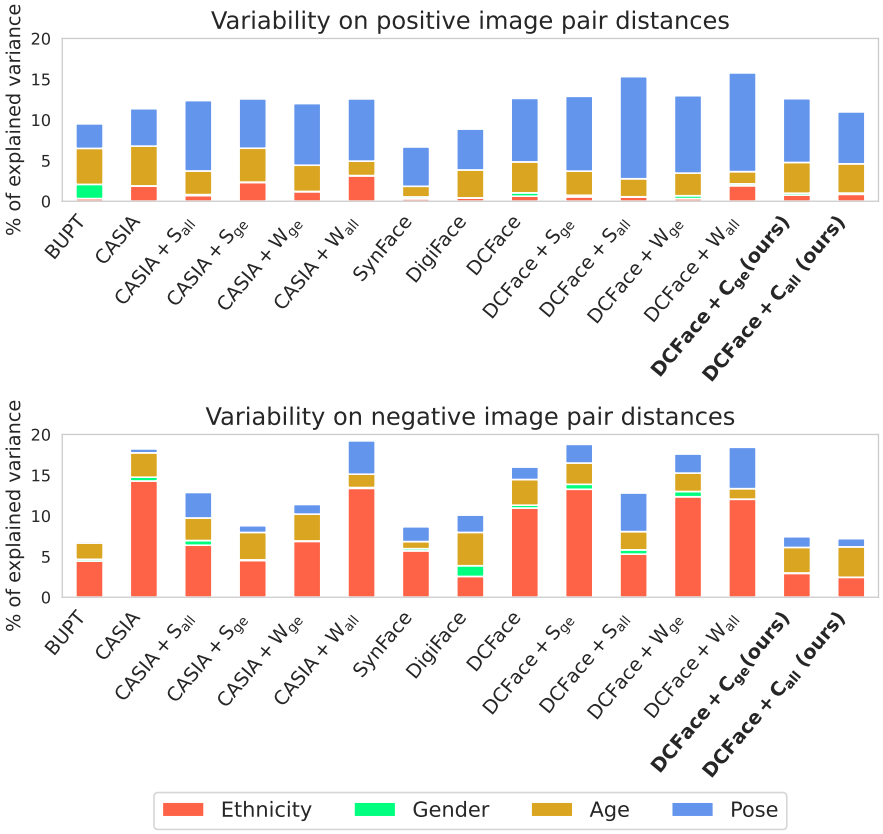


Figure 5.7: ANOVA results on RFW: total height corresponds to R^2, the explained variance by the variables. Each bar is decomposed into multiple η^2, i.e. the individual contributions to the variance.

0.18, meaning that 18% of variance in the distances in the latent space can be attributed solely to the considered people's attributes.

On the positive pairs, `pose` has the most influence: indeed, we can have a lot of variability in the pose of the same person. However, neither `ethnicity` nor `gender` plays a big role, meaning that across demographic segments, the spread of the latent vectors of a single person is very similar. This is expected since the training loss tries to bring closer the latent vectors of the same individual, who has only one `ethnicity` and `gender` value.

On the other hand, on the negative pairs, `ethnicity` is the attribute that is associated with the highest impact on the latent vectors. This means that the distances for negative pairs are much higher for some demographic segments than for others. This quantifies how much the demographic imbalance translates into the geometry of the latent space. Confirming previous works [Kärkkäinen and Joo, 2019; Sarridis et al., 2023b] with another approach, our analysis shows a significant impact of the demographic attributes on the spread of the latent vectors. Once more the impact of our approaches,DCFace$_{all}$ DCFace$_{eg}$, on the η^2 shows the effectiveness of our controlled generation. By contrast, traditional training strategies such as re-sampling and loss-weighting are not as good at mitigating the biases in the latent space.

Key Findings: Bias Mitigation with Controlled Generation

The proposed controlled generation method effectively reduces demographic biases in face recognition. It significantly lowers FMR bias for the African subgroup (from 35 to 12 points), reduces ethnicity-based disparities in latent space representations, and maintains

non-significant gender bias in recognition outcomes.

This approach outperforms traditional bias mitigation techniques, showing superior bias reduction in both recognition outcomes and latent space. It effectively addresses the 18% variance in latent space distances attributed to demographic attributes.

Importantly, the method balances fairness improvement with maintained or slightly improved accuracy across different verification datasets.

5.5.5 Model Diagnostics

We validate the regression we performed by plotting the diagnostics plots of the regressions for TMR and FMR. As explained in the last section of Chapter 3, since we use a logit regression, we use the DHARMa package Hartig [2018] in R to simulate interpretable residuals.

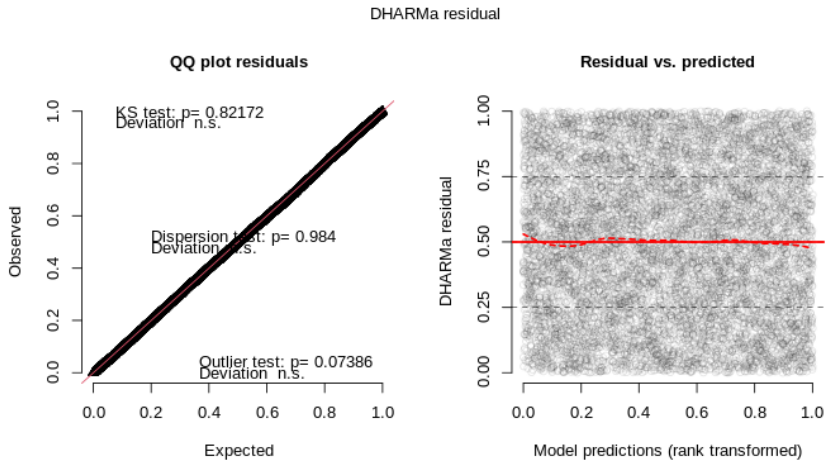


Figure 5.8: QQ-plot of residuals and Residual vs. predicted plot: logit model is adapted and log-odds are linear in the variables.

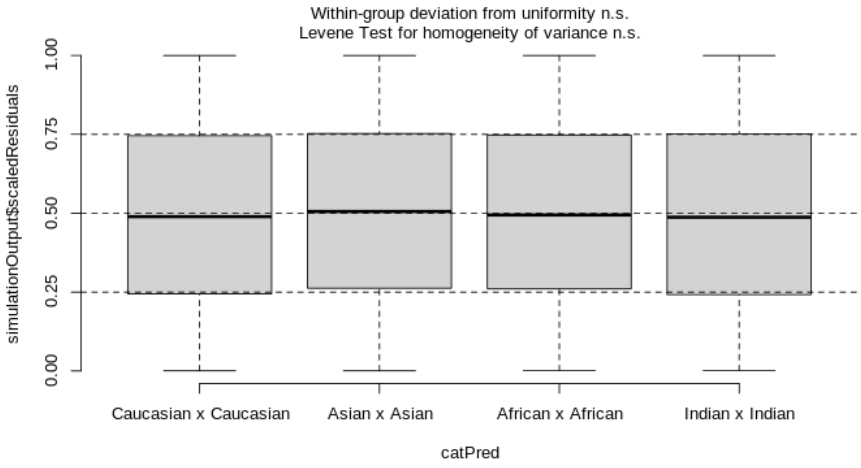


Figure 5.9: Residual vs. predictor plots: exogeneity is verified.

Figures 5.8 and 5.9 present the QQ-plot of residuals, the residuals vs. predicted plot, and the residual vs. predictor plot for the `ethnicity` variables. We see no particular specification problem. Additional plots such as overdispersion and zero-inflation can be found in the appendix.

5.6 Conclusion

We addressed the issue of fairness in FVT by evaluating the performance and bias of models trained on various real and synthetic datasets. We proposed a novel controlled generation approach to create balanced synthetic datasets, DCFace$_{eg}$ and DCFace$_{all}$, which prioritize attribute diversity. Our experiments demonstrated that models trained on these synthetic datasets significantly improved fairness metrics while maintaining competitive accuracy. Our proposed analysis based on logit regression and ANOVA revealed that our controlled generation method effectively reduces attribute-based biases in both model predictions and latent space representations. It also highlights a persistent disparity in fairness across all considered approaches that, in particular, penalizes the African sub-group.

The findings of this work have important implications for the development of fairer and more inclusive FVT systems. By demonstrating the effectiveness of attribute balancing in synthetic data generation and providing a comprehensive evaluation framework, we advocate for more efforts in addressing bias issues in computer vision applications. Future research could explore the integration of our approach with other bias mitigation techniques and investigate the generalizability of our findings to other computer vision tasks and datasets.

Men are mistaken in thinking themselves free; their opinion is made up of consciousness of their own actions, and ignorance of the causes by which they are determined.

— Spinoza, *Ethics*

6

Explaining Recommender Systems Performance with User Coherence Measures

Contents

6.1 Introduction

The increase in data generated through online interactions presents a significant challenge in presenting relevant information to users. Recommender systems (RS) address this issue by utilizing user behavior, preferences, and interaction history to generate personalized suggestions [Nilashi et al., 2013; Pavlidis, 2019]. These systems are essential in filtering and personalizing content across various domains, from e-commerce to entertainment and news consumption [Nilashi et al., 2013; Konstan, 2004].

RS can be categorized into three main types: content-based (CB), knowledge-based (KB), and in our study scope, collaborative filtering (CF) approaches [Burke, 2000]. On the one hand, content-based systems use machine learning to classify items likely to interest users based on available characteristics of previously consumed items (e.g. the genre of a film), whereas knowledge-based approaches aim to extract semantic representations in order to find products meeting user requirements [Burke, 2000; Jannach et al., 2010]. On the other hand, collaborative filtering has gained significant attention due to its ability to aggregate user preferences and make recommendations based on similarities in user behavior patterns [Konstan, 2004]. We concentrate on CF approaches because they have dominated the research landscape in recent years [Lops et al., 2019; Zhang et al., 2019; Batmaz et al., 2019] and require minimal information, analyzing only user-item interactions without the need for additional content or knowledge-based features [Ekstrand et al., 2011; Afoudi et al., 2018].

Traditionally, RS research has focused primarily on domains such as movie recommendations and e-commerce. However, to better understand the impact and effectiveness of recommender systems across various fields, it is crucial to expand our evaluation to other domains. One domain where the user experience could benefit from personalized recommendations is tourism. Points of interest (POIs) are a central part of tourist experiences, and ideally, tourists should receive personalized recommendations to discover new places that are most interesting to them. Such personalization can be achieved by leveraging user profiles that encode their tourist preferences Deldjoo et al. [2020]; Werneck et al. [2021]. However, creating rich and accurate user profiles in the tourism domain is challenging, particularly due to the sparsity of explicit user feedback and the visual nature of tourist experiences.

To address these challenges and provide a diverse dataset for evaluating recommender systems, we introduce Vis2Rec, a new visual dataset designed for POI recommendation. This dataset allows for examining recommender systems in the tourism domain and provides an opportunity to explore the use of visual data in creating user profiles and generating recommendations.

To rigorously evaluate the effectiveness of RS, the choice of adapted measures is crucial. Selecting the right measures is essential to ensure that RS delivers recommendations that are not only accurate but also align well with users' overall preferences and consumption patterns, ultimately enhancing user satisfaction. Despite the emergence of many new measures like diversity and novelty [Kaminskas and Bridge, 2016] that describe predicted items, the field lacks measures to accurately model the diversity of the user consumptions.

In this chapter, we introduce two coherence measures, **Surprise** and **Conditional Surprise**, that quantify different aspects of coherence and can be applied to both user interactions and model predictions. The former describes how surprising (uncommon) a user's consumptions' are, while the latter describes their internal coherence, independently of their uncommonness. We provide theoretical interpretations and properties of these measures and use them to improve our understanding of the performance of RS.

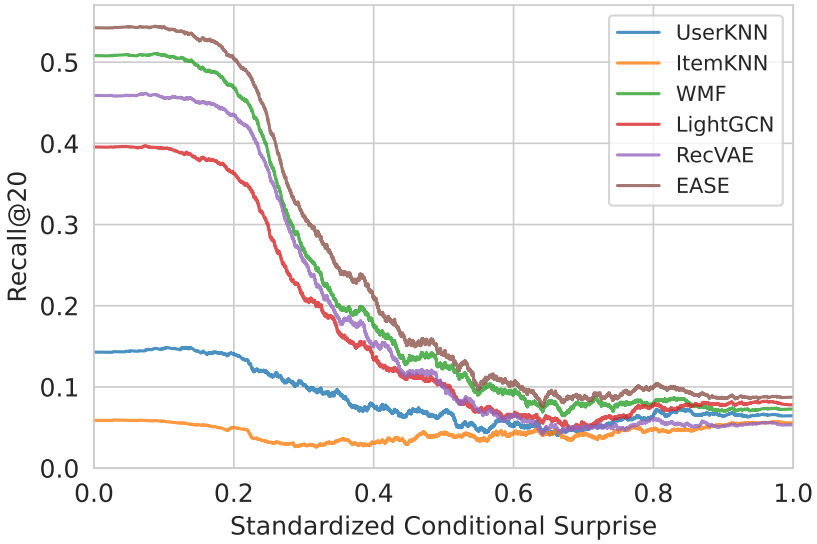


Figure 6.1: Performance in `Recall@20` of different RS, averaged over datasets, w.r. to our proposed **Conditional Surprise** ($CS(u)$) measure, standardized between 0 and 1, with a moving average smoothing. All RS performance collapse for high values of $CS(u)$.

Figure 6.1 illustrates the impact of our proposed Conditional Surprise measure on the performance of various recommendation algorithms across multiple datasets. The graph clearly demonstrates that as the Conditional Surprise increases (indicating less coherent user behavior), the performance of all recommender systems drastically declines. This relationship holds true across different algorithms and datasets, highlighting the importance of user coherence in recommendation tasks and the potential utility of our proposed measure for understanding and improving recommender system performance. Specifically, we address three fundamental questions in this regard:

Main Findings

1. **Can we develop a framework that describes user behavior and how it impacts the effectiveness of different recommendation algorithms?** We demonstrate that our measures capture nuanced user behavior patterns, revealing significant correlations between user coherence and algorithm performance across various domains.
2. **Can this framework allow for meaningful recommendation effectiveness comparisons across domains and algorithms?** Our proposed measures provide a common basis for evaluating recommender systems across diverse contexts, allowing us to compare RS performance across multiple scenarios.
3. **How can we leverage the proposed measures to enhance overall recommender system performance and adaptability?** Our analysis demonstrates that these measures can be used to optimize various aspects of recommender systems, from prediction quality and algorithm selection to coherence preservation, leading to more effective and efficient recommendations.

By combining the analysis of these coherence measures with the introduction of the Vis2Rec dataset, we aim to provide a comprehensive framework for evaluating and improving

recommender systems across diverse domains, including the challenging and underexplored field of tourism recommendations.

6.2 Related Work

RS have evolved significantly over the past decade [Wu et al., 2012; Roy et al., 2022]. However, recent studies still highlight persistent and major challenges within the RS field. More specifically, offline evaluation [Sun, 2023, 2024] and replicability [Dong et al., 2023] issues motivate the need for more rigorous evaluation approaches.

Recommendation algorithms

Traditional collaborative filtering approaches, such as user-based and item-based k-nearest neighbors (kNN), remain simple yet effective baselines [Nilashi et al., 2013]. However, the field has been predominantly shaped by Matrix Factorization (MF) techniques since the Netflix Prize challenge [Bennett et al., 2007b]. MF methods, including basic MF [Koren et al., 2009], Weighted MF [Hu et al., 2008], and extensions like GeoMF [Li et al., 2015], have shown superior performance in capturing latent factors of user preferences and item characteristics. More recently, the advent of deep learning has led to novel architectures in recommendation. Neural Collaborative Filtering [He et al., 2017] and Variational Autoencoders [Liang et al., 2018; Shenbin et al., 2020a] have demonstrated impressive results by capturing complex non-linear interactions. Graph-based methods, such as LightGCN [He et al., 2020b], leverage the inherent graph structure of user-item interactions to enhance recommendation quality. In the context of visual data, which is particularly relevant for domains like POI recommendation, approaches like VBPR [He and McAuley, 2016] incorporate visual features to enhance traditional collaborative filtering methods. More sophisticated models like CausalRec [Qiu et al., 2021b] attempt to address the causal relationships in recommendation, aiming for more robust and interpretable suggestions. These advancements reflect the field's progression towards more nuanced, context-aware, and potentially explainable recommendation models.

Recommendation datasets

Datasets play a major role in the development and evaluation of recommender systems. Several benchmark datasets have become standard in the field, each with its own characteristics and limitations. The MovieLens datasets, particularly the MovieLens 1M and 20M versions, have been widely used for benchmarking movie recommendation algorithms [Harper and Konstan, 2016]. These datasets offer a range of user-movie interactions, including ratings and timestamps. Another significant dataset is the Netflix Prize dataset, which spurred considerable advancements in collaborative filtering techniques [Bennett et al., 2007a]. In the e-commerce domain, the Amazon product reviews datasets cover various product categories and have been valuable for studying recommendation in retail contexts [Lakkaraju et al., 2013]. For music recommendations, the Last.fm dataset provides music listening data, offering insights into temporal patterns of user behavior [Konstan and Riedl, 2012]. However, in specific domains such as Point of Interest (POI) recommendation, there is a notable lack of comprehensive, large-scale datasets. Existing POI datasets, like those derived from Panoramio or Instagram, often face challenges related to data provenance, user privacy, and copyright issues [Wang et al., 2017a]. For instance, the Photo2Trip dataset, which included over 20 million geotagged images and 30,000 POIs, relied heavily on geotags, which are not always available or reliable [Lu et al., 2010]. The

scarcity of robust POI datasets highlights a gap in the field, particularly for tourism-oriented recommendation tasks that could benefit from visual and location-based data.

Evaluation protocols

The data collection process can also have an important impact on performance. Meng et al. [2020] examined data splitting strategies and their impact on evaluation outcomes while Ji et al. [2020] addressed data leakage in offline evaluation, highlighting the issue of evaluating on too few sets [Fan et al., 2024].

Several frameworks and analyses have thus been developed to enhance evaluation rigor and standardization. Sun et al. [2020] as well as Salah et al. [2020] propose complete benchmarking approaches for reproducible evaluation. More recently, new comprehensive frameworks like Elliot [Anelli et al., 2021] and ReChorus2.0 [Li et al., 2024] now provide standardized evaluation tools. While these frameworks offer valuable resources for consistent evaluation, they primarily focus on traditional performance metrics.

In this regard, evaluation metrics play a major role in assessing the performance and effectiveness of recommender systems. While traditional metrics focus primarily on accuracy, recent research has emphasized the importance of considering additional factors to provide a more comprehensive evaluation [Kuanr and Mohapatra, 2021]. In the context of collaborative filtering, the accuracy is usually measured through `Recall@K`, `Precision@K`, or `NDCG@K`, among others, which describe the RS's capacity to recover the user's interactions that were hidden in test-time. These metrics have been widely used [Silveira et al., 2019], but leave out other recommendation aspects that are in the users' interests, some of which have been tackled over the years. More precisely, Konstan and Riedl [2012] introduced user satisfaction in evaluations. Kaminskas and Bridge [2016] examined diversity, serendipity, and coverage metrics. The use of these metrics was democratized to evaluate model prediction in the following works [Silveira et al., 2019; Alhijawi et al., 2022]. Additional works highlighted the importance of considering these newer metrics in order to better match users' behavior and find better recommendations [Kim et al., 2021; Ping et al., 2024]. Using relevant metrics can also allow for analyzing group disparity in the behavior of the recommendation system. Recent work thus used these newer metrics in order to analyze such disparity in given user or item segments [Dong et al., 2023; Diricic et al., 2023].

At the same time time, concerns about domain-specific evaluation methodologies have emerged, highlighting the need for more nuanced and versatile evaluation approaches. Latifi et al. [2022] proposed metrics for session-based recommendations in streaming contexts, revealing that traditional static metrics often fail to adequately assess the performance of recommender systems in dynamic, session-based environments. Dietz et al. [2023] studied the influence of data characteristics on point-of-interest recommendation algorithms, demonstrating that factors such as geographical distribution and temporal patterns significantly impact algorithm performance. Their findings underscored the importance of considering domain-specific features in evaluation methodologies. Building on these insights, Sun [2024] questioned the cross-domain applicability of current evaluation practices, arguing that metrics and evaluation protocols optimized for one domain may not translate effectively to others. They highlighted the need for more generalizable evaluation frameworks that can account for domain-specific nuances while still enabling meaningful cross-domain comparisons.

Recent research advocates for improvements in fundamental aspects of recommender system evaluation, such as data handling, performance metrics, and cross-domain applicability [Fan et al., 2024; Shevchenko et al., 2024]. Additionally, emerging works highlight new concerns,

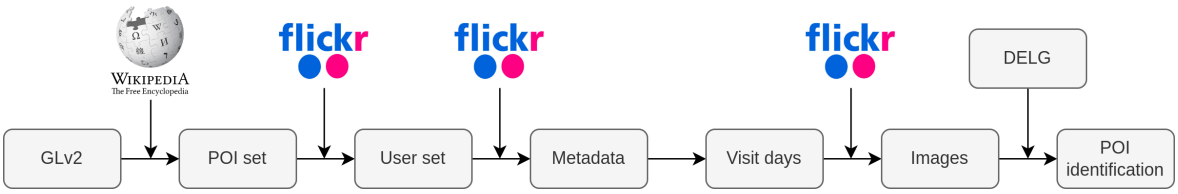


Figure 6.2: Data collection and annotation pipeline.

including the need to create RS tasks more aligned with real-world applications [Sun, 2024] and to develop a better understanding of users' true intents [Kleinberg et al., 2022].

Collectively, these studies depict a field in transition, showing the need for better evaluation practices to match the increasing sophistication and real-world impact of recommender systems.

6.3 Vis2Rec: A Visual Dataset for Visit Recommendation

The goal of Vis2Rec is to provide a realistic and sustainable testbed for visit recommendations based on user images. To meet this objective, we need to address technical, legal, and ethical challenges.

The dataset is built to propose recommendations at scale, and after the correct processing for recommendation, it caters to at least 36,111 POIs in 5,012 cities. These POIs are taken from Google Landmarks v2 (GLv2) [Weyand et al., 2020a] in order to enable large-scale visual POI recognition. The size of the user set is also important in order to capture diversified user preferences. Preprocessed Vis2Rec includes a total of 14,600 users, 829,673 POI-associated user images, and over 6M additional images.

Sustainability is ensured by implementing a legally compliant data collection and distribution process. The dataset includes only distributable images which were taken on visit days. Equally important, face de-identification was applied to ensure the anonymity of the users.

We describe the main steps of the dataset constitution and packaging below, and the data collection and structuring pipeline is summarized in Figure 6.2.

6.3.1 Initial data collection

POI set. GLv2 [Weyand et al., 2020a] is one of the largest publicly available POI-related datasets, which was collected from Wikimedia Commons. We use the "clean" subset, which includes a total of 1,580,470 images, which represent 81,313 POIs. GLv2 is therefore adapted for the creation of a comprehensive visit recommendation dataset, such as Vis2Rec. To perform efficient data queries, we need to enrich this dataset by mining information from the Wikipedia pages associated to POIs. The resulting dataset includes the name of the POI (with translations, when available), its associated GPS coordinates, and the closest city from the Geonames[1] list of 139,439 cities that have at least 1000 inhabitants.

User set. Flickr offers an easy-to-use API for a large collection of images and associated metadata, and is as such a very adapted data source to our work setup. We launch Flickr API queries with the POI name(s), using a 3 km radius around the coordinates. Queries are limited to photos distributed under Creative Commons licenses to ensure that they are redistributable. Metadata for up to 5000 photos is collected for each POI, containing photo ID, user ID, and user tags, as well as geographic coordinates of the photos. This process provides an initial list of 20k preselected users.

[1] `https://www.geonames.org/`

6.3.2 Domain-related data selection

The image collection should be focused on tourist visits. More specifically, we collect all the photos corresponding to a potential visit day, determined by generating coarse POI predictions for each image. A day is kept if it includes at least one POI name in the image tags. Since POI names are often ambiguous [Popescu et al., 2008; Serdyukov et al., 2009], further post-processing is needed to disambiguate potential POI matches. Whenever geolocation is available for at least one photo taken during one day, it is used to check for POIs which are located within a radius of 10 kilometers. If geolocation is not available, we resort to text-based matching which uses a probabilistic geographic language model [Serdyukov et al., 2009]. This model associates the visit day with a list of probable cities based on the aggregation of the location probabilities of the tags used during a tested day. A geolocated subset of metadata is used to determine a threshold which provides a good balance between precision and recall for detected visit days.

This matching provides a text-based profile of each user [Kurashima et al., 2013] which is used to select interesting users for the visual dataset. The direct use of text-based profiles for recommendation [Kurashima et al., 2013; Popescu and Grefenstette, 2011] is possible but is suboptimal since users are required to provide explicit textual annotations of their visits, which often leads to incomplete profiles. The resulting intermediate dataset includes 17k user profiles and a total of 27k text-annotated POIs.

6.3.3 Visual matching of POIs

Vis2Rec is intended for recommendation based on the sole use of photo content and we should make no assumption regarding the availability of textual annotations or geolocation for the dataset. This is important in practice in order to design a profiling pipeline that does not require any effort from the users. Consequently, we collect images for the visit days identified in the intermediate dataset based on tags (Subsection 6.3.2). These photos are then compared to POI images from Google Landmarks v2 dataset [Weyand et al., 2020a] using a DELG descriptor [Cao et al., 2020].

Visual matching procedure. Visual matching is performed using DELG [Cao et al., 2020], which achieves state-of-the-art single model instance-level recognition on GLv2. We use the model only for inference since the pretrained weights on GLv2 can be found in the official implementation [2]. The visual matching of candidate and reference images is done in two steps:

1. A 2048-dimensional global embedding is used to select a subset of similar reference images from GLv2 for each candidate image in which POI occurrences are searched. Following common practice, the top 20 most similar reference images are retained for the second step.
2. A geometric verification process based on 128-dimensional local descriptors provided by DELG is performed to refine the list of similar reference images. The final ranking is based on the number of matched keypoints between the candidate and the reference images.

This two-step process is needed since global retrieval is fast but potentially prone to errors, while geometric verification is slow but accurate. Each candidate image is paired with the reference image that has the highest matching score, and attributed with the POI represented by this reference image. The number of keypoints can be used as a confidence estimator for the quality of visual matching.

Results. Since DELG was pretrained on the same POI set as Vis2Rec , the visual matching procedure has good qualitative results (see Figure 6.3). Correct identification is possible for a

[2] `https://github.com/tensorflow/models/tree/master/research/delf`

Figure 6.3: Examples of visual matches provided by DELG. The model correctly recognizes (a) outdoor landscapes, (b) indoor scenes, and (c) different lighting conditions. Errors can be caused by (d) the same objects in different places, (e) Signs with identical features, and (f) similar architectures.

wide range of setups, including outdoor landscapes, indoor architectures, as well as difficult lighting conditions. However, this process is far from perfect and fails in particular situations (Figure 6.3). By analyzing the results of the visual matching, we can identify three types of recurring errors: (1) objects which occur in different regions of the world and are representative for POIs (Figure 6.3 (d)); (2) visually similar objects which are specific to a city (Figure 6.3 (e)); and (3) visually similar POIs (Figure 6.3 (f)).

The first type of error can be reduced by removing GLv2 reference images which match target images located in different parts of the world. To do this, we use a geolocated validation set and remove any reference image which was matched only to POIs farther than 15km away at least 5 times. The remaining spatial aberrations are removed by selecting the most confident POI detection for each day and removing detections corresponding to POIs farther than 100km from it. This geographic filtering removes over 1 million images.

The second type of error is the most difficult to handle since neither a spatial criterion nor a good matching score threshold can deal with them.

The third type of error is usually associated to lower matching scores. By manually verifying a few hundred matched image pairs, we observe that a matching score of 30 leads to an accuracy of at least 98%. Interestingly, this coincides with the threshold chosen in the GLv2 article [Weyand et al., 2020a] to generate the "clean" subset, and to the threshold that leads to the best recommendation results. In the rest of this work, this is the default chosen threshold.

6.3.4 Data distribution

We present dataset-level statistics for detected POIs and user visits. These statistics are obtained after applying the visual matching error mitigation measures described in Subsection 6.3.3, and lead to a dataset comprised of 36,111 unique POIs, depicted on 820,593 images, corresponding to 421,065 unique user visits. Since these statistics highly depend on the chosen matching threshold, the distributed dataset contains all of the POI predictions without any filtering to allow for further research and POI discovery.

Identified POIs. Figure 6.4 illustrates the distribution of identified visits across the world, along with the associated number of detections. The obtained distribution is in line with global tourist visit trends [UNWTO, 2019], and shows a strong concentration of POIs in Western Europe, East and West coasts of North America, and Eastern and South-Eastern Asia. The distribution is also influenced by Flickr usage trends, and confirms previous analyses of geolocated photos shared on this platform [Crandall et al., 2009; Popescu et al., 2008].

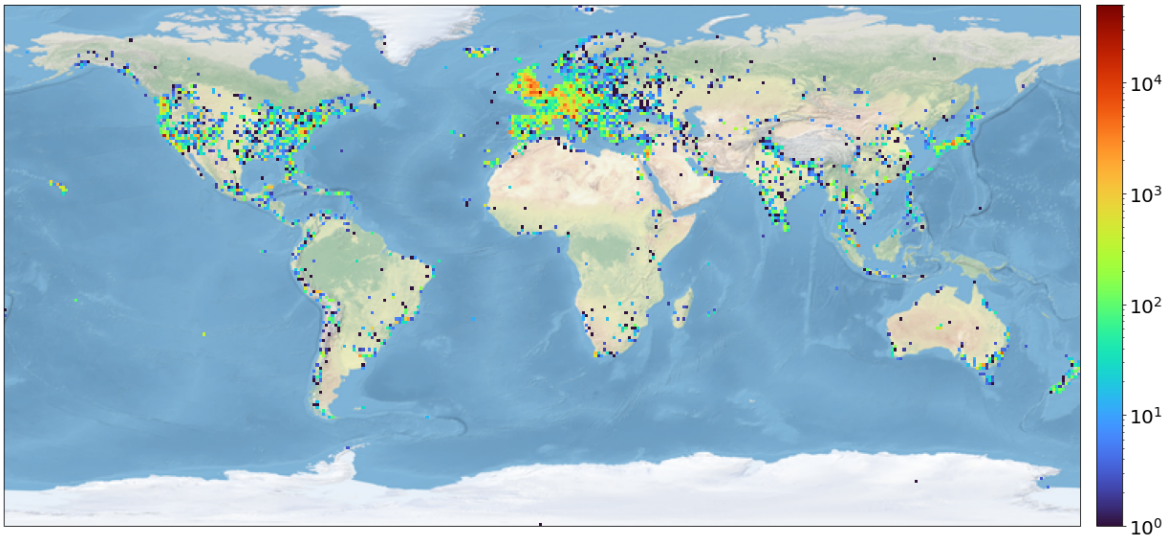


Figure 6.4: Spatial distribution of the density of identified visits

The distributions of the number of identified POIs and the number of visits per city are proposed in Figures 6.5 and 6.6, respectively.

Both of them exhibit long-tail shapes, with a large number identified POIs and of visits concentrated in large tourist hotspots, such as London, Paris, New York City, and significantly fewer visits associated to the other cities. More details about the visited POIS and visits in the different cities are provided in the supplementary material.

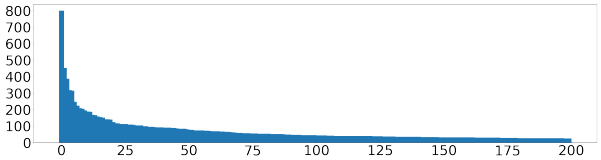


Figure 6.5: Distribution of the number of identified POIs in the top 200 cities.

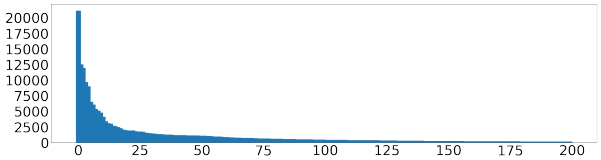


Figure 6.6: Distribution of the number of user visits in the top 200 cities

User visits. User profiles generated in Vis2Rec are rich and diversified. First of all, 84% of the users visited at least 5 POIs, a threshold commonly used in recommender systems for filtering purposes, while the median user visited 16 distinct POIs. Secondly, 95% of the users visited more than one city, 8 being the number of cities visited by the median user, resulting in rich travel profiles. These observations can be easily explained by the fact that travel images are often uploaded to Flickr to highlight their extraordinary nature. Therefore, one should keep in mind that Vis2Rec does not contain images that are representative of the everyday life of its users, but more of their vacation travels.

Additional images. Confident POI detections account for 11% of the 7,158,454 total images. We estimate that between 1 and 2 million other images could depict POIs, and counting

them as valid by lowering the matching score threshold would increase the POI set to around 60k unique POIs. However, this introduces many false positives in the user profiles, resulting in lower recommendation performances. As per this observation, a threshold of 30 matching keypoints is kept throughout our work. The remaining images are non-POI personal user photos and are distributed for potential further work.

6.3.5 Dataset annotation

In preliminary experiments, we analyzed random samples of target-reference image pairs provided by the geometric matching process. We partitioned the matched pairs into bins based on their matching score, each bin corresponding to a 10-keypoints window. We then drew 500 random samples from each bin and performed a manual verification of the matched pairs. The results showed that the visual matching has an accuracy of over 99% when the number of matched keypoints is larger than 40.

A manual annotation process is run base on this observation. A total of 10k image pairs with a matching score lower than 40 are manually verified. The task is relatively simple since annotators need to decide whether the two images of a target-reference pair depict the same POI or not. Three annotators verify each pair, and we consider the match to be correct if at least two of them label it correctly.

This allows us to get the number of True and False positives with for different matching scores, as presented in Figure 6.7. Based on these observations, we choose a threshold of 30 keypoints for a positive matching.

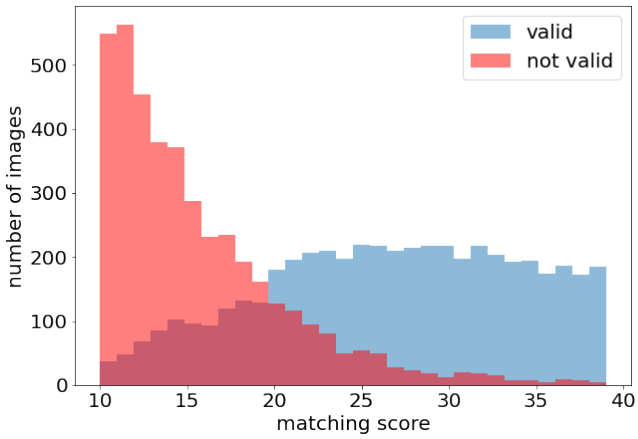


Figure 6.7: Distribution of the annotated image pairs

6.3.6 Dataset compliance

First, Vis2Rec was collected via the official Flickr API, a data source that allows the constitution of datasets made of data originally shared by its users. For instance, the well-known YFCC100M dataset [Thomee et al., 2016] was also collected from Flickr and is still available today. Second, we keep only images that are shared under Creative Commons (CC) licenses in order to enable lawful redistribution of content. The dataset will be published using a license that is compatible with the most restrictive CC licenses included in Vis2Rec, and commercial reuse will be notably not permitted. Third, we will enforce the data minimization

principle defined in Article 5 of the General Data Protection Regulation[3], and share only the data needed for the POI recommendation task. The dataset includes only images taken on days that correspond to tourist visits. A qualitative exploration of Vis2Rec showed that it contains many personal images. As such faces will be de-identified [Ma et al., 2021] in the dataset to protect the anonymity of the depicted persons.

6.4 Analyzing User Coherence in Recommender System

6.4.1 Notations

We denote the set of users as $\mathcal{U}$ and the set of items as $\mathcal{I}$. Let $n = |\mathcal{U}|$ be the number of users and $m = |\mathcal{I}|$ the number of items. We place ourselves in a binary setting where each user $u \in \mathcal{U}$ can either interact ($x_{ui} = 1$) or not interact ($x_{ui} = 0$) with an item $i \in \mathcal{I}$, leading to the binary interaction matrix $X \in \{0,1\}^{n \times m}$. We identify each user to its item set so that "u" refers equivalently to a user id and to their item set.

A set of test users $\mathcal{U}_{test}$ is randomly sampled. For each test user, their last interaction is isolated as the test target $x_{ui_{test}}$. All other interactions are part of the training set. The goal of a recommender system is to learn a function $f : \mathcal{U} \times \mathcal{I} \to \mathbb{R}$ that assigns a score $f(u,i)$ to each user-item pair (u,i), indicating the likelihood of user u interacting with item i. The system's performance is evaluated by its ability to rank the test item i_{test} highly among all items not in the user's training set.

6.4.2 Coherence measures

Our goal is to understand and quantify individual user behavior in recommender systems. While traditional approaches often simplify users into entries in a user-item matrix, this can overlook important nuances in user preferences and consumption patterns. By modeling individual user behavior more comprehensively, we can gain insights into why certain recommendations succeed or fail, and potentially tailor our approaches to different types of users.

Definition: Coherence in Recommender Systems

We define coherence as "*the degree to which a user's interactions form a consistent and predictable pattern*". A highly coherent user would have a set of interactions that align well with each other and with common consumption patterns. In contrast, a less coherent user might have more erratic or diverse interactions.

Specifically, we want to measure how the performance of recommender algorithms is impacted by how surprising or unpredictable a user's behavior is. To model surprise, one natural way is to first assign probabilities to items. In previous works [Kaminskas and Bridge, 2016], the probabilities used to compute existing measures describe the prediction distribution. However, this approach is limited as it focuses on the model's output rather than the inherent characteristics of user behavior. Here, we adopt another view and consider the item probability as their frequency among the users:

$$p_i^* = \frac{\left|\{x_{ui} = 1\}\right|}{n} \tag{6.1}$$

[3] https://gdpr-info.eu/art-5-gdpr/

This quantity is insufficient since a user can interact with very rare items but still have a very coherent set of items (e.g. niche movies from the same director). Therefore, we consider the second-order statistics:

$$p^*_{i,j} = \frac{\left|\{x_{ui} = x_{uj} = 1, u \in \mathcal{U}\}\right|}{n} \quad \text{and} \quad p^*_{i|j} = \frac{p^*_{i,j}}{p^*_j} \tag{6.2}$$

The probability $p^*_{i|j}$ represents how much a user is likely to interact with i when they also interacted with j.

Then, we can define the **Surprise** (or Information Content) of an item as $-\log(p^*_i)$ and the **Conditional Surprise** $-\log(p^*_{i|j})$. The first natural quantity to study is the mean empirical binary cross-entropies:

$$\widetilde{S}(u) = -\frac{1}{m}\sum_{i=1}^{m} \log(p^*_i) x_{ui} \tag{6.3}$$

$$\widetilde{CS}(u) = -\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} \log(p^*_{i|j}) x_{ui} x_{uj} \tag{6.4}$$

However, these existing definitions have a problem: they are non-decreasing when a new item is added to a user's set, regardless of the item's characteristics. This behavior is counter-intuitive, as we would expect the overall surprise to potentially decrease if a highly predictable or common item is added to the user's profile. Thus, we define our user coherence measures, called **Mean Surprise** and **Mean Conditional Surprise**:

Definition: Coherence Measures

Mean Surprise and **Mean Conditional Surprise** are defined by:

$$S(u) = -\frac{1}{|u|}\sum_{i \in u} \log(p^*_i) \tag{6.5}$$

$$CS(u) = -\frac{1}{|u|^2}\sum_{i \in u}\sum_{j \in u} \log(p^*_{i|j}) \tag{6.6}$$

where $|\cdot|$ is the L^1 norm.

To show the relevance of our measures, we also compute an Oracle version of our measures on the test items. In our leave-one-out setup, they simplify to $S(i_{test}) = -\log(p_{i_{test}})$ and $CS(i_{test}) = -\frac{1}{|u|}\sum_{j \in u} \log(p_{i_{test}|j})$.

6.4.3 Interpretation and Properties

The quantities in Equations 6.5 and 6.6 have a similar form as in Equations 6.3 and 6.4, but have a dynamic rescaling dependent on the user. This ensures that the measures are comparable across users with different numbers of interactions, providing a fair basis for comparison. Unlike the previous formulations, these measures can decrease when a user interacts with a common item, better reflecting intuitive notions of surprise and coherence. The Mean Surprise $S(u)$ describes at the first order how much a user deviates from the popular items, on a scale from the ***unsurprising*** users to the ***surprising*** users. The Mean Conditional Surprise $CS(u)$ indicates whether the co-occurrences in the user's consumption set are far

from frequent co-occurrences, capturing the internal consistency of a user's choices, on a scale from the ***coherent*** users to the ***incoherent*** users.

In RS data, user behavior and item consumption detection can be quite noisy [Amatriain et al., 2009]. We can verify how our measures behave on average:

Average Bounds on S and CS

Let π_u be the distribution from which u is drawn, and $\pi_u^{\geq 1}$ be the distribution of u conditioned on $|u| \geq 1$. Let $S^*(u) = \mathbb{E}_{\pi_u^{\geq 1}}[\widetilde{S}(u)]$ and $CS^*(u) = \mathbb{E}_{\pi_u^{\geq 1}}[\widetilde{CS}(u)]$. Then:

$$\frac{m}{\mathbb{E}_{\pi_u^{\geq 1}}[|u|]} \leq \frac{\mathbb{E}_{\pi_u^{\geq 1}}[S(u)]}{S^*(u)} \leq \mathbb{E}_{\pi_u^{\geq 1}}\left[\frac{m}{|u|}\right] \tag{6.7}$$

$$\frac{m^2}{\mathbb{E}_{\pi_u^{\geq 1}}[|u|^2]} \leq \frac{\mathbb{E}_{\pi_u^{\geq 1}}[CS(u)]}{CS^*(u)} \leq \mathbb{E}_{\pi_u^{\geq 1}}\left[\frac{m^2}{|u|^2}\right] \tag{6.8}$$

Proof. See appendix. □

In particular, we see that the lower bound for the scaling depends only on the expected value of the number of items. The upper bound, however, can get bigger than $m/\mathbb{E}_{\pi_u^{\geq 1}}[|u|]$, due to the Jensen inequality. Sadly, there is no way of simply bounding it without additional hypotheses. Since $|u|$ is the number of items consumed by u, we can model it as a Poisson variable with parameter λ (see appendix). We have the following bound:

Expectancy Bound of an Inverse Poisson Variable

If X is a Poisson variable of parameter $\lambda > 0$, we have:

$$\mathbb{E}_{\geq 1}\left[\frac{1}{X}\right] \leq \frac{2}{\mathbb{E}_{\geq 1}[X]} \tag{6.9}$$

Proof. See appendix. □

Empirically, we find a tighter upper bound for equation (6) with a numerator equal to 1.37 instead of 2. This upper bound is met for $\lambda_{sup} \approx 2.9$. For smaller and larger values of λ, the numerator quickly drops to 1. This means that for surprising users, on average, the empirical estimation should not be too far from $S^*(u) \times m/\mathbb{E}_{\pi_u^{\geq 1}}[|u|]$, which is indeed the classical estimator expectancy re-scaled by the mean proportion of items that a user will consume. The same bound effects apply to $CS(u)$.

6.4.4 User Coherence Segmentation

We study how different recommendation algorithms react to different users. For each train dataset $\mathcal{D}$, we calculate $S(u)$ and $CS(u)$ for each user u. In particular, we have an estimate of the measures for the test users, using their train interactions. This allows us to segment the dataset into bins based on the value of the user measures. As we will see, the conditional surprise metric plays an important role in explaining user performance, so we mainly use $CS(u)$ for our user segmentation. We denote by $\mathcal{D}[\alpha, \beta]$ the dataset comprised of users with a $CS(u)$ between the α^{th} and the β^{th} percentiles. For example, $\mathcal{D}[0, 0.1]$ is the set of *coherent* users, i.e., users with a Mean Conditional Surprise in the first decile.

Dataset	Items	Users	Inter.	Density
ML 1M	3.1K	6K	562K	$3 \cdot 10^{-2}$
ML 10M	9.4K	69K	5.7M	$9 \cdot 10^{-3}$
Netflix Small	2.7K	8.3K	320K	$1 \cdot 10^{-2}$
Netflix	18K	463K	59.9M	$7 \cdot 10^{-4}$
Vis2Rec	9.3K	9.1K	200K	$2 \cdot 10^{-3}$
Tradesy	12K	6.6K	73K	$9 \cdot 10^{-4}$
Amazon Music	11K	8.6K	87K	$9 \cdot 10^{-4}$
Amazon Office	62K	20K	468K	$4 \cdot 10^{-4}$
Amazon Toys	143K	61K	1.2M	$1 \cdot 10^{-4}$

Table 6.1: Description of the datasets after processing

6.4.5 Regression Model as an Analytical tool

To accurately quantify the impact of our measures on RS performance, we employ logistic regression to model the relation between attributes and binary outcomes, as describes in Chapter 3.

When X is a variable estimating X^* with a certain variance σ^2, the plain regression on X becomes imprecise. We the SIMEX (Simulation-Extrapolation) method [Cook and Stefanski, 1994], which simulates many regressions with the added noise σ^2, and extrapolates to the case of no noise, providing more robust coefficient estimates.

6.5 Experimental Setup

6.5.1 Datasets

We perform our analysis on 9 datasets of various sizes and domains:

- **MovieLens 1M** and **MovieLens 10M** [Harper and Konstan, 2016]: movie ratings recommendation datasets, commonly used for benchmarking recommendation algorithms;
- **Netflix Small** and **Netflix** [Bennett et al., 2007a]: two versions of the Netflix Prize dataset, consisting of movie ratings;
- **Amazon Music**, **Amazon Office**, and **Amazon Toys** [Lakkaraju et al., 2013]: part of the Amazon product reviews collection, focusing on different product categories;
- **Tradesy** [Lakkaraju et al., 2013]: interactions for the Tradesy platform, which specializes in the resale of designer fashion;
- **Vis2Rec** [Soumm et al., 2023]: the dataset previously introduced.

We used different sizes from the same dataset, and datasets from the same domain but different sources, specifically to study the effects of dataset size and domain, and analyze recommender systems across a wide range of real-world applications.

6.5.2 Data Processing

Following standard practices [Meng et al., 2020], we binarize non-binary ratings, which are all in $[1, 5]$, by setting $x_{ui} = \mathbb{1}(r_{ui} > 3)$. A 5-code is extracted from the datasets, i.e. a subset of users and items with at least 5 interactions, by sequentially filtering out users and items with less than 5 interactions until convergence. This pre-processing reduces the effective sizes of the datasets, leading to a size distribution described in Table 6.1.

We sample 10'000 test users from each dataset, from which the last interaction is isolated as the test set, and the second-to-last interaction is isolated in a validation set. For datasets with less than 10'000 users, all users are considered test users.

For each dataset, the Mean Surprise and Conditional Surprise are computed for each user. This allows us to easily create the segments $\mathcal{D}[\alpha, \beta]$ for multiple values of α and β, both in the train and test set. Since most algorithms cannot handle users that are not in the train set, we make sure that the test users are always part of the train set. However, some items of the test set may not always be present in the train set, which makes the task more complicated. All relevant code is provided in the appendix.

6.5.3 Recommender Algorithms

We benchmark 7 different recommendation algorithms:

- `MostPop` is the baseline algorithm that recommends the most popular items to every user;
- `UserKNN` is a neighborhood approach that relies on a similarity between pairs of users.
- `ItemKNN` is a neighborhood approach that relies on a similarity between pairs of items;
- `WMF` is a weighted matrix factorization approach that learns user and item embeddings with gradient descent, minimizing a reconstruction loss;
- `EASE` [Steck, 2019] is a matrix factorization approach that computes an item-item weight matrix with a closed-form formula;
- `LightGCN` [Shenbin et al., 2020b] is an approach that learns user and item embeddings by aggregating information from the user-item interaction graph using a light graph convolution;
- `RecVAE` [Shenbin et al., 2020b] is a variational auto-encoder approach inspired by β-VAE [Higgins et al., 2017] and denoising-VAE [Im et al., 2016].

These algorithms have been chosen to represent a wide range of possible usages, depending on how well they scale with the number of users, or the number of items, training times, or how well they adapt to new users or items.

6.5.4 Training

In the leave-one-out protocol, as there is only 1 relevant test item per user, the two reference metrics `Recall@K` and `Precision@K` [Herlocker et al., 2004] are equivalent, up to a constant, since they are proportional to the number of relevant items. Therefore, we only use the metric `Recall@K`, which corresponds in this case to a binary variable 0/1. For each experiment (i.e. algorithm trained on a dataset or a dataset segment), we perform a hyperparameter search using `optuna` with 50 rounds maximizing the `Recall@20` on the validation set.
The models are trained on 256 AMD EPYC 9554 64-Core CPUs and 1.4TB of RAM for the algorithms that run on CPU, while others are run on a single NVIDIA A100 GPU with 40GB of VRAM.

6.6 Results and Analysis

6.6.1 Experimental Properties of the Measures

6.6.1.1 Measure Distribution.

Figure 6.8 presents the distribution of our proposed measures across the datasets. A key observation is that the distribution of $S(u)$ characterizes the domain: the movie-related datasets exhibit comparable Mean Surprise values. This suggests a uniformity in user behavior patterns within the movie recommendation domain, even if they come from a different source and collection process. In contrast, all e-commerce datasets demonstrate higher $S(u)$ values, indicating higher

diversity in user consumptions. Vis2Rec representing a unique domain of tourism recommendations, falls between the movie and e-commerce clusters, highlighting its distinct nature.

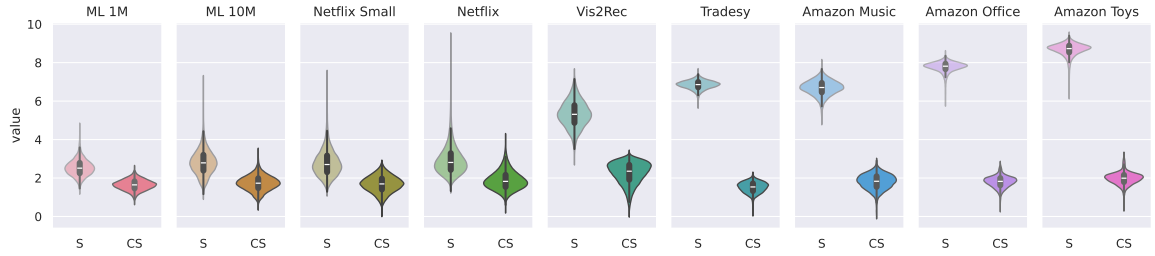


Figure 6.8: Distribution of the measures across datasets. S denotes the Surprise measure, and CS the Conditional Surprise measure. CS shows remarkable stability across all datasets.

Interestingly, despite the variations in the distribution of $S(u)$ across domains, we observe consistency in the distribution of $CS(u)$ across all datasets. This suggests that the $CS(u)$ measure is a good candidate for a domain-agnostic coherence measure.

6.6.1.2 Comparison with naive measures.

We start by verifying that our measures behave better than existing measures, such as the mean-cross entropies $\widetilde{S}(u)$ and $\widetilde{CS}(u)$ defined in Equations 6.3 and 6.4. We graphically inspect the relationship between candidate measures and $|u|$. As we see in Figure 6.9, $\widetilde{S}(u)$ diverges with the number of items, whereas $S(u)$ tends to stabilize as $|u|$ increases.

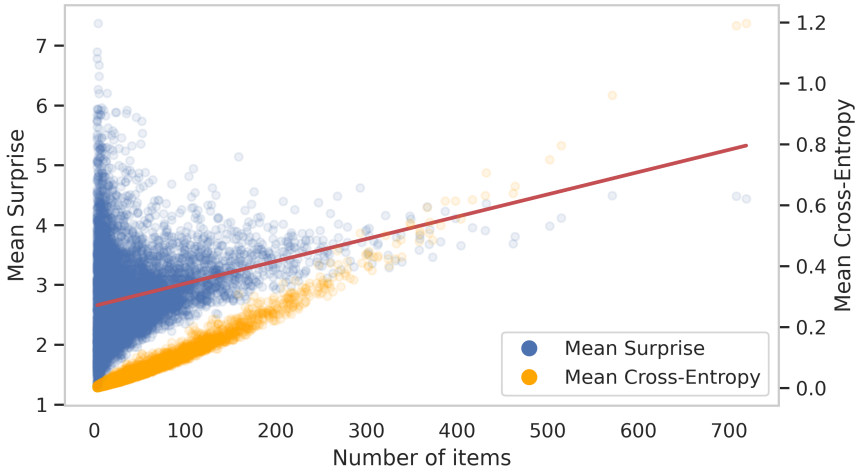


Figure 6.9: Comparison of $S(u)$ and $\widetilde{S}(u)$ against $|u|$ on the Netflix dataset, with a regression fo $S(u)$ on $|u|$. Similar graphs are produced for other datasets and for $CS(u)$ (see supplementary material).

A closer inspection reveals that $(S(u) - \hat{\mathbb{E}}_u[S(u)])$ still increases with $|u|$, but with a constant asymptotic behavior. This is coherent with the idea that users with fewer items can either over-consume very popular items, or on the contrary deviate into specific tastes, whereas users with more items have preference on average for both popular and specific items, converging to a single limit distribution. $CS(u)$ behaves the same way.

6.6.1.3 Correlation between measures

Table 6.2 presents the linear dependence between $CS(u)$ and $S(u)$ across the tested datasets.

Dataset	Coefficient	R^2
ML 1M	0.58	0.79
ML 10M	0.55	0.86
Netflix Small	0.49	0.67
Netflix	0.6	0.84
Vis2Rec	-0.12	0.02
Tradesy	-0.66	0.28
Amazon Music	-0.66	0.43
Amazon Office	-0.68	0.44
Amazon Toys	-0.66	0.51

Table 6.2: Linear dependence between the information measures across datasets

As before, we see a clear distinction between the different domains. For movie datasets, we observe positive correlation coefficients ranging from 0.49 to 0.6. This suggests that in movie datasets, unsurprising users also happen to be coherent. The higher R^2 value for these datasets indicates that this relationship is quite significant and consistent within the movie domain.

In contrast, e-commerce datasets demonstrate strong negative correlations, with coefficients consistently between -0.66 and -0.68. This negative relationship implies that in online shopping contexts, the coherent users are also the most surprising ones. The moderate R^2 values suggest that while this inverse relationship is significant, it's not as deterministic as in the movie domain.

The Vis2Rec dataset stands out with a weak negative correlation (-0.12) and a very low R^2. This implies that for this dataset, there is only a very weak link between surprise and coherence.

Main Experimental Properties of S and CS

The Surprise and Conditional Surprise are measures that:

- **Can be interpreted in absolute**: S clearly indicates raw diversity in user consumption, and is higher of e-commender datasets than in movie datasets. CS is mostly stable across datasets and allows us to compare user behaviour between different domains.
- **Have better properties than traditional entropies**: users with fewer items have a broader range of values fo S and CS. The dependence between the measures and the number of items is less deterministic.
- **Characterize domains based on their correlation**: looking at the correlation between S and CS provides a lot of information about user's mean behavior in the datasets.

6.6.2 Overall Performance

The overall results are presented in Figure 6.10. All combinations of datasets and algorithms have been benchmarked except UserKNN on Netflix. Due to the amount of users, the method saturates the 1.4 TB of RAM at our disposal. Figure 6.10 shows two global trends. First, there are substantial differences in performances between movie datasets and e-commerce datasets. This can be explained by the density of these datasets (see Table 6.1), which is much higher for movie datasets. Second, performances are better when the dataset size increases: the performances of the algorithms are overall better on ML 10M and Netflix than, respectively, on ML 1M and Netflix Small. This is not trivial since increasing the size not only increases the number of samples (users) but also the number of items, which theoretically makes the task harder.

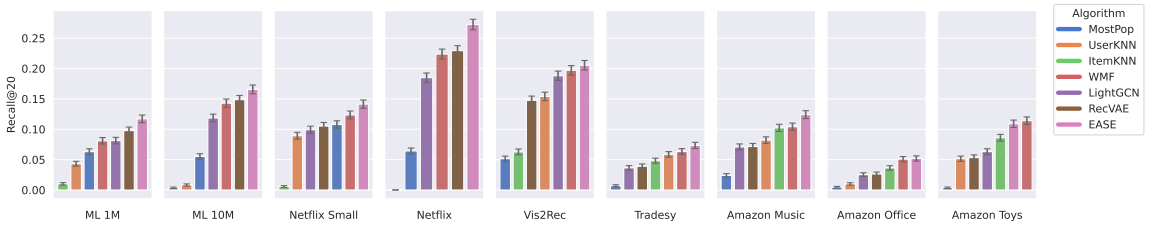


Figure 6.10: Overall algorithms performance across datasets, measured in `Recall@20` with confidence intervals at 95%.

`EASE` provides the best performances on almost all datasets. `WMF` is second everywhere except for the biggest sets, where `RecVAE` seems, as expected from a deep approach, to leverage better the amount of data at its disposal. However `RecVAE` as well as `ItemKNN` show themselves sensitive to the number of items and users. In particular, `RecVAE` performs better when the number of items is below the number of users. Conversely, `ItemKNN` performs better when the number of users is below the number of items.

6.6.3 Validating Coherence Measures

For each dataset, we demonstrate the usefulness of our coherence measures by performing a logistic regression of the binary `Recall@20` metric denoted $Rec(u)$ on the oracle test information measures and the profile density, controlling for the used algorithm:

$$\texttt{Rec(u)} \sim \sigma(\texttt{algo} + \texttt{|u|} \times \texttt{S}(i_{test}) \times \texttt{CS}(\texttt{i}_{test})) \tag{6.10}$$

We compute the average marginal effects (AME) of the variables and the McFadden's R^2 of the model, reporting the values in Table 6.3.

Dataset	$\|u\|$	$S(i_{test})$	$CS(i_{test})$	R^2
ML 1M	-0.02	0.02	-0.22	0.34
ML 10M	-0.02	0.06	-0.26	0.50
Netflix Small	0.02	-0.00	-0.19	0.53
Netflix	-0.05	-0.04	0.02	0.14
Vis2Rec	0.05	-0.02	-0.17	0.4
Tradesy	0.03	-0.03	-0.05	0.32
Amazon Music	0.04	-0.05	-0.08	0.33
Amazon Office	0.02	-0.01	-0.03	0.31
Amazon Toys	0.05	-0.03	-0.09	0.35

Table 6.3: Marginal effects of the variables for equation 6.10. All values are significant at p-value $< .05$. For e.g. on ML 1M, the increase of 1 std. in $CS(i_{test})$ causes the `Recall@20` to decrease of 22 points on average.

The regressions have overall very high R^2 values, which shows the power of our regression[4]. We additionally noticed that the marginal effects of the measures, in particular $CS(u)$, are as important as the algorithm's. The mostly negative coefficients highlight the sensitivity of the algorithms to the coherence of the test item with respect to the input items.

[4]For logistic regression, McFadden $R^2 > 0.2$ is considered an excellent fit [Allison, 2014]

6.6.4 Impact on Performance

We now directly estimate how our measures computed on the train set impact RS performance. Figure 6.1 shows the relation of the `Recall@20` to $CS(u)$. This graph reveals several important insights:

- There is a clear negative correlation between $CS(u)$ and recommendation performance across all algorithms.
- The performance gap between different algorithms is most pronounced for coherent users, i.e. for users with low $CS(u)$ values.
- As $CS(u)$ increases, the performance of all algorithms converges to a similarly low level.

Notably, the convergence of algorithm performance for high $CS(u)$ values suggests that for highly incoherent users, the choice of algorithm becomes less important. This observation has significant implications for recommender system design and deployment. It indicates that most gains in overall performance primarily come from improvements in recommendations for coherent users. For incoherent users, even sophisticated algorithms struggle to outperform simpler approaches.

To get the true marginal effect independent of other variables, we estimate the model:

$$\texttt{Rec(u)} \sim \sigma(\texttt{|u|} \times \texttt{S(u)} \times \texttt{CS(u)}) \tag{6.11}$$

To get a fine analysis, we perform one regression on each dataset and algorithm pair. We model the variability in $\log(p_i)$ and $\log(p_{i|j})$ for a given user by using SIMEX with a variance estimated on each user set of interactions. Empirically, this yields to a model with much more statistically significant effects, and with a larger effect norm.

The AME for the regression of each dataset and algorithm are reported in Figure 6.11, showing that the most important effects come from profile density and $CS(u)$.

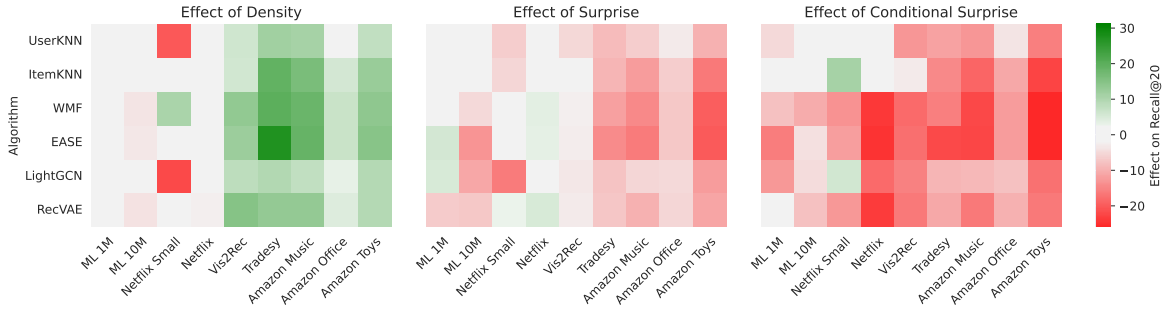


Figure 6.11: Average marginal effect of the variables on the performance. Each value corresponds to the causal variation in `Recall@20` when the variable goes up by 1 standard deviation.

The AME of $|u|$ is non-significant or negative for movie sets while being positive for e-commerce. This can be explained by the difference in density averages between movie e-commerce datasets. Since e-commerce profiles are more sparse, each new item adds useful information for RS. On the other hand, adding items to already dense profiles only adds complexity.

The $S(u)$ measure is the less impactful variable, meaning RS adapt (to some degree) to users with niche tastes. Mean Surprise still holds a negative effect on e-commerce sets. This could be explained by the fact that these datasets have a higher mean $S(u)$. When a user deviates from the popular items, they would buy very rare items, which are not well-modeled by the algorithm.

Mean Conditional Surprise $CS(u)$ greatly impacts the performance negatively in most scenarios. This highlights the importance of the measure in quantifying the difficulty of a user. The distinction between e-commerce and movie datasets is not as clear as for the previous graphs, showing the cross-domain applicability of the measure.

Main Factors Influencing RS's Performance

All other things kept constant:

- The **profile density** can have positive or negative influence based on the domain: for datasets with rich profiles, adding items to the recommendation list adds little useful information. However, in cases with profile scarcity, such as with e-commerce and Vis2Rec, profile density is really important.
- **Surprising users** have some impact on performance, but are on average somewhat handled by the algorithms.
- **Incoherent users** significantly drop RS's performance. Algorithms that perform better on average usually just perform better on coherent users, i.e. on the "easy" part of the data.

6.6.5 Coherence Reproduction

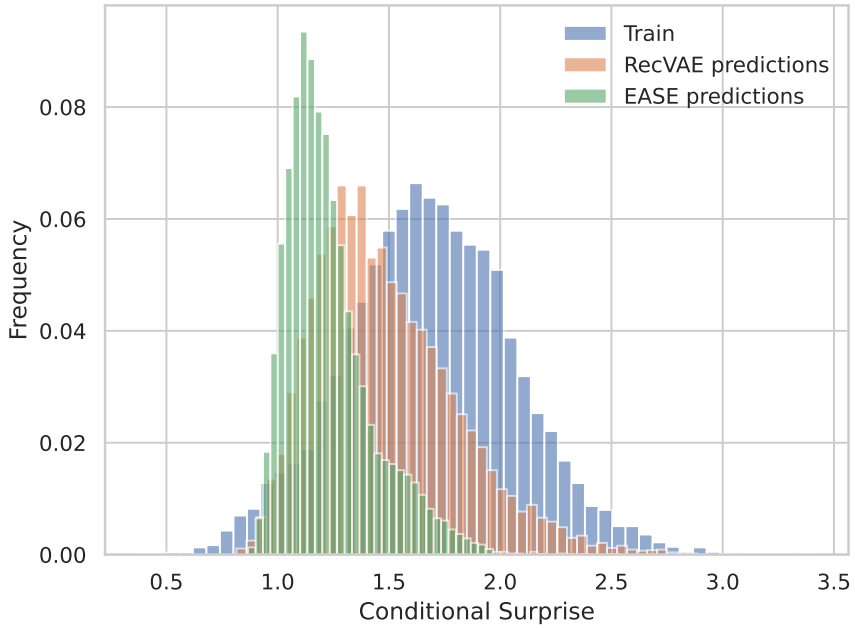


Figure 6.12: Distributions of CS(u) for ML 10M (Train) set and predictions of both RecVAE and EASE

While `Recall@K` and other discrete metrics are broadly used to evaluate RS, the exclusive use of discrete performance metrics is adapted to the field. A good recommendation system might not exactly recommend the test target, as long as its predictions are coherent with the user behavior. Thus, we compare in Figure 6.12 the distribution of $CS(u)$ for a given train set (ML 10M) and for the prediction sets given by `EASE` and `RecVAE` for each user. While `RecVAE` achieves a lower recall score compared to `EASE`, it more closely reproduces the $CS(u)$ distribution of the training set. This observation suggests that `RecVAE` may be better at capturing the underlying coherence patterns of user behavior, even if it doesn't always predict the exact test items. This dual approach to evaluation offers several advantages: it provides a more comprehensive view of recommendation quality, helps identify algorithms that better maintain user behavior patterns, and may reveal strengths overlooked by traditional metrics. This approach encourages the development of more nuanced recommendation strategies.

Coherence Level of the Predictions

Most algorithms, since they do not make "batched" predictions, but compute marginal user-item proximity scores, fail to produce a coherent set of predictions. A notable exception is RecVAE, that predicts an entire item set at once, which is in general more aligned with the coherence level of the training set.

In particular, methods with the best `Recall@K` score are not the best at producing a coherent set of predicted items.

6.6.6 Specialized Models for Coherent Users

Since we showed that $CS(u)$ explains most of RS performance, and since it is well-behaved across datasets, it is a good candidate to segment datasets using the protocol described in Section 6.4.4. From our experiments, we found that the best segmenting strategy was to evaluate on the coherent test users set $\mathcal{D}[0, 0.1]$. We train models on $\mathcal{D}[0, \beta]$ for Netflix, with $\beta \in \{0.1, 0.2, 0.3\}$ independently chosen for each algorithm, and evaluate them on the coherent users.

	ItemKNN	LightGCN	WMF	RecVAE	EASE
Vanilla	0.0	33.8	47.4	46.0	53.0
Spec.	**3.2**	**39.3**	**49.2**	**47.8**	**56.0**

Table 6.4: `Recall@20` on the coherent users of Netflix, for the Vanilla models and the specialized ones, trained on a small coherent subset.

Table 6.4 shows that despite training on at most 30% for the training set, specialized models achieve better performances on coherent users than models trained on the whole dataset. This is probably due to a reduction in the distribution shift between the train and the test set.

6.7 Conclusion

We introduced two information measures for analyzing recommender systems across diverse domains. Our study shows that these measures effectively capture nuanced user behavior patterns that are consistent across different recommendation contexts. These measures provide a domain-agnostic framework for quantifying user coherence, offering insights into the relationship between user behavior and recommendation difficulty. By revealing how user coherence impacts algorithm performance, our approach enables a more nuanced understanding of recommender system dynamics. This work shows the importance of coherence user modeling in recommender systems, potentially leading to adaptive architectures that can better align with diverse user behaviors across various domains.

Not everything that can be counted counts.
Not everything that counts can be counted.

— William Bruce Cameron

7 Conclusion

7.1 Summary and Contributions

Throughout this thesis, we have examined how statistical tools, particularly those adapted from econometrics, can significantly enhance our understanding and evaluation of machine learning systems. While traditional evaluation methods provide valuable baseline measurements, they often fail to capture the nuanced relationships between performance outcomes and model components, training strategies, or data characteristics. Our work shows that by applying rigorous statistical methodologies, we can move beyond simple performance metrics and ablation studies to understand the causal relationships that drive model behavior.

In exemplar-free class-incremental learning (EFCIL), we conducted the first large-scale systematic study of pre-training strategies. Through careful experimental design, we demonstrated that while no single strategy dominates across all scenarios, pre-training with external data consistently improves performance when the domain gap is reasonable. Our analysis revealed that self-supervised pre-training can significantly boost incremental learning performance, particularly when the pre-trained model is fine-tuned on initial classes. We show that pre-training is the main factor influencing incremental accuracy, and we highlight how different metrics are differently impacted by explanatory factors. These findings provide concrete guidance for practitioners implementing EFCIL systems.

In face verification, we made two key contributions to address fairness concerns. First, we introduced a novel controlled generation approach, resulting in the DCFace dataset variants, demonstrating that synthetic data can be leveraged to improve fairness while maintaining competitive accuracy. Second, we developed a new analysis framework combining logit regression and ANOVA that provides deeper insights into the sources and nature of demographic biases, enabling researchers to quantify not just the presence of bias, but its specific impacts across different demographic segments.

In recommendation systems, we introduced Vis2Rec, a new visual dataset for visit recommendation, addressing the scarcity of comprehensive, visually rich datasets in the tourism domain. We also developed novel coherence measures – Surprise and Conditional Surprise – that quantify different aspects of user behavior patterns. These measures provide a domain-agnostic framework for understanding user behavior, offering insights into why certain recommendations succeed

or fail across various algorithms and domains. In particular, we show that the performance of the highest-ranked algorithms is mostly due to their performance on the "easy" users. We also provide guidance on how we can use these metrics to assess batch recommendation coherence, and enhance the performance on some subsets of the data by training on a more aligned training set.

Our work enables rigorous cross-model comparison through careful control of confounding variables and appropriate statistical tests. This was evidenced in comparing EFCIL strategies, face recognition approaches, and recommendation algorithms. Statistical analysis reveals the impact of modeling choices on performance, as demonstrated in our EFCIL study where we could disentangle the effects of pre-training, architecture choice, and dataset characteristics. Our framework provides tools for measuring and understanding biases beyond simple metrics, as shown in our face recognition work where we could analyze bias both in model predictions and latent space representations. Statistical tools enable validation of data modeling approaches, exemplified by our coherence measures in recommender systems which provided insights into user behavior patterns that traditional metrics couldn't capture.

This synthesis of domain-specific contributions and statistical methodology provides a foundation for a more reliable and interpretable evaluation of machine learning systems, enabling researchers and practitioners to make more informed decisions about model development and deployment.

7.2 Perspectives

The methodological advances and empirical findings presented in this thesis open up several promising directions for future research, both in terms of immediate methodological extensions and broader research questions.

7.2.1 Short-term Methodological Extensions

A few possible direct applications or enhancements of our methodology can be developed in the short term.

The impact of Data Augmentation in Self-Supervised Learning

During our research on EFCIL, we trained self-supervised models in various settings. The choice of data augmentations to perform for each algorithm and to apply them on different downstream datasets was a significant source of variability in the performance. To standardize our study, we chose a fixed set of augmentations, but understanding the dependence between self-supervised training and augmentations is a promising research topic. While certain transformations like random cropping appear fundamental, quantifying their individual importance is complicated by strong interdependencies and the computational cost of pre-training. Traditional hyperparameter optimization approaches iterate on known successful combinations, creating correlations that mask individual effects. Our statistical framework could be extended to address this challenge by developing methods for estimating true marginal effects from historically biased optimization data. This would require adapting our causal analysis tools to handle partially observed high-dimensional spaces, potentially combining observational analysis of existing pre-training results with targeted experiments. Such analysis

could reveal which augmentations are truly fundamental versus those that are only effective in specific combinations, guiding more efficient pre-training strategies.

This application is particularly relevant as it combines the key elements of our framework – causal analysis, efficient experimental design, and handling of confounding effects – while addressing a current need in modern deep learning.

Sequential Analysis Through Panel Data Methods

Many machine learning scenarios naturally generate sequential data, such as incremental learning trajectories in EFCIL or user interaction in recommender systems. While current evaluation methods often reduce these sequences to aggregate metrics, econometrics offers sophisticated tools for analyzing panel data that could significantly enhance our understanding of temporal patterns in ML. Panel data methods, which track multiple units (e.g., models or users) over time, provide powerful frameworks for causal inference in sequential settings. For instance, in EFCIL, fixed effects models could help separate the impact of architectural choices from step-specific challenges, controlling for time-invariant model characteristics while identifying which factors truly affect catastrophic forgetting. In recommender systems, dynamic panel models could disentangle how user preferences naturally evolve from how they are influenced by recommendations, accounting for temporal dependencies in user behavior.

These methods are particularly valuable because they can handle both observed and unobserved confounding factors that remain constant over time, enabling more reliable causal inference in sequential settings than traditional before-after comparisons. Adapting these econometric tools to ML evaluation would provide a rigorous framework for analyzing the increasing amount of sequential data in modern ML applications.

From Statistical Control to Causal Understanding

While our statistical framework successfully measures true correlations by controlling for confounding factors, establishing true causation requires explicitly modeling causal relationships. This is particularly relevant in complex ML systems where performance improvements might operate through multiple mechanisms. Mediation analysis, which distinguishes between direct and indirect effects, offers a promising direction for understanding these mechanisms. For instance, in EFCIL, initial accuracy appears to mediate the relationship between pre-training strategy and average incremental accuracy: pre-training might improve incremental learning both directly (through better feature representations) and indirectly (through higher initial accuracy). Formally modeling such mediation effects would help distinguish which improvements come from better starting points versus better adaptation capabilities. A similar analysis could reveal what ethnic features tend to create biases in face recognition systems. In recommender systems, the modeling used in this thesis could be completed with other user-wise behavioral metrics, leading to a better understanding of how behavioral patterns affect performance.

This extension requires developing explicit causal models for ML evaluation, using tools like Directed Acyclic Graphs to formalize our assumptions about how different components influence each other. Such causal modeling would complement our statistical framework by providing a theoretical foundation for interpreting the correlations we observe.

7.2.2 Broader Research Directions

More broadly, several research directions emerge from this work. First, the rise of large language models and multi-modal systems presents new challenges for evaluation, which will be crucial in the next few years as these models continue to develop. Multi-modal and large language models exhibit varying capabilities across different types of tasks, making traditional single-metric evaluations insufficient. Vision-language models like GPT-4V or DALL-E show varying capabilities across different types of tasks. While these models can perform well on many visual tasks, systematic evaluation of their performance across different levels of reasoning (from literal description to cultural understanding) remains an open challenge. Our statistical framework could be extended to analyze such behavioral patterns by formally defining and measuring different types of capabilities and then using regression analysis to understand how architectural choices affect each capability. This could include analyzing how the depth and structure of cross-attention mechanisms influence different types of visual-language tasks, or how the dimensionality and quality of visual feature spaces affect downstream language generation. Such analysis would provide quantitative insights into architectural design choices that are currently often made qualitatively.

The statistical framework could be particularly valuable for analyzing modern architectures that combine multiple specialized components. Consider Mixture-of-Experts (MoE) models, where understanding how experts specialize and interact is crucial for architectural optimization. Statistical analysis could quantify expert specialization patterns through input distribution analysis, while variance decomposition techniques could reveal how different expert combinations contribute to model performance. For instance, in models like Switch Transformer or GLaM, we could analyze whether performance variations across tasks are explained more by expert selection patterns or by expert parameter counts. Similarly, for sparse architectures where only a subset of parameters is active for each input, statistical tools could help quantify the relationship between sparsity patterns and performance, potentially guiding more efficient architecture design. This approach could be particularly valuable for studying scale-dependent behaviors: rather than simply observing that capabilities improve with scale, we could analyze which architectural components contribute most to performance improvements at different scales, providing quantitative guidance for scaling decisions.

While existing methods like ensemble techniques or Bayesian neural networks effectively quantify predictive uncertainty, understanding the sources of this uncertainty remains challenging. Our statistical framework could be extended to analyze how different factors contribute to model uncertainty in critical applications. For instance, in autonomous driving, we could use variance decomposition to understand whether uncertainty in depth prediction stems primarily from physical conditions (lighting, weather), scene complexity (occlusions, object density), or domain shifts (novel scene types).

A similar analysis could be applied to medical imaging, where understanding whether diagnostic uncertainty arises from image quality, rare pathologies, or patient characteristics could guide both model improvement and clinical deployment decisions. This approach would complement existing uncertainty quantification methods by providing actionable insights into the root causes of prediction uncertainty, enabling more targeted improvements in model robustness for safety-critical applications.

7.2.3 Practical Considerations

Integration in frameworks

Integrating statistical tools into machine learning workflows requires extending existing ML infrastructure rather than developing separate statistical frameworks. Modern experiment tracking platforms like Weights & Biases (W&B) and MLflow already provide sophisticated metadata tracking and basic statistical analysis through Bayesian optimization. These platforms could be extended to incorporate more advanced statistical tools directly in their analysis pipeline. For example, W&B's parameter importance analysis could be complemented with formal ANOVA to decompose performance variance across different factors. The platform could automatically perform regression analysis on the collected metrics, providing not just correlation between hyperparameters and performance but actual causal analysis when the experimental design permits it.

One possible enhancement would be the integration of statistical assumption verification. When performing regression analysis, the system could automatically generate diagnostic plots for residual analysis, test for heteroscedasticity, and alert users when statistical assumptions are violated. For ANOVA, the system could verify the normality of residuals within groups and homogeneity of variance between groups. These checks would help ensure that statistical conclusions drawn from experiments are valid and that appropriate corrections (like using robust standard errors or non-parametric tests) are applied when needed.

Results reporting

Standardizing statistical evaluation in machine learning requires clear protocols adapted to each subfield's specific needs. While reporting statistical significance has become common practice, the field needs more rigorous standards for statistical analysis and reporting. Such standards should specify not just which statistical tests to use but how to report their results comprehensively.

For instance, when comparing model architectures, papers should report both within-run variance (from different initializations) and between-run variance (from different hyperparameter settings), with appropriate corrections for multiple comparisons. In fairness evaluation, standards should specify how to aggregate metrics across demographic groups while accounting for different group sizes and intersectional effects. For incremental learning, guidelines should detail how to analyze performance trajectories across learning steps, accounting for the temporal dependence in measurements. These guidelines should be formalized through:

1. Standardized reporting templates for common statistical analyses (ANOVA, regression)
2. Required documentation of assumption verification (e.g., residual plots for regression)
3. Field-specific effect size measures that capture meaningful performance differences

Standardization would enable more reliable meta-analyses and facilitate the reproduction of statistical findings across different studies.

Such integration would make statistical rigor a natural part of ML development rather than an additional burden. This approach would also enable systematic meta-analysis across

experiments and facilitate the standardization of statistical best practices in ML research.

These research directions highlight an important truth: as machine learning systems become more sophisticated, our evaluation methods must evolve accordingly. The statistical tools and frameworks presented in this thesis provide a foundation for this evolution. Still, much work remains to be done to ensure that our evaluation methods keep pace with advances in machine learning technology.

8
Appendices

A Implementation Details and additional Results for EFCIL

A.1 Datasets

We select thirteen datasets containing 100 classes and three datasets containing 1000 classes as follows. The list of datasets is summarized in Table 8.1. The datasets IMN100_1 and IMN100_2 are obtained by randomly sampling 100 classes from ImageNet-21k [Deng et al., 2009] which are not present in ILSVRC [Russakovsky et al., 2015]. Flora is a thematic subset of ImageNet obtained by sampling 100 classes under the concept " flora" without intersection with ILSVRC. We also used 100-classes subsets of WikiArt [Saleh and Elgammal, 2015] (Art100), Casia-align [Yi et al., 2014] (Casia100), Food101 [Bossard et al., 2014] (Food100), FGVC-Aircraft [Maji et al., 2013] (Air100), MTSD [Madani and Yusof, 2016] (MTSD100), Google Landmarks v2 [Weyand et al., 2020b] (Land100), Logo2K [Wang et al., 2020a] (Logo100) and Quickdraw [Ha and Eck, 2017] (Qdraw100). We build two fine-grained subsets from iNaturalist [Van Horn et al., 2018] (2018 version) by selecting (i) amphibia species (Amph100) and (ii) fungi species (Fungi100) which do not intersect with the ILSVRC dataset. Finally, we also use three 1000-classes subsets of Casia-align (Casia1k), Google Landmarks v1 [Noh et al., 2017] (Land1k), and iNaturalist (iNat1k), respectively.

A.2 Comparing performance in multiple scenarios

Factors influencing the average incremental accuracy Let us recall the overall pairwise comparisons from Figure 4.4 and Figure 4.5. We explore the effects of other variables by splitting the data with respect to a studied variable and report the regression results separately. Figure 8.1 presents the results for each target dataset. Figure 8.2 presents the results for each incremental algorithm. Figure 8.3 presents the results depending on the number of classes in the initial state.

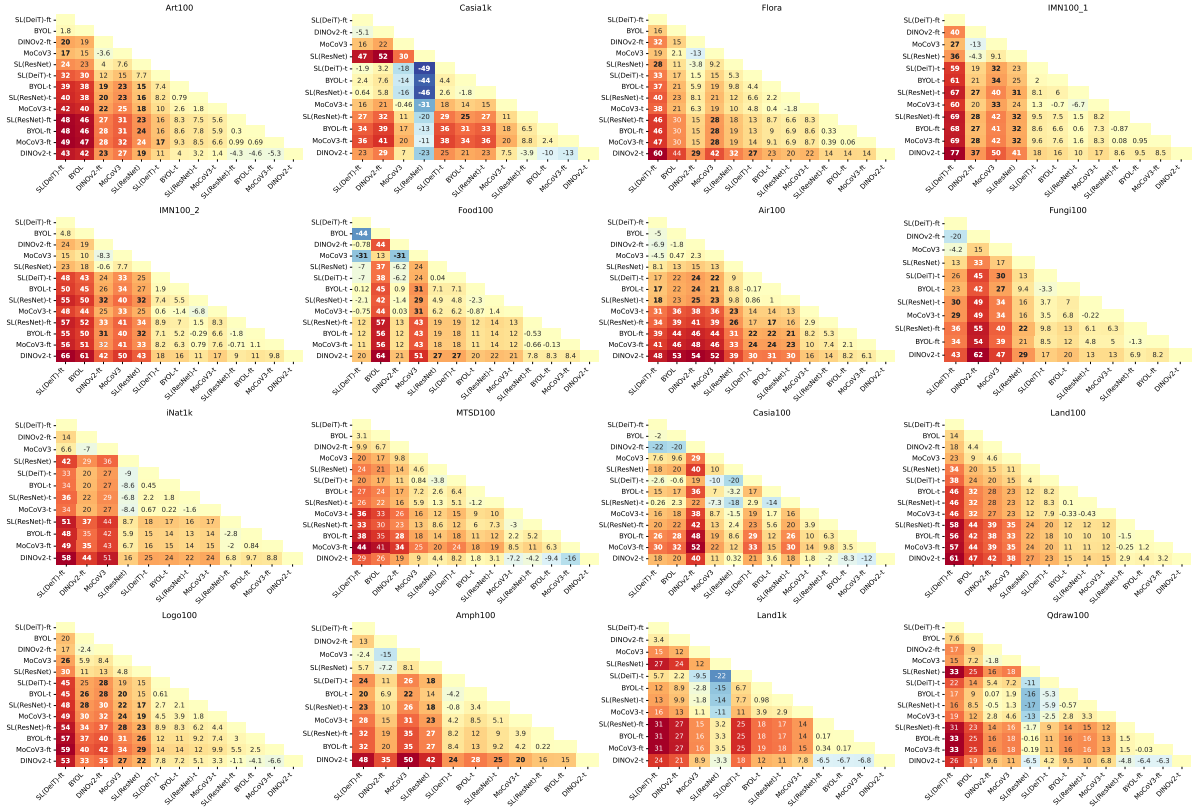


Figure 8.1: Pairwise accuracy gain per dataset. Significant values in **bold** (black or white font).

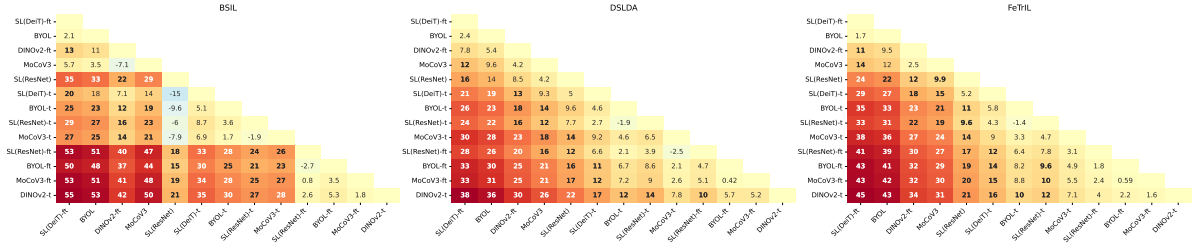


Figure 8.2: Pairwise accuracy gain per EFCIL algorithm. Significant values in **bold** (black or white font).

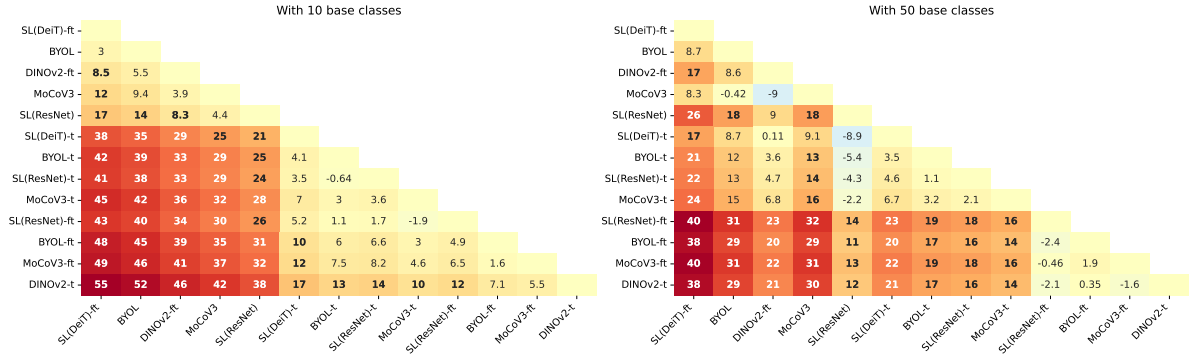


Figure 8.3: Pairwise accuracy gain per proportion of classes in the initial state. Significant values in **bold** (black or white font).

Name	Source	n_{train}	n_{test}	Topic
$IMN100_1$	ImageNet-21k [Deng et al., 2009]	340	60	Random 100 classes
$IMN100_2$	ImageNet-21k [Deng et al., 2009]	340	60	Random 100 classes
Flora	ImageNet-21k [Deng et al., 2009]	340	60	Flora species
Art100	WikiArt [Saleh and Elgammal, 2015]	150	25	Art works (paintings)
Casia100	Casia-align [Yi et al., 2014]	250	50	Faces
Food100	Food101 [Bossard et al., 2014]	750	250	Food
Air100	FGVC-Aircraft [Maji et al., 2013]	80	20	Aircraft
MTSD100	MTSD [Madani and Yusof, 2016]	100	20	Traffic Signs
Land100	Google Landmarks v2 [Weyand et al., 2020b]	300	50	Landmarks
Logo100	Logo2K [Wang et al., 2020a]	80	15	Logos
Qdraw100	Quickdraw [Ha and Eck, 2017]	500	100	Sketches
Amph100	iNaturalist [Van Horn et al., 2018]	300	10	Amphibia species
Fungi100	iNaturalist [Van Horn et al., 2018]	300	10	Fungi species
Casia1k	Casia-align [Yi et al., 2014]	60	28	Faces
Land1k	Google Landmarks v1 [Noh et al., 2017]	374	20	Landmarks
iNat1k	iNaturalist [Van Horn et al., 2018]	300	10	Natural species

Table 8.1: Datasets used in the experiments of Chapter 4.

B Implementation Details and Additional Results for Face Recognition

B.1 Parameters for training and generation

For training the face classifier, we use the Adaface training pipeline Kim et al. [2022]. We apply the same augmentations, crop, and low-resolution augmentations, for all training sets, with an exception on DigiFace , where we also use the augmentation of the authors to reach optimal performances. We perform the training on 4 GPUs with a batch size of 256 (i.e. 64 per GPU), the optimizer is the standard SGD with a learning rate of 0.1 and a momentum of 0.9. We use as a scheduler a multi-step learning rate decay whose milestones are the epochs 12,20,24 and the decay coefficient is 0.1. The training loss is that of Adaface Kim et al. [2022]. The margin parameter m is set to 0.4, and the control concentration constant h to 0.333 as recommended by Kim et al. [2022]. On each training set, the training lasts 60 epochs.

For generating the DCFACE set and its variants, we use the generation pipeline of Kim et al. [2023]. We impose the X_{id} image and the X_{sty} to be of the same demographic group as we found that mismatching is likely to induce non-convergence of the resnet50 model when training on the resulting dataset (in particular when mismatching in gender). Randomly sampling the style image within the CASIA dataset thus results in a non-decreasing loss of the ResNet network. Within the code of Kim et al. [2023], there is a sampling strategy we haven't tested: combining DDPM images with the closer CASIA faces. This approach was and still is, unfortunately, non-usable due to incomplete critical files [1] Moreover, this strategy is not mentioned in the original paper and, since it combines similar CASIA and DDPM faces in a resnet100 latent space, it seems to be in contradiction with what is stated within the ID Image Sampling subsection of Kim et al. [2023]. We thus chose to ignore this strategy, our study being primarily an analysis of fairness and improvement research in this regard.

For all methods, similarly to what the original paper did, we introduce variability within the considered DDPM X_{id} pictures by using a similar F_{eval} model as in Kim et al. [2023]. However,

[1]The provided center_ir_101_adaface_webface4m_faces_webface _112x112.pth file doesn't have a required "similarity_df" field. Also, the dcface_3x3.ckpt file doesn't seem to store the following property: recognition_model.center.weight.data

one should be aware that the Cosine Similarity Threshold might vary depending on the training of the F_{eval} network. We used the network trained on Zhu et al. [2021c] provided by the Adaface Github repository and found 0.6 as an effective threshold to filter similar images. We also get rid of faces wearing glasses with the following solution Birškus [2024].

B.2 Statistical Analysis on FAVCI2D

We present here the results of our statistical analysis on FAVCI2D . Be aware that while the metadata of this dataset contains gender information, it doesn't provide ethnicity. We infer it using FairFace. We consider the prediction of FairFace robust enough to compute macro metrics such as the Diversity metric of the main paper however for a finer study such as ours, it might introduce some uncertainty due to model prediction error (Table 8.2). With that in mind, we still get consistent results for the effects of demographic attributes on the models (Figure 8.4). Our approach shows even more insensitiveness on FAVCI2D than BUPT, by contrast to the results obtained on RFW. The increase of the BUPT-trained model's sensitivity with regard to the inferred labels on FAVCI2D might come from the dataset balancing done on the same labeling system as RFW. Results obtained regarding the TMR (Figure 8.5) and FMR are coherent with the idea that models tend to predict positive outcomes for certain protected ethnical sub-groups. They thus have a high recall for these groups (high TMR and high FMR). With the gender provided by the metadata, we can thus confirm the impact of the balancing on fairness relative to this attribute. While most of the models are sensitive to gender, the model trained on DCFace$_{all}$ DCFace has close to no sensitivity for this attribute, both being close to perfectly balanced concerning gender.

Figure 8.6 shows the result of ANOVA on the distances in the latent space of the FAVCI2D dataset, both on the positive and negative pairs. The results are coherent with the ANOVA computed on RFW. It furthermore highlights the sensitivity of some models' latent space to gender, while our balancing approach allows for more insensitivity about demographic attributes.

B.3 Statistical Analysis on BFW

To tackle the issue of the lack of metadata, in addition to BFW, other alternatives exist such as BFW Robinson et al. [2023] and DemogPairs Hupont and Fernández [2019]. While these datasets provide some ground-truth metadata, they are composed of significantly fewer identities compared to datasets like FAVCI2D or RFW. This is a limitation of our analysis: Having too few identities might bring instability within Anova or marginal effect studies due to redundancy. We report the results obtained with BFW on as similar number of pairs as RFW and FAVCI2D (24k), meaning every single identity appears in around 30 evaluated pairs. The impact of the number of identities within benchmarking should be studied in future works as this might affect the obtained analysis of performance and fairness.

Figure 8.9 shows the ANOVA analysis performed on BFW. As before, on the negative image pairs, our conditional generation methods greatly reduces the variance explained by the sensitive attributes.

Figures 8.8 and 8.7 present the marginal effects of the attributes, respectively, on TMR and FMR. As we see, the fairness gain mostly comes from a fairer FMR between ethnicities: the FMR of the Asian and Black subgroups are 8 and 12 points higher than for the White subgroup in the original DCFace , and become non-significant with DCFace$_{all}$. For the TMR, however, just as for RFW, becomes slightly more unfair between ethnicities. Still, as shown in Table 2 of the paper, on all fairness metrics except EOR, our method outperforms the other synthetic data approaches on BFW.

ethnicity	Black	White	East-Asian	Indian	Latino-Hispanic	Middle-Eastern	South-Asian
Prediction accuracy	0.863	0.777	0.784	0.724	0.581	0.631	0.641

Table 8.2: FairFace model accuracy when inferring on the Fairface validation set. Available Metadata only provides the race7 variable ground truth while we are considering the race variable (whose values are White, Black, Asian, and Indian). The robustness of the model for this latter should be thus greater.

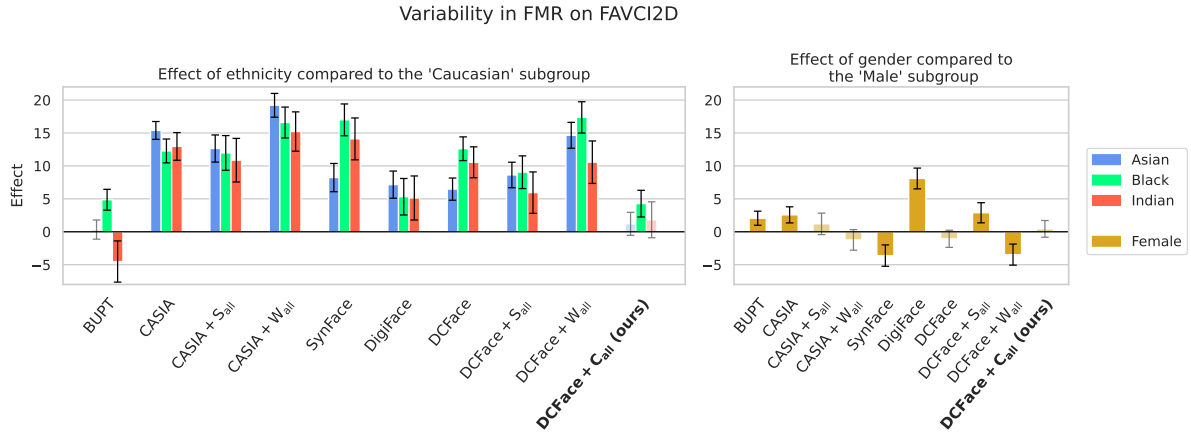


Figure 8.4: Marginal effect on FMR (lower is better) for each method compared to the unprotected group. Analysis done on FAVCI2D

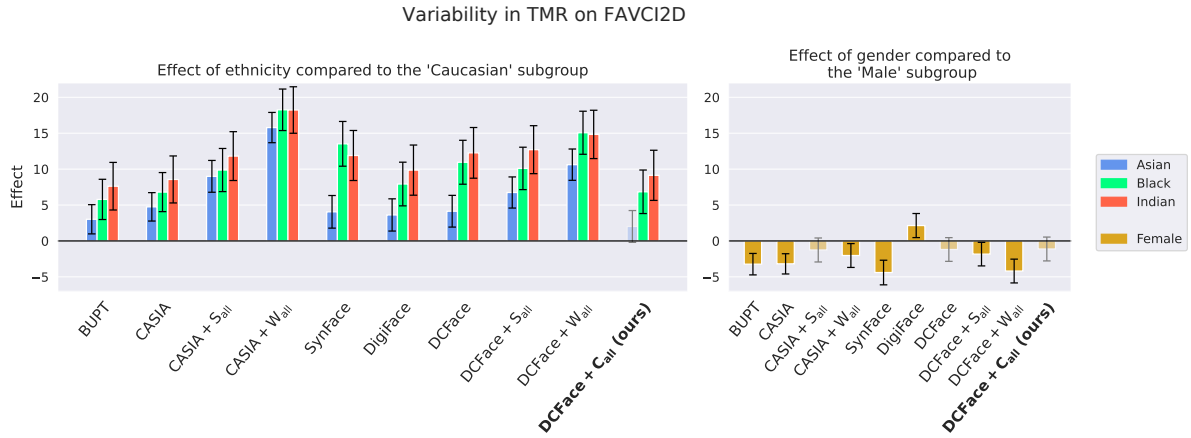


Figure 8.5: Marginal effect on TMR (lower in absolute is better) for each method compared to the unprotected group. Analysis done on FAVCI2D

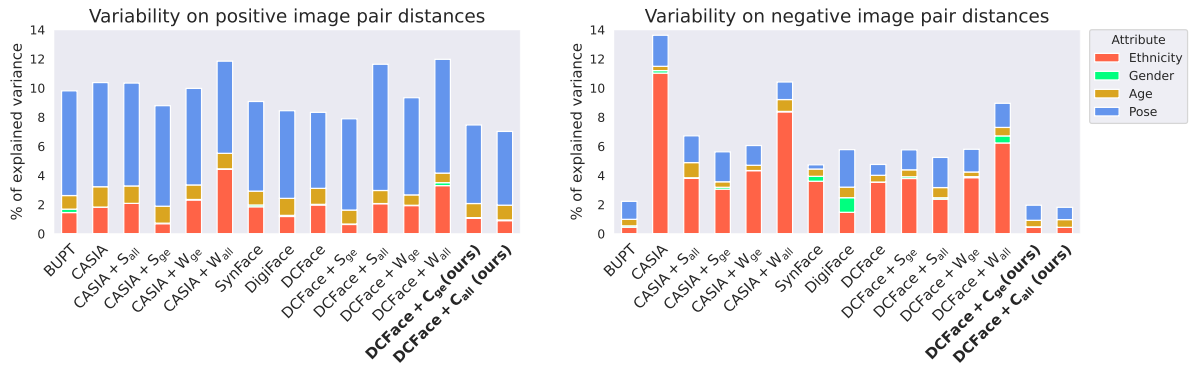


Figure 8.6: ANOVA results on FAVCI2D : total height corresponds to R^2, the explained variance by the variables. Each bar is decomposed into multiple η^2, i.e. the individual contributions to the variance

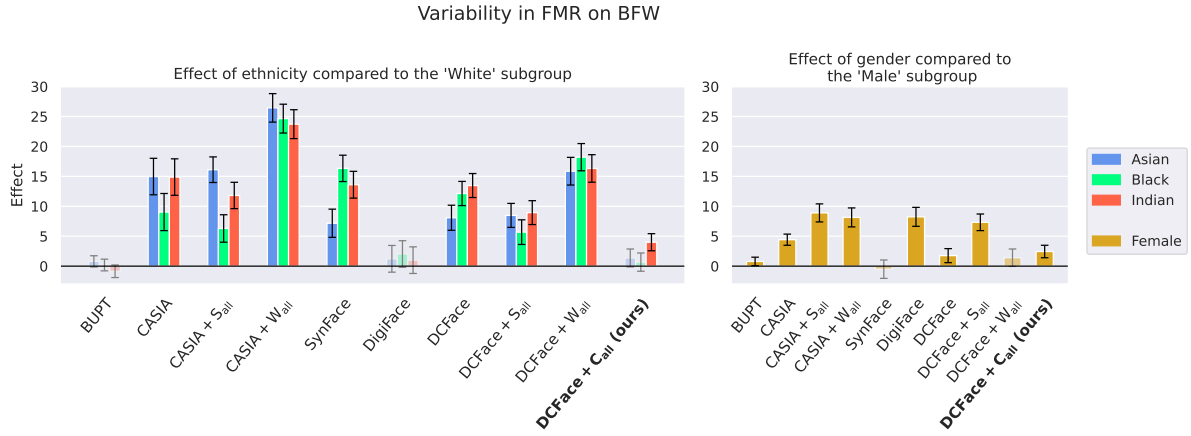


Figure 8.7: Marginal effect on FMR (lower is better) for each method compared to the unprotected group. Analysis done on BFW

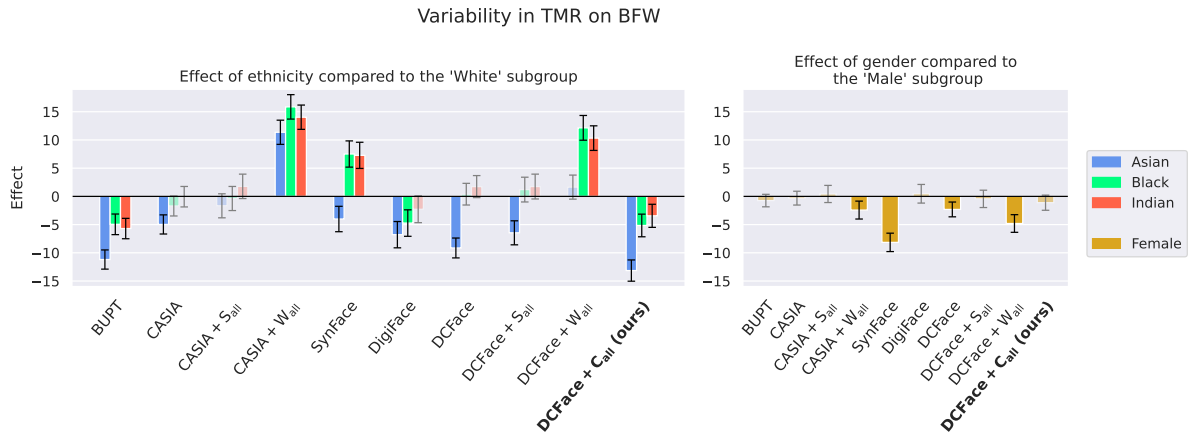


Figure 8.8: Marginal effect on TMR (lower in absolute is better) for each method compared to the unprotected group. Analysis done on BFW

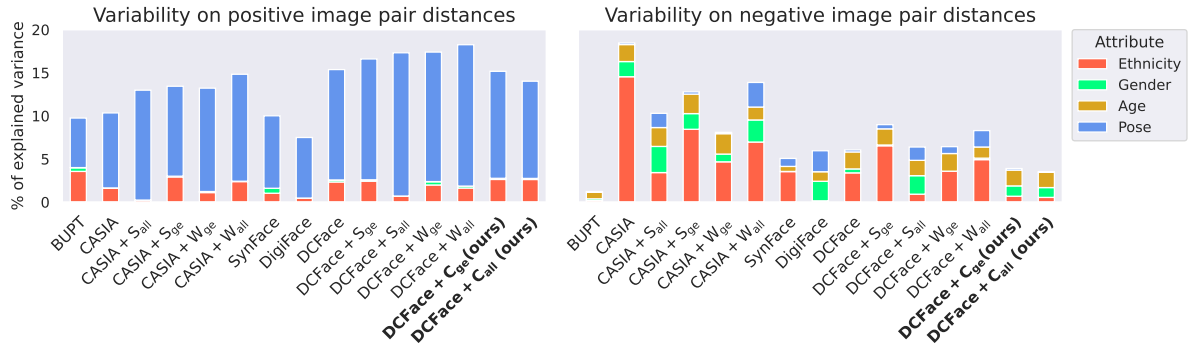


Figure 8.9: ANOVA results on BFW: total height corresponds to R^2, the explained variance by the variables. Each bar is decomposed into multiple η^2, i.e. the individual contributions to the variance

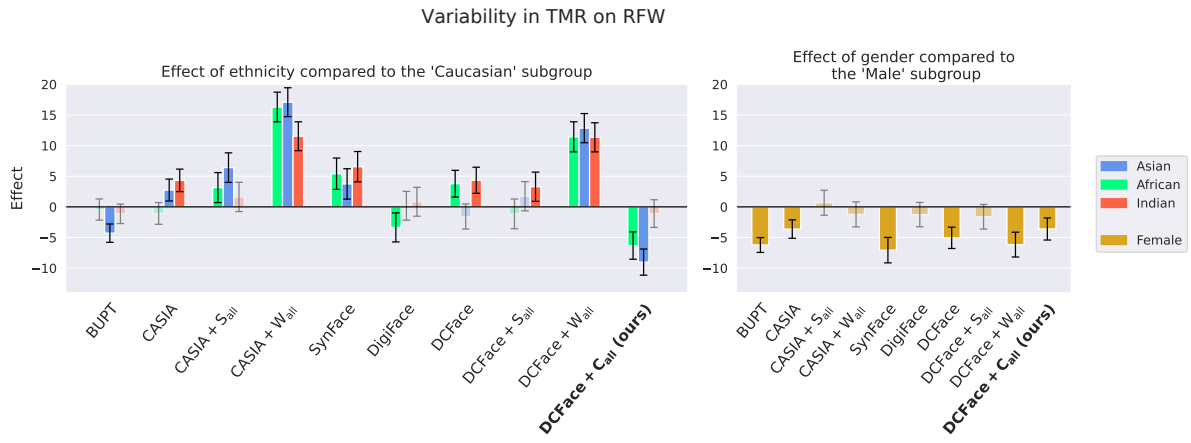


Figure 8.10: Marginal effects on TMR (lower in absolute is better) for each method compared to the unprotected group. Analysis done on RFW

B.4 Datasets Images examples

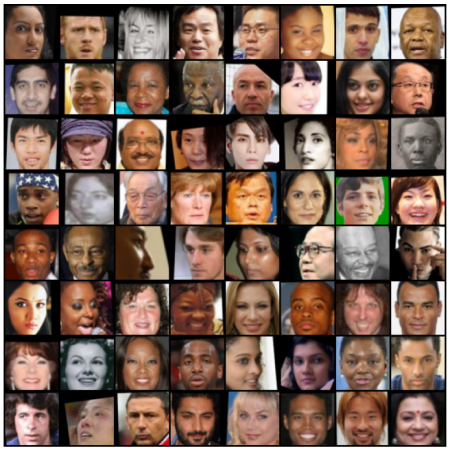
(a) Examples of images within our proposed DCFace$_{all}$ approach. We notice a greater diversity of images.

(b) Examples of images generated with the original DCFace Kim et al. [2023] pipeline

(c) Examples of images generated with the SynFace pipeline Qiu et al. [2021a]

(d) Examples of images within the DigiFace dataset Bae et al. [2023]

(e) Examples of images within the CASIA dataset Yi et al. [2014]



(f) Examples of images within the BUPT dataset Wang et al. [2021]

C Implementation Details and additional Results for Recommender Systems

C.1 Theoretical Elements

Orthogonality of the measures

Since $\log(p_{i|i}) = 0$, effectively, the pairs (i, i) do not intervene in the definition of $SC(u)$. In fact, if we denote by $PS(u)$ the **Mean Pair Surprise** by replacing $p_{i|j}$ by $p_{i,j}$ in the definition of $CS(u)$, then we have:

$$CS(u) = PS(u) - S(u) \tag{8.1}$$

Effectively, we remove the effect of the surprise at the first order from the surprise of the pairs.

Proof of Proposition 1

Let $p_{ui} = \mathbb{P}(x_{ui} = 1 | |u| > 0)$. First, consider that:

$$\begin{aligned}\mathbb{E}_{\pi_u^{\geq 1}}[\widetilde{S}(u)] &= \frac{1}{m}\sum_{i=1}^{m}\log(p_i^*)\mathbb{E}_{\pi_u^{\geq 1}}[x_{ui}] \\ &= \frac{1}{m}\sum_{i=1}^{m}\log(p_i^*)p_{ui}\end{aligned}$$

since x_{ui} is a Bernoulli variable. Then for any user u with $|u| > 0$:

$$\begin{aligned}\mathbb{E}_{\pi_u^{\geq 1}}[S(u)] &= -\sum_{i}^{m}\log(p_i^*)\mathbb{E}_{\pi_u^{\geq 1}}\left[\frac{x_{ui}}{|u|}\right] \\ &= -\sum_{i}^{m}\log(p_i^*)\mathbb{E}_{x_{ui}}\left[\mathbb{E}_{\pi_u^{\geq 1}}\left[\frac{x_{ui}}{|u|}\middle| x_{ui}\right]\right] \\ &= -\sum_{i}^{m}\log(p_i^*)p_{ui}\mathbb{E}_{\pi_u^{\geq 1}}\left[\frac{x_{ui}}{|u|}\middle| x_{ui} = 1\right] \\ &= -\sum_{i}^{m}\log(p_i^*)p_{ui}\mathbb{E}_{\pi_u^{\geq 1}}\left[\frac{1}{|u|}\middle| x_{ui} = 1\right]\end{aligned}$$

where the second line is a consequence of the law of iterated expectations. Then, since $x \mapsto 1/x$ is convex on $\mathbb{R}_+^*$, by Jensen inequality:

$$\begin{aligned}\mathbb{E}_{\pi_u^{\geq 1}}[S(u)] &\geq -\sum_i^m \log(p_i^*) p_{ui} \frac{1}{\mathbb{E}_{\pi_u^{\geq 1}}[|u|]} \\ &\geq \mathbb{E}_{\pi_u^{\geq 1}}[\widetilde{S}(u)] \frac{m}{\mathbb{E}_{\pi_u^{\geq 1}}[|u|]}\end{aligned}$$

which directly gives the left-hand side of Proposition 1. The right-hand-side follows from the fact that $|u| = \sum_i x_{ui}$. So, for any random events ω_0, ω_1 that only differ in $x_{ui}(\omega_0) = 0$ and $x_{ui}(\omega_1) = 1$, we have $\frac{1}{|u|}(\omega_0) \geq \frac{1}{|u|}(\omega_1)$. Taking the expected value from both sides gives:

$$\begin{aligned}\mathbb{E}_{\pi_u^{\geq 1}}[S(u)] &\leq -\sum_i^m \log(p_i^*) p_{ui} \mathbb{E}_{\pi_u^{\geq 1}}\left[\frac{1}{|u|}\right] \\ &\leq -\mathbb{E}_{\pi_u^{\geq 1}}\left[\frac{1}{|u|}\right] \sum_i^m \log(p_i^*) p_{ui} \\ &\leq \mathbb{E}_{\pi_u^{\geq 1}}\left[\frac{m}{|u|}\right] \mathbb{E}_{\pi_u^{\geq 1}}[\widetilde{S}(u)] \quad \square\end{aligned}$$

For $CS(u)$, the convexity of $x \mapsto 1/x^2$ on $\mathbb{R}_+^*$ and the fact that $\mathbb{E}_{\pi_u^{\geq 1}}\left[\frac{1}{|u|^2}|x_{ui} = 1\right] \leq \mathbb{E}_{\pi_u^{\geq 1}}\left[\frac{1}{|u|^2}\right]$ proves the bounds.

Discussion about the Poisson model for $|u|$

The Poison distribution, also known as "the law of rare events", is an adapted model to count the frequency of events that occur rarely. In particular, in recommendation data, users consume only a small fraction of the possible items, giving a motivation for Poisson modelization. Moreover, if we choose a finer description of the user's choices, for example, assigning known oracle probability p_{ui} of observing the item i in the user's u set, then π_u becomes, by definition, a multivariate Bernoulli distribution. Therefore, $|u| = \sum x_{ui}$ is, by definition, a Poisson-Binomial variable, which is well-approximated by a Poisson distribution of parameter $\lambda = \sum p_{ui}$, in virtue of Le Cam's theorem Cam [1960].

Proof of Proposition 2

If X is a poison variable of parameter $\lambda > 0$, then :

$$\forall k \in \mathbb{N}, \mathbb{P}[X = k] = \frac{e^{-\lambda}\lambda^k}{k!}$$

which yields:

$$\begin{aligned}\forall k \in \mathbb{N}, \mathbb{P}[X = k | X > 0] &= \frac{\mathbb{P}[X > 0 | X = k]\mathbb{P}[X = k]}{1 - \mathbb{P}[X = 0]} \\ &= \frac{\mathbb{1}(k > 0)}{1 - e^{-\lambda}} \mathbb{P}[X = k] \\ &= \mathbb{1}(k > 0) \frac{e^{-\lambda}\lambda^k}{(1 - e^{-\lambda})k!}\end{aligned}$$

Then, we can consider the fact that we have:

$$\forall k \in \mathbb{N}^*, \frac{1}{k} \leq \frac{2}{k+1}$$

Taking the expectancy conditioned on $X > 0$ (i.e. $X \geq 1$) gives us:

$$\begin{aligned}
\mathbb{E}_{X\geq 1}\left[\frac{1}{X}\right] &\leq \mathbb{E}_{X\geq 1}\left[\frac{2}{X+1}\right] \\
&\leq 2\sum_{k=1}^{\infty}\frac{\mathbb{P}[X=k|X>0]}{k+1} \\
&\leq 2\sum_{k=1}^{\infty}\frac{e^{-\lambda}}{1-e^{-\lambda}}\frac{\lambda^k}{(k+1)k!} \\
&\leq \frac{2e^{-\lambda}}{1-e^{-\lambda}}\frac{1}{\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k+1}}{(k+1)!} \\
&\leq \frac{2e^{-\lambda}}{1-e^{-\lambda}}\frac{1}{\lambda}(e^{\lambda}-1-\lambda) \\
&\leq \frac{2(1-e^{-\lambda})}{\lambda} \\
&\leq \frac{2}{\mathbb{E}_{\geq 1}[X]} \quad \square
\end{aligned}$$

To get a similar bound for $CS(u)$, i.e bounding $\mathbb{E}_{X\geq 1}[X^2]\mathbb{E}_{X\geq 1}\left[\frac{1}{X^2}\right]$ we first find a constant K such that :

$$\begin{aligned}
\forall k \in \mathbb{N}^*, \frac{1}{k^2} &\leq \frac{K}{(k+1)(k+2)} \\
0 &\leq (1-\frac{1}{K})k^2 - \frac{3}{K}k - \frac{2}{K}
\end{aligned}$$

The biggest root of the RHS is given by $\frac{3+\sqrt{1+8K}}{2K-2}$, which is smaller than 1 for $K \geq 6$. The rest of the proof follows the same calculations as for $S(u)$.

C.2 Empirical Bounds

We empirically estimate an upper bound for $\mathbb{E}_{X\geq 1}[X]\mathbb{E}_{X\geq 1}\left[\frac{1}{X}\right]$ and $\mathbb{E}_{X\geq 1}[X^2]\mathbb{E}_{X\geq 1}\left[\frac{1}{X^2}\right]$ for a Poisson variable by varying λ. The results are presented in Figures 8.13 and 8.14.

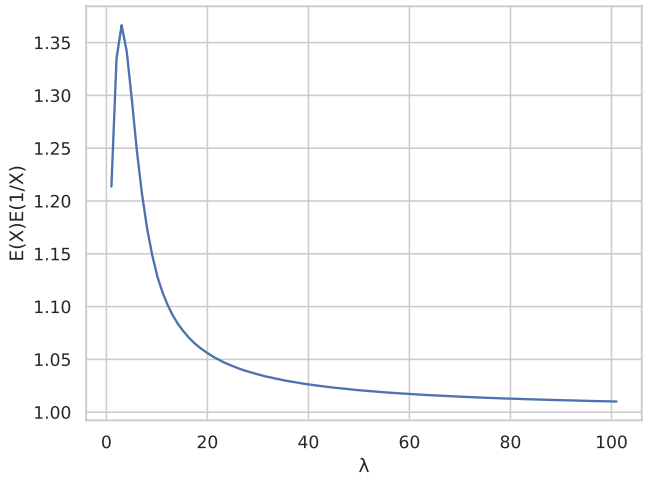


Figure 8.13: Monte-Carlo estimation of $\mathbb{E}_{X\geq 1}[X]\mathbb{E}_{X\geq 1}\left[\frac{1}{X}\right]$

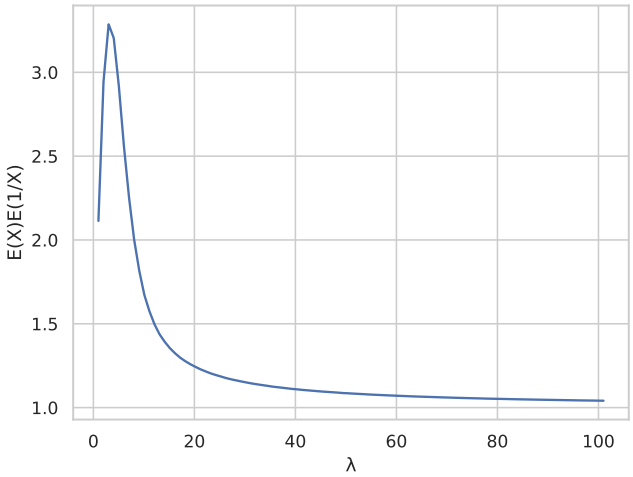


Figure 8.14: Monte Carlo estimation of $\mathbb{E}_{X\geq 1}[X^2]\mathbb{E}_{X\geq 1}\left[\frac{1}{X^2}\right]$

C.3 Data Processing and Training

Data Processing

We found that in most implementations of recommender data pre-processing, the data was filtered by first removing the items that were consumed by less than k users, then removing the users that consumed less than k items. This is a problem since, with this ordering, the remaining items could be consumed less than k times. Thus, the k-core is extracted using the algorithm 1.

Algorithm 1 k-core extraction

Input: A DataFrame `df` in format `['user'; 'item']`
Parameters: An integer k
Output: A k-core for `df`

Let $n_1 =$ `df.groupby('user').len().min()`
Let $n_2 =$ `df.groupby('item').len().min()`
while $\max(n_1, n_2) > k$ **do**
    `df` $\leftarrow$ `df.groupby('user').filter(len(x) >= k )`
    `df` $\leftarrow$ `df.groupby('item').filter(len(x) >= k )`
    $n_1 \leftarrow$ `df.groupby('user').len().min()`
    $n_2 \leftarrow$ `df.groupby('item').len().min()`
end while
return `df`

Optimal Hyperparameters

The optimal hyperparameters found by `optuna` on 50 runs for each combination of dataset and algorithm, optimizing on the `Recall@20` of the validation set, are presented in Table 8.3.

C.4 Additional Results

Experimental Properties of $S(u)$ and $CS(u)$

We show a non zero relationship between $S(u)$ or $CS(u)$, and $|u|$, in Figure 8.15 and 8.16. As we see, the relationship is much less clear than for $\widetilde{S}(u)$ or $\widetilde{CS}(u)$, which are almost perfectly linear or quadratic. Graphs 8.17 and 8.18 show the standard deviation dependence of $S(u)$ and

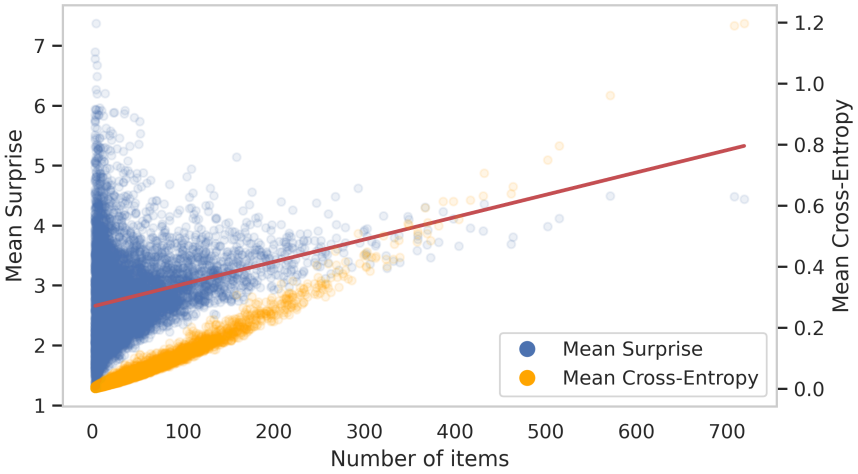


Figure 8.15: $S(u)$ and $\widetilde{S}(u)$ vs $|u|$ on the Netflix Small dataset, along with the linear fit of $S(u)$ on $|u|$

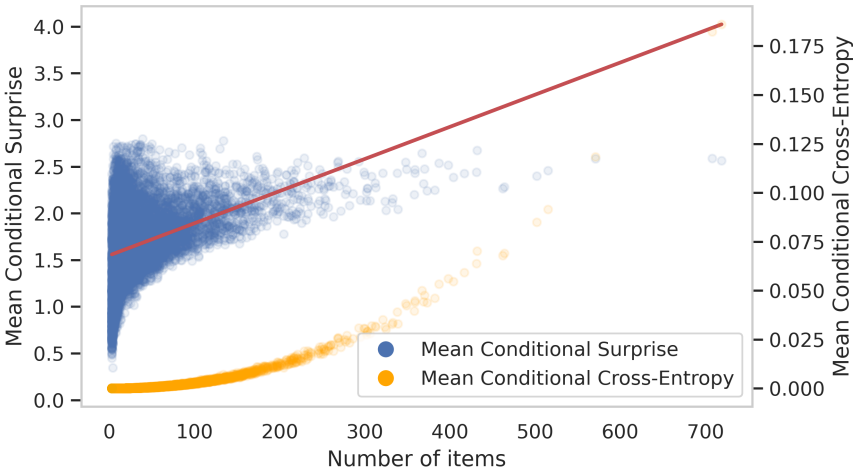


Figure 8.16: $CS(u)$ and $\widetilde{CS}(u)$ vs $|u|$ on the Netflix Small dataset, along with the linear fit of $CS(u)$ on $|u|$

$CS(u)$ on $|u|$, where the deviation is the standard deviation of the surprises (and conditional surprises) values used to compute the sums $S(u)$ and $CS(u)$. As $|u|$ increases, the standard deviation values stabilize, slightly increasing with with $|u|$. As we see, users with few items can consume either very low-variance items or, on the contrary, have a very erratic behavior.

C.5 Impact on Performance

All regression are run in R, using the `glm` function, with a `binomial` law with `logit` link. The `simex` package is used to incorporate the variance of the variables, and `margins` to get the marginal effects.

The choice of the regression to make, in particular which dependencies between variables (such as $S(u) \times CS(u)$), was motivated primarily by the model with the lowest AIC score. Consistently, the model with all the product variables met our criterion.

We also found an important aspect in modeling these logit regressions was to put a threshold on the variable $|u|$. In accordance with what we stated in the main chapter for the impact of our measures on performances, this can be justified by the fact that up until a certain point, adding

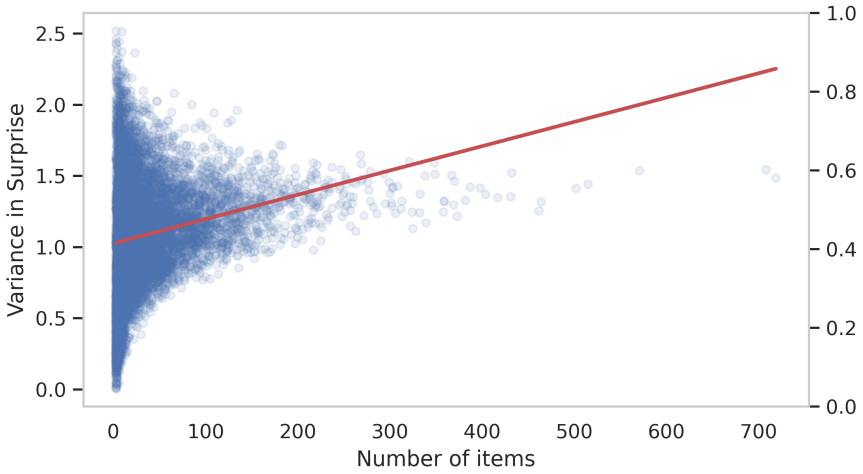


Figure 8.17: $std(S(u))$vs $|u|$ on the Netflix Small dataset, along with the linear fit

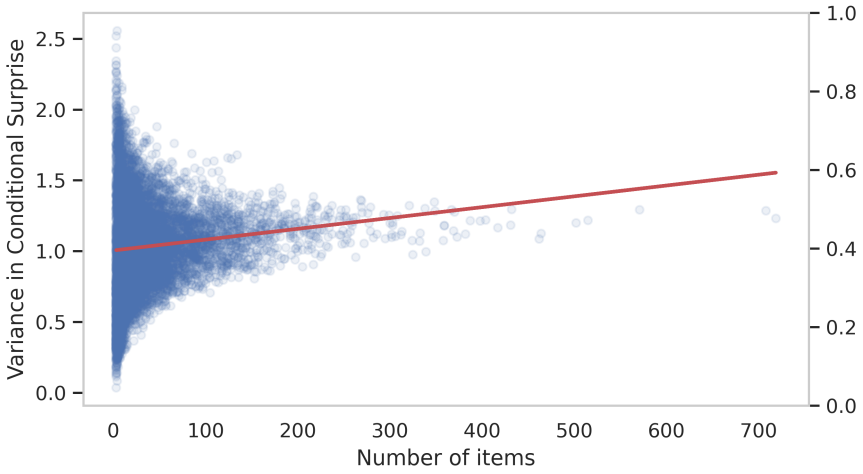


Figure 8.18: $std(CS(u))$ vs $|u|$ on the Netflix Small dataset, along with the linear fit.

more items to the item set of a user can only help us to cover all their tastes. Once all the interests of a user are well represented in their item set, then we expect $|u|$ to have less importance. Indeed, thresholding $|u|$ led to models with a lower AIC but also removed the heteroscedasticity of the residuals, i.e., the dependence between the variance of the errors and the predictor variables.

Coherence Reproduction

As mentioned in the Coherence Reproduction section of the main chapter, Figure 8.19 shows the correlation between the information measures of the user u, and the information measures evaluated on their predicted set $\hat{u}$. As we see, most algorithms (except `MostPop`), generate predictions that are more correlated with the Mean Surprise level of the known set of the user. One notable exception is `ItemKNN`, which quite poorly reproduces the Mean Surprise level of the movie datasets. This can be linked to the overall poor performance of `ItemKNN` on these datasets, with scores of recall way smaller than those of `MostPop`. For the Conditional Surprise, we find that Deep Learning methods, such as `LightGCN` and `RecVAE`, have a stronger correlation between the input and the predictions than more traditional methods (especially for datasets

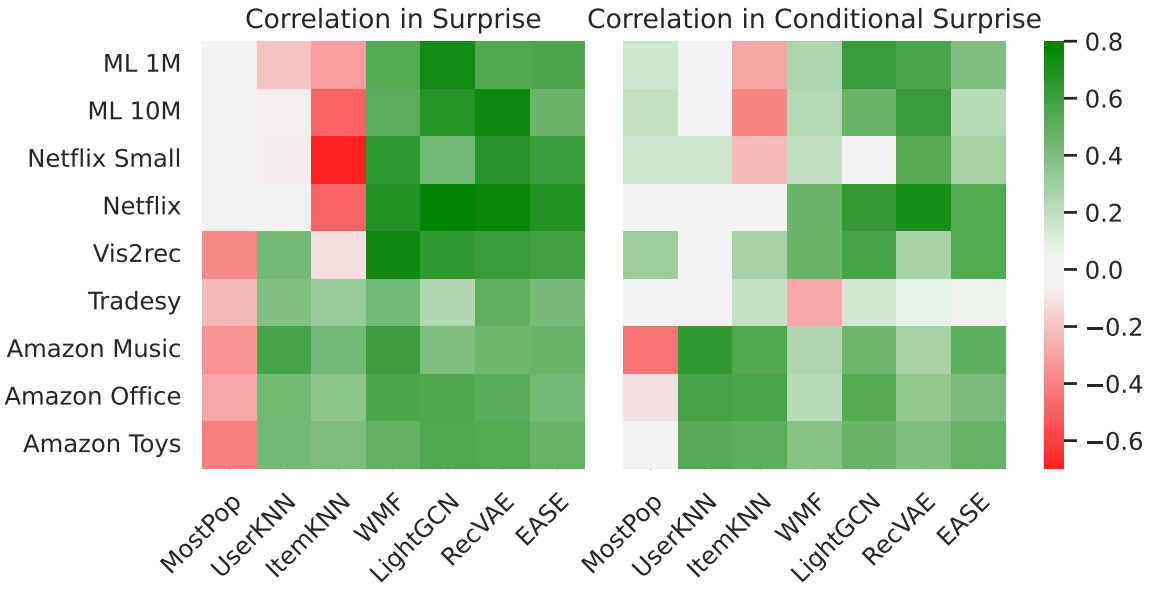


Figure 8.19: Correlation between $S(u)$ and $S(\hat{u})$; and $CS(u)$ and $CS(\hat{u})$, where $\hat{u}$ is the predicted item set for u.

with a lot of users). For these methods, the predictions are computed in a highly non-linear fashion, which could enable more complex interaction modelization.

Dataset	UserKNN	ItemKNN	WMF	LightGCN	RecVAE	EASE
ML 1M	k: 926 centered: 1 sim: cosine	k: 170 centered: 0 sim: cosine	b: 9.50e-01 batch_size: 16000 k: 518 lambda_u: 4.92e-03 lambda_v: 2.30e-04	batch_size: 16000 emb_size: 64 epochs: 1000 num_layers: 3	batch_size: 1024 hidden_dim: 919 latent_dim: 254 n_epochs: 100	lamb: 200
ML 10M	k: 705 centered: 1 sim: pearson	k: 594 centered: 1 sim: cosine	b: 9.07e-01 batch_size: 16000 k: 878 lambda_u: 3.54e-02 lambda_v: 5.54e-02	batch_size: 16000 emb_size: 64 epochs: 1000 num_layers: 3	batch_size: 1024 hidden_dim: 611 latent_dim: 541 n_epochs: 100	lamb: 147
Netflix S.	k: 930 centered: 1 sim: cosine	k: 610 centered: 1 sim: cosine	b: 9.52e-01 batch_size: 16000 k: 689 lambda_u: 3.91e-02 lambda_v: 4.08e-03	batch_size: 16000 emb_size: 64 epochs: 1000 num_layers: 3	batch_size: 1024 hidden_dim: 349 latent_dim: 696 n_epochs: 100	lamb: 787
Netflix	NaN	k: 841 centered: 0 sim: cosine	b: 7.20e-01 batch_size: 16000 k: 705 lambda_u: 8.36e-03 lambda_v: 2.79e-04	batch_size: 16000 emb_size: 32 epochs: 200 num_layers: 1	batch_size: 1024 hidden_dim: 778 latent_dim: 164 n_epochs: 100	lamb: 298
Vis2Rec	k: 547 centered: 0 sim: pearson	k: 814 centered: 1 sim: cosine	b: 9.25e-01 batch_size: 16000 k: 702 lambda_u: 1.81e-02 lambda_v: 3.32e-03	batch_size: 16000 emb_size: 64 epochs: 1000 num_layers: 3	batch_size: 1024 hidden_dim: 589 latent_dim: 614 n_epochs: 100	lamb: 63
Tradesy	k: 9 centered: 0 sim: cosine	k: 773 centered: 1 sim: pearson	b: 9.2e-01 batch_size: 16000 k: 485 lambda_u: 1.02e-02 lambda_v: 1.02e-04	batch_size: 16000 emb_size: 64 epochs: 1000 num_layers: 3	batch_size: 1025 hidden_dim: 325 latent_dim: 464 n_epochs: 101	lamb: 86
A. Music	k: 25 centered: 0 sim: cosine	k: 568 centered: 1 sim: pearson	b: 7.76e-01 batch_size: 16000 k: 588 lambda_u: 4.96e-02 lambda_v: 1.65e-04	batch_size: 16000 emb_size: 64 epochs: 1000 num_layers: 3	batch_size: 1024 hidden_dim: 217 latent_dim: 336 n_epochs: 100	lamb: 31
A. Office	k: 53 centered: 1 sim: pearson	k: 837 centered: 1 sim: cosine	b: 3.91e-02 batch_size: 16000 k: 985 lambda_u: 8.93e-01 lambda_v: 4.51e-02	batch_size: 16000 emb_size: 64 epochs: 1000 num_layers: 3	batch_size: 1024 hidden_dim: 479 latent_dim: 585 n_epochs: 100	lamb: 49
A. Toys	k: 159 centered: 0 sim: pearson	k: 524 centered: 0 sim: cosine	b: 2.34e-02 batch_size: 16000 k: 1000 lambda_u: 1.15e-02 lambda_v: 2.06e-04	batch_size: 16000 emb_size: 64 epochs: 1000 num_layers: 3	batch_size: 1024 hidden_dim: 660 latent_dim: 492 n_epochs: 100	lamb: 35

Table 8.3: Optimal hyperparameters on each dataset with each algorithm, optimizing the `Recall@20` in the leave-one-out setup.

Bibliography

General data protection regulation. https://gdpr.eu/. Accessed: 2020-11-12.

S. Abnar, M. Dehghani, B. Neyshabur, and H. Sedghi. Exploring the limits of large scale pre-training. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=V3C8p78sDa.

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Afoudi et al. Collaborative filtering recommender system. *Advances in intelligent systems and computing*, 2018. doi: 10.1007/978-3-030-11928-7_30.

A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. A reductions approach to fair classification. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/agarwal18a.html.

A. Agarwal, M. Dudik, and Z. S. Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 120–129. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/agarwal19d.html.

T. Ahmad, A. R. Dhamija, S. Cruz, R. Rabinowitz, C. Li, M. Jafarzadeh, and T. E. Boult. Few-shot class incremental learning leveraging self-supervised features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3900–3910, 2022.

H. Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY, 1998. ISBN 978-1-4612-1694-0. doi: 10.1007/978-1-4612-1694-0_15. URL https://doi.org/10.1007/978-1-4612-1694-0_15.

V. Albiero and K. Bowyer. Is face recognition sexist? no, gendered hairstyles and biology are, 08 2020.

V. Albiero, K. Zhang, and K. W. Bowyer. How does gender balance in training data affect face recognition accuracy? In *2020 ieee international joint conference on biometrics (ijcb)*, pages 1–10. IEEE, 2020.

V. Albiero, K. Zhang, M. C. King, and K. W. Bowyer. Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. *IEEE Transactions on Information Forensics and Security*, 17:127–137, 2021.

B. Alhijawi, A. Awajan, and S. Fraihat. Survey on the objectives of recommender systems: Measures, solutions, evaluation methodology, and new perspectives. *ACM Computing Surveys*, 2022. doi: 10.1145/3527449.

R. Aljundi, K. Kelchtermans, and T. Tuytelaars. Task-free continual learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 11254–11263. Computer Vision Foundation / IEEE, 2019.

P. D. Allison. Measures of Fit for Logistic Regression. 2014.

G. Amato, F. Falchi, C. Gennaro, F. V. Massoli, N. Passalis, A. Tefas, A. Trivilini, and C. Vairo. Face verification and recognition for digital forensics and information security. In *2019 7th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–6. IEEE, 2019.

X. Amatriain, J. M. Pujol, and N. Oliver. I like it... i like it not: Evaluating user ratings noise in recommender systems. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 247–258. Springer, 2009.

S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz. Guidelines for human-ai interaction. In *CHI 2019*. ACM, May 2019. URL `https://www.microsoft.com/en-us/research/publication/guidelines-for-human-ai-interaction/`. CHI 2019 Honorable Mention Award.

A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Y. A. Hawalah, and A. Hussain. Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. *IEEE Access*, 4:7940–7957, 2016. URL `https://api.semanticscholar.org/CorpusID:6526065`.

D. Amodei and D. Hernandez. Ai and compute, 2018. URL `https://openai.com/research/ai-and-compute`.

D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

V. W. Anelli, A. Bellogín, A. Ferrara, D. Malitesta, F. A. Merra, C. Pomo, F. M. Donini, and T. D. Noia. Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation. 2021. doi: 10.1145/3404835.3463245.

J. D. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009. ISBN 9780691120348. URL `http://www.jstor.org/stable/j.ctvcm4j72`.

S. Arora, N. Cohen, and E. Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization, 2018. URL `https://arxiv.org/abs/1802.06509`.

G. Bae, M. de La Gorce, T. Baltrušaitis, C. Hewitt, D. Chen, J. Valentin, R. Cipolla, and J. Shen. Digiface-1m: 1 million digital face images for face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3526–3535, 2023.

D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate, 2016. URL `https://arxiv.org/abs/1409.0473`.

S. Barattin, C. Tzelepis, I. Patras, and N. Sebe. Attribute-preserving face dataset anonymization via latent code optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8001–8010, 2023.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

J. T. Barron. A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4331–4339, 2019.

B. Barz and J. Denzler. Do we train on test data? purging cifar of near-duplicates. *Journal of Imaging*, 6(6):41, 2020.

Z. Batmaz, A. Yurekli, A. Bilge, and C. Kaleli. A review on deep learning for recommender systems: challenges and remedies. *Artificial Intelligence Review*, 52:1–37, 2019.

H. Behzadi-Khormouji and J. Oramas. A protocol for evaluating model interpretation methods from visual explanations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1421–1429, 2023.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32): 15849–15854, July 2019a. ISSN 1091-6490. doi: 10.1073/pnas.1903070116. URL `http://dx.doi.org/10.1073/pnas.1903070116`.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32): 15849–15854, 2019b. doi: 10.1073/pnas.1903070116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1903070116.

I. Bello, W. Fedus, X. Du, E. D. Cubuk, A. Srinivas, T.-Y. Lin, J. Shlens, and B. Zoph. Revisiting resnets: Improved training and scaling strategies, 2021. URL https://arxiv.org/abs/2103.07579.

E. Belouadah and A. Popescu. Scail: Classifier weights scaling for class incremental learning. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 03 2020.

E. Belouadah, A. Popescu, and I. Kanellos. A comprehensive study of class incremental learning algorithms for visual tasks. *Neural Networks*, 135:38–54, 2021.

Bennett et al. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35, 2007a.

J. Bennett, S. Lanning, et al. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. Citeseer, 2007b.

K. E. Bennin, J. W. Keung, P. Phannachitta, A. Monden, and S. Mensah. Mahakil: Diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction. *IEEE Transactions on Software Engineering*, 44:534–550, 2018. URL https://api.semanticscholar.org/CorpusID:47016996.

A. Birhane and V. U. Prabhu. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546. IEEE, 2021.

M. Birškus. Glasses Detector, 3 2024. URL https://github.com/mantasu/glasses-detector.

M. Borg, C. Englund, K. Wnuk, B. Duran, C. Levandowski, S. Gao, Y. Tan, H. Kaijser, H. Lönn, and J. Törnqvist. Safely entering the deep: A review of verification and validation for machine learning and a challenge elicitation in the automotive industry. *arXiv preprint arXiv:1812.05389*, 2018.

A. Borji. Pros and cons of gan evaluation measures, 2018. URL https://arxiv.org/abs/1802.03446.

L. Bossard, M. Guillaumin, and L. Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, N. Sepah, E. Raff, K. Madan, V. Voleti, S. E. Kahou, V. Michalski, D. Serdyuk, T. Arbel, C. Pal, G. Varoquaux, and P. Vincent. Accounting for variance in machine learning benchmarks, 2021. URL https://arxiv.org/abs/2103.03098.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In T. Theeramunkong, B. Kijsirikul, N. Cercone, and T.-B. Ho, editors, *Advances in Knowledge Discovery and Data Mining*, pages 475–482, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-01307-2.

J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

R. Burke. Knowledge-based recommender systems. *Encyclopedia of library and information systems*, 69, 05 2000.

R. Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12:331–370, 2002.

L. L. Cam. An approximation theorem for the Poisson binomial distribution. *Pacific Journal of Mathematics*, 10(4):1181 – 1197, 1960.

B. Cao, A. Araujo, and J. Sim. Unifying deep local and global features for image search. In *European Conference on Computer Vision*, pages 726–743. Springer, 2020.

Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. 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. IEEE, 2018.

N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari. End-to-end incremental learning. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*, pages 241–257, 2018.

A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. S. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, volume 11215 of *Lecture Notes in Computer Science*, pages 556–572. Springer, 2018.

N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, 16:321–357, 06 2002. doi: 10.1613/jair.953.

S. Chen, Y. Liu, X. Gao, and Z. Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Chinese Conference on Biometric Recognition*, pages 428–438. Springer, 2018.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020a. URL `https://proceedings.mlr.press/v119/chen20j.html`.

T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton. Big self-supervised models are strong semi-supervised learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22243–22255. Curran Associates, Inc., 2020b. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/fcbc95ccdd551da181207c0c1400c655-Paper.pdf`.

X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

R. Cohendet, C.-H. Demarty, N. Duong, M. Sjöberg, B. Ionescu, and T.-T. Do. Mediaeval 2018: Predicting media memorability task. *arXiv preprint arXiv:1807.01052*, 2018.

J.-R. Conti, N. Noiry, S. Clemencon, V. Despiegel, and S. Gentric. Mitigating gender bias in face recognition using the von mises-fisher mixture model. In *International Conference on Machine Learning*, pages 4344–4369. PMLR, 2022.

J. R. Cook and L. A. Stefanski. Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association*, 89(428):1314–1328, Dec. 1994. ISSN 0162-1459. doi: 10.1080/01621459.1994.10476871. URL `https://doi.org/10.1080/01621459.1994.10476871`. Publisher: ASA Website _eprint: https://doi.org/10.1080/01621459.1994.10476871.

D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World wide web*, pages 761–770, 2009.

B. Crepon and N. Jacquemet. *Econométrie : Méthodes et Applications*. 07 2010. ISBN 9782804153236.

E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation strategies from data. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 113–123, 2019. doi: 10.1109/CVPR.2019.00020.

I. Deandres-Tame, R. Tolosana, P. Melzi, R. Vera-Rodríguez, M. Kim, C. Rathgeb, X. Liu, A. Morales, J. Fiérrez, J. Ortega-Garcia, Z. Zhong, Y. Huang, Y. Mi, S. Ding, S. Zhou, S. He, L. Fu, H. Cong, R. Zhang, Z. Xiao, E. Smirnov, A. Pimenov, A. Grigorev, D. Timoshenko, K. Asfaw, C. Low, H. Liu, C. Wang, Q. Zuo, Z. He, H. O. Shahreza, A. George, A. Unnervik, P. Rahimi, S. Marcel, P. C. Neto, M. Huber, J. Kolf, N. Damer, F. Boutros, J. S. Cardoso, A. F. Sequeira, A. Atzori, G. Fenu, M. Marras, V. vStruc, J. Yu, Z. Li, J. Li, W. Zhao, Z. Lei, X. Zhu, X.-Y. Zhang, B. Biesseck, P. Vidal, L. Coelho, R. Granada, and D. Menotti. Second edition frcsyn challenge at cvpr 2024: Face recognition challenge in the era of synthetic data. 2024. doi: 10.48550/ARXIV.2404.10378.

M. Dehghani, Y. Tay, A. A. Gritsenko, Z. Zhao, N. Houlsby, F. Diaz, D. Metzler, and O. Vinyals. The benchmark lottery, 2021. URL `https://openreview.net/forum?id=5Str2l1vmr-`.

M. Delange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Software Engineering*, PP, Feb. 2021. ISSN 0098-5589. doi: 10.1109/TPAMI.2021.3057446.

Y. Deldjoo, M. Schedl, P. Cremonesi, and G. Pasi. Recommender systems leveraging multimedia content. *ACM Computing Surveys (CSUR)*, 53(5):1–38, 2020.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30, 2006. URL `http://jmlr.org/papers/v7/demsar06a.html`.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL `https://arxiv.org/abs/1810.04805`.

P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Dietz et al. Understanding the influence of data characteristics on the performance of point-of-interest recommendation algorithms. *arXiv.org*, 2023. doi: 10.48550/ARXIV.2311.07229.

M. D'Incà, C. Tzelepis, I. Patras, and N. Sebe. Improving fairness using vision-language driven image augmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4695–4704, 2024.

T. Diricic, D. Kowald, E. Lacic, and E. Lex. Beyond-accuracy: a review on diversity, serendipity, and fairness in recommender systems based on graph neural networks. *Frontiers in big data*, 2023. doi: 10.3389/FDATA.2023.1251072.

J. Dodge, G. Ilharco, R. Schwartz, A. Farhadi, H. Hajishirzi, and N. Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*, 2020.

Dong et al. When newer is not better: Does deep learning really benefit recommendation from implicit feedback? 2023. doi: 10.1145/3539618.3591785.

F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning*, 2017.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

A. Douillard, A. Ramé, G. Couairon, and M. Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9285–9295, June 2022.

C. Dwork, M. P. Kim, O. Reingold, G. N. Rothblum, and G. Yona. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1095–1108, 2021.

Ekstrand et al. Collaborative filtering recommender systems. *Found. Trends Hum. Comput. Interact.*, 2011. doi: 10.1561/1100000009.

K. Ethayarajh and D. Jurafsky. Utility is in the eye of the user: A critique of nlp leaderboards, 2021. URL `https://arxiv.org/abs/2009.13888`.

S. Fabbrizzi, S. Papadopoulos, E. Ntoutsi, and I. Kompatsiaris. A survey on bias in visual datasets. *Computer Vision and Image Understanding*, 223:103552, 2022.

Y. Fan, Y. Ji, J. Zhang, and A. Sun. Our model achieves excellent performance on movielens: What does it mean? *ACM transactions on office information systems*, 2024. doi: 10.1145/3675163.

E. Feillet, G. Petit, A. Popescu, M. Reyboz, and C. Hudelot. Advisil - a class-incremental learning advisor. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2400–2409, January 2023.

E. Feillet, A. Popescu, and C. Hudelot. Recommendation of data-free class-incremental learning algorithms by simulating future data, 2024. URL `https://arxiv.org/abs/2403.18132`.

Q. Feng, C. Guo, F. Benitez-Quiroz, and A. M. Martinez. When do gans replicate? on the choice of dataset size. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6701–6710, 2021.

K. R. M. Fernando and C. P. Tsokos. Dynamically weighted balanced loss: Class imbalanced learning and confidence calibration of deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2940–2951, 2022. doi: 10.1109/TNNLS.2020.3047335.

M. Ferrari Dacrema, P. Cremonesi, and D. Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, page 101–109, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362436. doi: 10.1145/3298689.3347058. URL `https://doi.org/10.1145/3298689.3347058`.

E. Fini, V. G. T. Da Costa, X. Alameda-Pineda, E. Ricci, K. Alahari, and J. Mairal. Self-supervised models are continual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9621–9630, 2022.

A. Fournier-Montgieux†, M. Soumm†, A. Popescu, B. Luvison, and H. L. Borgne. Toward fairer face recognition datasets, 2024. URL `https://arxiv.org/abs/2406.16592`.

J. Frankle and M. Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=rJl-b3RcF7`.

R. M. French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4): 128–135, 1999.

S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 329–338, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287589. URL `https://doi.org/10.1145/3287560.3287589`.

B. Friedman and H. Nissenbaum. Bias in computer systems. *ACM Transactions on information systems (TOIS)*, 14(3):330–347, 1996.

J. Gallardo, T. L. Hayes, and C. Kanan. Self-supervised training enhances online continual learning. In *British Machine Vision Conference (BMVC)*, 2020.

J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

J. Gareth, W. Daniela, H. Trevor, and T. Robert. *An introduction to statistical learning: with applications in R*. Spinger, 2013.

M. Ge, C. Delgado-Battenfeld, and D. Jannach. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 257–260, 2010.

T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

S. Gong, X. Liu, and A. Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *ECCV*, pages 330–347, 2020.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014a. URL `https://arxiv.org/abs/1406.2661`.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b.

D. Goswami, Y. Liu, B. Twardowski, and J. van de Weijer. Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. *Advances in Neural Information Processing Systems*, 36, 2024.

P. Goyal, P. Dollár, R. B. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *ArXiv*, abs/1706.02677, 2017. URL `https://api.semanticscholar.org/CorpusID:13905106`.

M. Grandini, E. Bagli, and G. Visani. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020.

J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

P. J. Grother, M. L. Ngan, K. K. Hanaoka, et al. Face recognition vendor test part 3: demographic effects. Technical report, National Institute of Standards and Technology, 2019.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1321–1330, 2017.

Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.

S. Gupta, A. Jalan, G. Ranade, H. Yang, and S. Zhuang. Too many fairness metrics: Is there a solution? *Web Technology eJournal*, 2020. URL `https://api.semanticscholar.org/CorpusID:219381013`.

D. Ha and D. Eck. A neural representation of sketch drawings. *CoRR*, abs/1704.03477, 2017. URL `http://arxiv.org/abs/1704.03477`.

H. Han, W.-Y. Wang, and B.-H. Mao. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In D.-S. Huang, X.-P. Zhang, and G.-B. Huang, editors, *Advances in Intelligent Computing*, pages 878–887, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31902-3.

D. J. Hand. Classifier Technology and the Illusion of Progress. *Statistical Science*, 21(1):1 – 14, 2006. doi: 10.1214/088342306000000060. URL `https://doi.org/10.1214/088342306000000060`.

M. Hardt, E. Price, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

M. A. Hardy. *Regression with dummy variables*. Number 93. Sage, 1993.

F. M. Harper and J. Konstan. The movielens datasets: History and context. *TIIS*, 2016. doi: 10.1145/2827872.

Hartig. Dharma: Residual diagnostics for hierarchical (multi-level / mixed) regression models., 2018. URL `https://cir.nii.ac.jp/crid/1370580229833186830`.

J. Harvey, Adam. LaPlace. Exposing.ai, 2021. URL `https://exposing.ai`.

T. L. Hayes and C. Kanan. Lifelong machine learning with deep streaming linear discriminant analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 220–221, 2020.

T. L. Hayes and C. Kanan. Online continual learning for embedded devices. In *Conference on Lifelong Learning Agents*, pages 744–766. PMLR, 2022.

T. L. Hayes, K. Kafle, R. Shrestha, M. Acharya, and C. Kanan. Remind your neural network to prevent catastrophic forgetting. In *European Conference on Computer Vision*, pages 466–483. Springer, 2020.

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.*, 21(9): 1263–1284, 2009.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, CVPR, 2016a.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016b.

K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020a.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, June 2022.

R. He and J. McAuley. Vbpr: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua. Neural collaborative filtering, 2017. URL `https://arxiv.org/abs/1708.05031`.

X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation, 2020b. URL `https://arxiv.org/abs/2002.02126`.

T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2496–2500. IEEE, 2022.

P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, jan 2004. ISSN 1046-8188. doi: 10.1145/963770.963772. URL https://doi.org/10.1145/963770.963772.

A. Hernandez-Garcia and P. König. Data augmentation instead of explicit regularization, 2020. URL https://openreview.net/forum?id=H1eqOnNYDH.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. doi: 10.1126/science.1127647. URL https://www.science.org/doi/abs/10.1126/science.1127647.

D. E. Ho, E. Black, M. Agrawala, and F.-F. Li. Domain shift and emerging questions in facial recognition technology. *HAI Policy Brief*, 2020a. URL https://hai.stanford.edu/sites/default/files/2020-11/HAI_FRT_WhitePaper_PolicyBrief_Nov2020.pdf.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. ISSN 1530-888X. doi: 10.1162/neco.1997.9.8.1735. Place: US Publisher: MIT Press.

S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim. A benchmark for interpretability methods in deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf.

S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin. Learning a unified classifier incrementally via rebalancing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 831–839, 2019.

Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*, pages 263–272. Ieee, 2008.

C. Huang, Y. Li, C. C. Loy, and X. Tang. Learning deep representation for imbalanced classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5375–5384, 2016. doi: 10.1109/CVPR.2016.580.

G. B. Huang and E. Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, May 2014.

I. Hupont and C. Fernández. Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019.

B. Hutchinson, N. Rostamzadeh, C. Greer, K. Heller, and V. Prabhakaran. Evaluation gaps in machine learning practice. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022. URL https://api.semanticscholar.org/CorpusID:248693488.

M. Hutson. Artificial intelligence faces reproducibility crisis. *Science*, 359(6377):725–726, 2018. doi: 10.1126/science.359.6377.725. URL https://www.science.org/doi/abs/10.1126/science.359.6377.725.

B. Y. Idrissi, M. Arjovsky, M. Pezeshki, and D. Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In B. Schölkopf, C. Uhler, and K. Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 336–351. PMLR, 11–13 Apr 2022. URL https://proceedings.mlr.press/v177/idrissi22a.html.

D. J. Im, S. Ahn, R. Memisevic, and Y. Bengio. Denoising criterion for variational auto-encoding framework, 2016. URL https://arxiv.org/abs/1511.06406.

A. Z. Jacobs and H. M. Wallach. Measurement and fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2019. URL https://api.semanticscholar.org/CorpusID:209202216.

A. K. Jain and S. Z. Li. *Handbook of face recognition*, volume 1. Springer, 2011.

A. K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology*, 14(1):4–20, 2004.

D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich. *Knowledge-based recommendation*, page 81–123. Cambridge University Press, 2010.

D. Jannach, M. Zanker, M. Ge, and M. Gröning. Recommender systems in computer science and information systems – a landscape of research. volume 123, 09 2012. ISBN 978-3-642-32272-3. doi: 10.1007/978-3-642-32273-0_7.

P. Janson, W. Zhang, R. Aljundi, and M. Elhoseiny. A simple baseline that questions the use of pretrained-models in continual learning. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022.

N. Japkowicz. Why question machine learning evaluation methods ? (an illustrative review of the shortcomings of current methods). 2006. URL https://api.semanticscholar.org/CorpusID:18364863.

L. A. Jeni, J. F. Cohn, and F. De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *Affective Computing and Intelligent Interaction (ACII)*. 2013.

Y. Ji, A. Sun, J. Zhang, and C. Li. A critical study on data leakage in recommender system offline evaluation. *ACM Transactions on Information Systems*, 2020. doi: 10.1145/3569930.

Z. Jiang, C. Zhang, K. Talwar, and M. Mozer. Characterizing structural regularities of labeled data in overparameterized models. *International Conference on Machine Learning*, 2020.

M. Jiménez-Guarneros and R. Alejo-Eleuterio. A class-incremental learning method based on preserving the learned feature space for eeg-based emotion recognition. *Mathematics*, 10(4), 2022. ISSN 2227-7390. doi: 10.3390/math10040598. URL https://www.mdpi.com/2227-7390/10/4/598.

L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020.

Q. Jodelet, X. Liu, and T. Murata. Balanced softmax cross-entropy for incremental learning. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part II*, pages 385–396. Springer, 2021.

Q. Jodelet, X. Liu, Y. J. Phua, and T. Murata. Class-incremental learning using diffusion model for distillation and replay. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3425–3433, 2023.

M. Kaminskas and D. Bridge. Diversity, serendipity, novelty, and coverage. *ACM Trans. Interact. Intell. Syst.*, 2016. doi: 10.1145/2926720.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv: Learning*, 2020.

K. Kärkkäinen and J. Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

M. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pages 2564–2572. PMLR, 2018.

I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4873–4882, 2016.

S. Khusro, Z. Ali, and I. Ullah. Recommender systems: issues, challenges, and research opportunities. In *Information science and applications (ICISA) 2016*, pages 1179–1189. Springer, 2016.

J.-K. Kim, I. Choi, and Q. Li. Customer satisfaction of recommender system: Examining accuracy and diversity in several types of recommendation approaches. *Sustainability*, 2021. doi: 10.3390/SU13116165.

M. Kim, A. K. Jain, and X. Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759, 2022.

M. Kim, F. Liu, A. Jain, and X. Liu. Dcface: Synthetic face generation with dual condition diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12715–12725, 2023.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2022. URL `https://arxiv.org/abs/1312.6114`.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, and A. K. Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012. doi: 10.1109/TIFS.2012.2214212.

J. Kleinberg, S. Mullainathan, M. Raghavan, J. Kleinberg, S. Mullainathan, and M. Raghavan. The challenge of understanding what users want. 2022. doi: 10.1145/3490486.3538365.

P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. A. Earnshaw, I. Haque, S. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, and P. Liang. Wilds: A benchmark of in-the-wild distribution shifts. *International Conference on Machine Learning*, 2020.

J. Konstan. Introduction to recommender systems: Algorithms and evaluation. *TOIS*, 2004. doi: 10.1145/963770.963771.

J. A. Konstan and J. Riedl. Recommender systems: from algorithms to user experience. *User Modeling and User-adapted Interaction*, 2012. doi: 10.1007/S11257-011-9112-X.

Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

S. Kornblith, J. Shlens, and Q. V. Le. Do better imagenet models transfer better? *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2656–2666, 2018. URL `https://api.semanticscholar.org/CorpusID:43928547`.

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

M. Kuanr and P. Mohapatra. Assessment methods for evaluation of recommender systems: A survey. *Foundations of Computing and Decision Sciences*, 2021. doi: 10.2478/FCDS-2021-0023.

A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

T. Kurashima, T. Iwata, G. Irie, and K. Fujimura. Travel route recommendation using geotagged photos. *Knowledge and information systems*, 37(1):37–60, 2013.

R. D. Labati, A. Genovese, E. Muñoz, V. Piuri, F. Scotti, and G. Sforza. Biometric recognition in automated border control: a survey. *ACM Computing Surveys (CSUR)*, 49(2):1–39, 2016.

Lakkaraju et al. What's in a name? understanding the interplay between titles, content, and communities in social media. *ICWSM*, 2013. doi: 10.1609/ICWSM.V7I1.14408.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

M. D. Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. G. Slabaugh, and T. Tuytelaars. Continual learning: A comparative study on how to defy forgetting in classification tasks. *CoRR*, abs/1909.08383, 2019.

F. Last, G. Douzas, and F. Bação. Oversampling for imbalanced learning based on k-means and smote. *ArXiv*, abs/1711.00837, 2017. URL https://api.semanticscholar.org/CorpusID:52875483.

S. Lathuiliere, P. Mesejo, X. Alameda-Pineda, and R. Horaud. A comprehensive analysis of deep regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(9):2065–2081, Sept. 2020. ISSN 1939-3539. doi: 10.1109/tpami.2019.2910523. URL http://dx.doi.org/10.1109/TPAMI.2019.2910523.

Latifi et al. Streaming session-based recommendation: When graph neural networks meet the neighborhood. *ACM Conference on Recommender Systems*, 2022. doi: 10.1145/3523227.3548485.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Y. LeCun, K. Kavukcuoglu, and C. Farabet. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 253–256, Paris, France, May 2010. IEEE. ISBN 978-1-4244-5308-5. doi: 10.1109/ISCAS.2010.5537907. URL http://ieeexplore.ieee.org/document/5537907/.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

K.-Y. Lee, Y. Zhong, and Y.-X. Wang. Do pre-trained models benefit equally in continual learning? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6485–6493, January 2023.

Y. Lee, J. R. Willette, J. Kim, and S. J. Hwang. Visualizing the loss landscape of self-supervised vision transformer, 2024. URL https://arxiv.org/abs/2405.18042.

D. Lehmann and M. Ebner. Subclass-based undersampling for class-imbalanced image classification. In *VISIGRAPP*, 2022. URL https://api.semanticscholar.org/CorpusID:246839869.

B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2018.

H. Li, P. Chaudhari, H. Yang, M. Lam, A. Ravichandran, R. Bhotika, and S. Soatto. Rethinking the hyperparameters for fine-tuning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1g8VkHFPH.

J. Li, H. Li, Z. He, W. Ma, P. Sun, M. Zhang, and S. Ma. Rechorus2.0: A modular and task-flexible recommendation library. *arXiv.org*, 2024. doi: 10.48550/ARXIV.2405.18058.

X. Li, G. Cong, X. li, T.-A. Pham, and S. Krishnaswamy. Rank-geofm: A ranking based geographical factorization method for point of interest recommendation. *SIGIR*, 09 2015.

Z. Li and D. Hoiem. Learning without forgetting. In *European Conference on Computer Vision*, ECCV, 2016.

D. Liang, R. G. Krishnan, M. D. Hoffman, and T. Jebara. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 world wide web conference*, pages 689–698, 2018.

Z. C. Lipton. The mythos of model interpretability, 2017. URL https://arxiv.org/abs/1606.03490.

C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*, 2016.

H. Liu, J. Z. HaoChen, A. Gaidon, and T. Ma. Self-supervised learning is more robust to dataset imbalance. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=4AZz9osqrar.

X.-y. Liu, J. Wu, and Z.-h. Zhou. Exploratory under-sampling for class-imbalance learning. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 965–969, 2006. doi: 10.1109/ICDM.2006.68.

D. Lopez-Paz and M. Ranzato. Gradient episodic memory for continual learning, 2022. URL https://arxiv.org/abs/1706.08840.

P. Lops, D. Jannach, C. Musto, T. Bogers, and M. Koolen. Trends in content-based recommendation. *User Modeling and User-Adapted Interaction*, 2019. doi: 10.1007/S11257-019-09231-W.

I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

X. Lu, C. Wang, J.-M. Yang, Y. Pang, and L. Zhang. Photo2trip: generating travel routes from geo-tagged photos for trip planning. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 143–152, 2010.

T. Ma, D. Li, W. Wang, and J. Dong. Cfa-net: Controllable face anonymization network with identity representation manipulation. *arXiv preprint arXiv:2105.11137*, 2021.

D. Madaan, H. Yin, W. Byeon, J. Kautz, and P. Molchanov. Heterogeneous continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15985–15995, 2023.

A. Madani and R. Yusof. Malaysian traffic sign dataset for traffic sign detection and recognition systems. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 8(11):137–143, 2016.

S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

M. Masana, X. Liu, B. Twardowski, M. Menta, A. D. Bagdanov, and J. van de Weijer. Class-incremental learning: survey and performance evaluation on image classification, 2021.

I. Masi, Y. Wu, T. Hassner, and P. Natarajan. Deep face recognition: A survey. In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 471–478, 2018. doi: 10.1109/SIBGRAPI.2018.00067.

D. K. McClish. Analyzing a portion of the roc curve. *Medical decision making*, 9(3):190–195, 1989.

M. Mccloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation*, 24:104–169, 1989.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall / CRC, London, 1989.

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 2021. doi: 10.1145/3457607.

Z. Meng, R. McCreadie, C. Macdonald, and I. Ounis. Exploring data splitting strategies for the evaluation of recommendation models. 2020. doi: 10.1145/3383313.3418479.

M. Merler, N. Ratha, R. S. Feris, and J. R. Smith. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019.

M. Mermillod, A. Bugaiska, and P. Bonin. The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in Psychology*, 4:504–504, 2013.

R. Meyes, M. Lu, C. W. de Puiseau, and T. Meisen. Ablation studies in artificial neural networks, 2019. URL `https://arxiv.org/abs/1901.08644`.

S. I. Mirzadeh, A. Chaudhry, D. Yin, H. Hu, R. Pascanu, D. Gorur, and M. Farajtabar. Wide neural networks forget less catastrophically. In *International Conference on Machine Learning*, pages 15699–15717. PMLR, 2022.

K. Mishchenko, A. Khaled, and P. Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33:17309–17320, 2020.

A. Misra, D. Hwang, Z. Huo, S. Garg, N. Siddhartha, A. Narayanan, and K. C. Sim. A comparison of supervised and unsupervised pre-training of end-to-end models. In *Interspeech*, pages 731–735, 2021.

M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application*, 8(1):141–163, 2021.

C. Molnar, G. König, J. Herbinger, T. Freiesleben, S. Dandl, C. A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, and B. Bischl. Pitfalls to avoid when interpreting machine learning models. *arXiv: Machine Learning*, 2020.

S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017.

M. Mousavi, A. Khanal, and R. Estrada. Ai playground: Unreal engine-based data ablation tool for deep learning. In *International Symposium on Visual Computing*, pages 518–532. Springer, 2020.

P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever. Deep double descent: where bigger models and more data hurt*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021 (12):124003, dec 2021. doi: 10.1088/1742-5468/ac3a74. URL `https://dx.doi.org/10.1088/1742-5468/ac3a74`.

P. C. Neto, E. Caldeira, J. S. Cardoso, and A. F. Sequeira. Compressed models decompress race biases: What quantized models forget for fair face recognition. *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2023. URL `https://api.semanticscholar.org/CorpusID:261076232`.

B. Neyshabur, H. Sedghi, and C. Zhang. What is being transferred in transfer learning? *arXiv preprint arXiv:2008.11687*, 2020.

B. Neyshabur, H. Sedghi, and C. Zhang. What is being transferred in transfer learning?, 2021. URL https://arxiv.org/abs/2008.11687.

C. Nguyen, T. Hassner, M. Seeger, and C. Archambeau. Leep: A new measure to evaluate transferability of learned representations. In *ICML 2020*, 2020. URL https://www.amazon.science/publications/leep-a-new-measure-to-evaluate-transferability-of-learned-representations.

M. Nilashi, K. Bagherifard, O. Ibrahim, H. Alizadeh, L. A. Nojeem, and N. Roozegar. Collaborative filtering recommender systems. *Research Journal of Applied Sciences, Engineering and Technology*, 5: 4168–4182, 2013. URL https://api.semanticscholar.org/CorpusID:9525761.

J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran. Measuring calibration in deep learning. In *CVPR workshops*, volume 2, 2019.

H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pages 3456–3465, 2017.

C. G. Northcutt, A. Athalye, and J. Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.

M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

A. Panos, Y. Kobe, D. O. Reino, R. Aljundi, and R. E. Turner. First session adaptation: A strong replay-free baseline for class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18820–18830, October 2023.

G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference, 2021. URL https://arxiv.org/abs/1912.02762.

G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2019.01.012. URL https://www.sciencedirect.com/science/article/pii/S0893608019300231.

G. Pavlidis. Recommender systems, cultural heritage applications, and the way forward. *Journal of Cultural Heritage*, 2019. doi: 10.1016/J.CULHER.2018.06.003.

J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018. ISBN 046509760X.

L. Pellegrini, V. Lomonaco, G. Graffieti, and D. Maltoni. Continual learning at the edge: Real-time training on smartphone devices. *arXiv preprint arXiv:2105.13127*, 2021.

F. Pelosin. Simpler is better: off-the-shelf continual learning through pretrained backbones. *arXiv preprint arXiv:2205.01586*, 2022.

M. V. Perera and V. M. Patel. Analyzing bias in diffusion-based face generation models. *arXiv preprint arXiv:2305.06402*, 2023.

G. Petit, A. Popescu, H. Schindler, D. Picard, and B. Delezoide. Fetril: Feature translation for exemplar-free class-incremental learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3911–3920, January 2023.

G. Petit†, M. Soumm†, E. Feillet†, A. Popescu, B. Delezoide, D. Picard, and C. Hudelot. An analysis of initial training strategies for exemplar-free class-incremental learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1837–1847, January 2024.

P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer. The good, the bad, and the ugly face challenge problem. *Image and Vision Computing*, 30(3):177–185, 2012.

D. A. Pierce and D. W. Schafer. Residuals in generalized linear models. *Journal of the American Statistical Association*, 81(396):977–986, 1986. ISSN 01621459, 1537274X. URL `http://www.jstor.org/stable/2289071`.

J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Lariviere, A. Beygelzimer, F. d'Alche Buc, E. Fox, and H. Larochelle. Improving reproducibility in machine learning research(a report from the neurips 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021. URL `http://jmlr.org/papers/v22/20-303.html`.

Y. Ping, Y. Li, and J. Zhu. Beyond accuracy measures: the effect of diversity, novelty and serendipity in recommender systems on user engagement. *Electronic Commerce Research*, 2024. doi: 10.1007/S10660-024-09813-W.

F. Pinto, P. H. Torr, and P. K. Dokania. An impartial take to the cnn vs transformer robustness contest. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 466–480. Springer, 2022.

A. Popescu and G. Grefenstette. Mining social media to create personalized recommendations for tourist visits. In *Proceedings of the 2nd international conference on computing for geospatial research & applications*, pages 1–6, 2011.

A. Popescu, G. Grefenstette, and P. A. Moëllic. Gazetiki: automatic creation of a geographical gazetteer. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 85–93, 2008.

A. Popescu, L.-D. Ștefan, J. Deshayes-Chossart, and B. Ionescu. Face verification with challenging imposters and diversified demographics. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3357–3366, 2022.

D. M. W. Powers. Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

A. Prabhu, P. H. Torr, and P. K. Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *European Conference on Computer Vision*, pages 524–540. Springer, 2020.

J. Purshouse and L. Campbell. Automated facial recognition and policing: a bridge too far? *Legal Studies*, 42(2):209–227, 2022.

H. Qiu, B. Yu, D. Gong, Z. Li, W. Liu, and D. Tao. Synface: Face recognition with synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10880–10890, 2021a.

R. Qiu, S. Wang, Z. Chen, H. Yin, and Z. Huang. CausalRec: Causal inference for visual debiasing in visually-aware recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. ACM, oct 2021b. doi: 10.1145/3474085.3475266. URL `https://dl.acm.org/doi/10.1145/3474085.3475266`.

J. Quang. Does training ai violate copyright law? *Berkeley Tech. LJ*, 36:1407, 2021.

J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. Dataset shift in machine learning. 2009. doi: 10.7551/MITPRESS/9780262170055.001.0001.

I. Rabiu, N. Salim, A. Da'u, and A. Osman. Recommender system based on temporal models: A systematic review. *Applied Sciences*, 10(7), 2020. ISSN 2076-3417. doi: 10.3390/app10072204. URL `https://www.mdpi.com/2076-3417/10/7/2204`.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. A. Hamprecht, Y. Bengio, and A. Courville. On the spectral bias of neural networks, 2019. URL `https://arxiv.org/abs/1806.08734`.

I. D. Raji and G. Fried. About face: A survey of facial recognition evaluation, 2021. URL `https://arxiv.org/abs/2102.00813`.

I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151, 2020.

A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

L. Ravaglia, M. Rusci, D. Nadalini, A. Capotondi, F. Conti, and L. Benini. A tinyml platform for on-device continual learning with quantized latent replays. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 11(4):789–802, 2021.

S. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. icarl: Incremental classifier and representation learning. In *Conference on Computer Vision and Pattern Recognition*, CVPR, 2017.

V. J. Reddi, C. Cheng, D. Kanter, P. Mattson, G. Schmuelling, C.-J. Wu, B. Anderson, M. Breughe, M. Charlebois, W. Chou, R. Chukka, C. Coleman, S. Davis, P. Deng, G. Diamos, J. Duke, D. Fick, J. S. Gardner, I. Hubara, S. Idgunji, T. B. Jablin, J. Jiao, T. S. John, P. Kanwar, D. Lee, J. Liao, A. Lokhmotov, F. Massa, P. Meng, P. Micikevicius, C. Osborne, G. Pekhimenko, A. T. R. Rajan, D. Sequeira, A. Sirasao, F. Sun, H. Tang, M. Thomson, F. Wei, E. Wu, L. Xu, K. Yamada, B. Yu, G. Yuan, A. Zhong, P. Zhang, and Y. Zhou. Mlperf inference benchmark, 2020. URL `https://arxiv.org/abs/1911.02549`.

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. URL `https://arxiv.org/abs/1506.01497`.

I. N. Rezende. Facial recognition in police hands: Assessing the 'clearview case'from a european perspective. *New Journal of European Criminal Law*, 11(3):375–389, 2020.

M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. 2020. doi: 10.18653/V1/2020.ACL-MAIN.442.

M. B. Ring. Child: A first step towards continual learning. *Machine Learning*, 28(1):77–104, 1997.

J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner. Face recognition: too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–1, 2020.

J. P. Robinson, C. Qin, Y. Henon, S. Timoner, and Y. Fu. Balancing biases and preserving privacy on balanced faces in the wild. *IEEE Transactions on Image Processing*, 32:4365–4377, 2023. doi: 10.1109/TIP.2023.3282837.

N. D. Rodríguez, V. Lomonaco, D. Filliat, and D. Maltoni. Don't forget, there is more than forgetting: new metrics for continual learning. *ArXiv*, abs/1810.13166, 2018. URL `https://api.semanticscholar.org/CorpusID:53106736`.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

D. Roy, M. Dutta, D. Roy, and M. Dutta. A systematic review and research perspective on recommender systems. 2022. doi: 10.1186/S40537-022-00592-5.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, Oct. 1986. ISSN 1476-4687. doi: 10.1038/323533a0. URL https://www.nature.com/articles/323533a0. Publisher: Nature Publishing Group.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

A. Salah, Q.-T. Truong, H. W. Lauw, and A. Mueller. Cornac: A comparative framework for multimodal recommender systems. *J. Mach. Learn. Res.*, 2020.

B. Saleh and A. Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015.

I. Sarridis, C. Koutlis, S. Papadopoulos, and C. Diou. Flac: Fairness-aware representation learning by suppressing attribute-class associations. *arXiv preprint arXiv:2304.14252*, 2023a.

I. Sarridis, C. Koutlis, S. Papadopoulos, and C. Diou. Towards fair face verification: An in-depth analysis of demographic biases. *arXiv preprint arXiv:2307.10011*, 2023b.

T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, Jan. 2015. ISSN 0893-6080. doi: 10.1016/j.neunet.2014.09.003. URL http://dx.doi.org/10.1016/j.neunet.2014.09.003.

B. Scholkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Towards causal representation learning. *ArXiv*, abs/2102.11107, 2021. URL https://api.semanticscholar.org/CorpusID:231986372.

F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

J. Schwarz, W. M. Czarnecki, J. Luketina, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, and R. Hadsell. Progress & compress: A scalable framework for continual learning. *ArXiv*, abs/1805.06370, 2018. URL https://api.semanticscholar.org/CorpusID:21718339.

N. Selwyn, M. Andrejevic, G. J. D. Smith, X. Gu, and C. O'Neill. Facial recognition technology: key issues and emerging concerns, 1 2023. URL https://bridges.monash.edu/articles/chapter/Facial_recognition_technology_key_issues_and_emerging_concerns/21965732.

S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016.

P. Serdyukov, V. Murdock, and R. Van Zwol. Placing flickr photos on a map. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 484–491, 2009.

A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein. Adversarial training for free! *Advances in neural information processing systems*, 32, 2019.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui. Towards out-of-distribution generalization: A survey. *arXiv: Learning*, 2021.

I. Shenbin, A. Alekseev, E. Tutubalina, V. Malykh, and S. I. Nikolenko. RecVAE: A new variational autoencoder for top-n recommendations with implicit feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. ACM, jan 2020a. doi: 10.1145/3336191.3371831. URL https://dl.acm.org/doi/10.1145/3336191.3371831.

I. Shenbin, A. Alekseev, E. Tutubalina, V. Malykh, and S. I. Nikolenko. Recvae: A new variational autoencoder for top-n recommendations with implicit feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM '20. ACM, Jan. 2020b. doi: 10.1145/3336191.3371831. URL http://dx.doi.org/10.1145/3336191.3371831.

V. Shevchenko, N. Belousov, A. Vasilev, V. Zholobov, A. Sosedka, N. Semenova, A. Volodkevich, A. Savchenko, and A. Zaytsev. From variability to stability: Advancing recsys benchmarking practices. *arXiv.org*, 2024. doi: 10.48550/ARXIV.2402.09766.

C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:1–48, 2019a. URL https://api.semanticscholar.org/CorpusID:195811894.

C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 2019b. doi: 10.1186/S40537-019-0197-0.

T. Silveira, M. Zhang, X. Lin, Y. Liu, and S. Ma. How good your recommender system is? a survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics*, 2019. doi: 10.1007/S13042-017-0762-9.

M. Singh, G. Ghalachyan, K. R. Varshney, and R. E. Bryant. An empirical study of accuracy, fairness, explainability, distributional robustness, and adversarial robustness. *ArXiv*, abs/2109.14653, 2021. URL https://api.semanticscholar.org/CorpusID:238227069.

A. Sinha, H. Namkoong, R. Volpi, and J. Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

D. Slack, A. Hilgard, S. Singh, and H. Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in neural information processing systems*, 34:9391–9404, 2021.

J. S. Smith, L. Karlinsky, V. Gutta, P. Cascante-Bonilla, D. Kim, A. Arbelle, R. Panda, R. Feris, and Z. Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11909–11919, 2023.

L. N. Smith. Cyclical learning rates for training neural networks, 2017. URL https://arxiv.org/abs/1506.01186.

L. N. Smith. A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay. *ArXiv*, abs/1803.09820, 2018. URL https://api.semanticscholar.org/CorpusID:4714223.

M. Smith, S. Miller, M. Smith, and S. Miller. Facial recognition and privacy rights. *Biometric Identification, Law and Ethics*, pages 21–38, 2021.

M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2009.03.002. URL https://www.sciencedirect.com/science/article/pii/S0306457309000259.

G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023.

C. Song and V. Shmatikov. Overlearning reveals sensitive attributes. *ArXiv*, abs/1905.11742, 2019. URL https://api.semanticscholar.org/CorpusID:167217888.

M. Soumm. Causal inference tools for a better evaluation of machine learning, 2024. URL https://arxiv.org/abs/2410.01392.

M. Soumm, A. Popescu, and B. Delezoide. Vis2rec: A large-scale visual dataset for visit recommendation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2987–2997, January 2023.

M. Soumm†, A. Fournier-Montgieux†, A. Popescu, and B. Delezoide. Quantifying user coherence: A unified framework for cross-domain recommendation analysis, 2024. URL https://arxiv.org/abs/2410.02453.

C. Staunton, S. Slokenberga, and D. Mascalzoni. The gdpr and the research exemption: considerations on the necessary safeguards for research biobanks. *European Journal of Human Genetics*, 27(8): 1159–1167, 2019.

H. Steck. Embarrassingly shallow autoencoders for sparse data. In *The World Wide Web Conference*, WWW '19. ACM, May 2019. doi: 10.1145/3308558.3313710. URL http://dx.doi.org/10.1145/3308558.3313710.

E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13693–13696, 2020.

A. Sun. On challenges of evaluating recommender systems in an offline setting. 2023. doi: 10.1145/3604915.3609495.

A. Sun. Beyond collaborative filtering: A relook at task formulation in recommender systems. *arXiv.org*, 2024. doi: 10.48550/ARXIV.2404.13375.

C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. *IEEE International Conference on Computer Vision*, 2017a. doi: 10.1109/ICCV.2017.97.

C. Sun, A. Shrivastava, S. Singh, and A. Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, 2017b. doi: 10.1109/ICCV.2017.97.

Z. Sun, D. Yu, H. Fang, J. Yang, X. Qu, J. Zhang, and C. Geng. Are we evaluating rigorously? benchmarking recommendation for reproducible evaluation and fair comparison. *ACM Conference on Recommender Systems*, 2020. doi: 10.1145/3383313.3412489.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks, 2014. URL https://arxiv.org/abs/1409.3215.

C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer, 2018.

M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le. Mnasnet: Platform-aware neural architecture search for mobile, 2019. URL https://arxiv.org/abs/1807.11626.

C. K. Tantithamthavorn, A. Hassan, and K. ichi Matsumoto. The impact of class rebalancing techniques on the performance and interpretation of defect prediction models. *IEEE Transactions on Software Engineering*, 46:1200–1219, 2018. URL https://api.semanticscholar.org/CorpusID:21317850.

R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. Measuring robustness to natural distribution shifts in image classification. *arXiv: Learning*, 2020.

M. Taskiran, N. Kahraman, and C. E. Erdem. Face recognition: Past, present and future (a review). *Digital Signal Processing*, 106:102809, 2020.

P. Terhörst, M. L. Tran, N. Damer, F. Kirchbuchner, and A. Kuijper. Comparison-level mitigation of ethnic bias in face recognition. In *2020 8th international workshop on biometrics and forensics (IWBF)*, pages 1–6. IEEE, 2020.

P. Terhörst, J. N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. Morales, J. Fierrez, and A. Kuijper. A comprehensive study on face recognition biases beyond demographics. *arXiv preprint arXiv:2103.01592*, 2021.

P. S. Thomas, B. Castro da Silva, A. G. Barto, S. Giguere, Y. Brun, and E. Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019.

B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528, 2011. doi: 10.1109/CVPR.2011.5995347.

H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

C.-F. Tsai, W.-C. Lin, Y.-H. Hu, and G.-T. Yao. Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. *Information Sciences*, 477:47–54, 2019. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2018.10.029. URL `https://www.sciencedirect.com/science/article/pii/S0020025518308478`.

A. Tsitsulin, M. Munkhoeva, and B. Perozzi. Unsupervised embedding quality evaluation. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pages 169–188. PMLR, 2023.

P. UE. Règlement (ue) 2024/1689 du parlement européen et du conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle et modifiant les règlements (ce) n° 300/2008, (ue) n° 167/2013, (ue) n° 168/2013, (ue) 2018/858, (ue) 2018/1139 et (ue) 2019/2144 et les directives 2014/90/ue, (ue) 2016/797 et (ue) 2020/1828 (règlement sur l'intelligence artificielle)texte présentant de l'intérêt pour l'eee., 2024. URL `https://eur-lex.europa.eu/eli/reg/2024/1689/oj`.

UNWTO. International tourist arrivals reach 1.4 billion two years ahead of forecasts. *United Nations World Tourism Organization*, 2019.

G. M. Van de Ven and A. S. Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding, 2019. URL `https://arxiv.org/abs/1807.03748`.

G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

R. Van Noorden. The ethical questions that haunt facial-recognition research. *Nature*, 587(7834): 354–358, 2020.

G. Varoquaux and O. Colliot. *Evaluating Machine Learning Models and Their Diagnostic Value*, pages 601–630. Springer US, New York, NY, 2023. ISBN 978-1-0716-3195-9. doi: 10.1007/978-1-0716-3195-9_20. URL `https://doi.org/10.1007/978-1-0716-3195-9_20`.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023. URL `https://arxiv.org/abs/1706.03762`.

T. Verma, L. Jin, J. Zhou, J. Huang, M. Tan, B. C. M. Choong, T. F. Tan, F. Gao, X. Xu, D. S. Ting, et al. Privacy-preserving continual learning methods for medical image classification: a comparative analysis. *Frontiers in Medicine*, 10, 2023.

J. Wang, W. Min, S. Hou, S. Ma, Y. Zheng, H. Wang, and S. Jiang. Logo-2k+: A large-scale logo dataset for scalable logo classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34 (4), pages 6194–6201, 2020a.

L. Wang, Z. Q. Lin, and A. Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific reports*, 10(1):19549, 2020b.

L. Wang, X. Zhang, H. Su, and J. Zhu. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

M. Wang and W. Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

M. Wang and W. Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9322–9331, 2020.

M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 692–702, 2019a.

M. Wang, Y. Zhang, and W. Deng. Meta balanced network for fair face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):8433–8448, 2021.

S. Wang, Y. Wang, J. Tang, K. Shu, S. Ranganath, and H. Liu. What your images reveal: Exploiting visual contents for point-of-interest recommendation. In *Proceedings of the 26th international conference on world wide web*, pages 391–400, 2017a.

T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5310–5319, 2019b.

Y.-X. Wang, D. Ramanan, and M. Hebert. Learning to model the tail. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 7032–7042, Red Hook, NY, USA, 2017b. Curran Associates Inc. ISBN 9781510860964.

Z. Wang, Z. Zhang, S. Ebrahimi, R. Sun, H. Zhang, C.-Y. Lee, X. Ren, G. Su, V. Perot, J. Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pages 631–648. Springer, 2022a.

Z. Wang, Z. Zhang, C.-Y. Lee, H. Zhang, R. Sun, X. Ren, G. Su, V. Perot, J. Dy, and T. Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022b.

Z. Wang, Z. Zhang, C.-Y. Lee, H. Zhang, R. Sun, X. Ren, G. Su, V. Perot, J. Dy, and T. Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 139–149, 2022c.

Z. Wang, Y. Bai, Y. Zhou, and C. Xie. Can CNNs be more robust than transformers? In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=TKIFuQHHECj`.

H. Werneck, N. Silva, M. Viana, A. C. Pereira, F. Mourão, and L. Rocha. Points of interest recommendations: methods, evaluation, and future directions. *Information Systems*, 101:101789, 2021.

T. Weyand, A. Araujo, B. Cao, and J. Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020a.

T. Weyand, A. Araujo, B. Cao, and J. Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020b.

A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht. The marginal value of adaptive gradient methods in machine learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4151–4161, 2017. URL `http://papers.nips.cc/paper/7003-the-marginal-value-of-adaptive-gradient-methods-in-machine-learning`.

E. Wood, T. Baltrušaitis, C. Hewitt, S. Dziadzio, T. J. Cashman, and J. Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021.

J. Wooldridge. *Introductory Econometrics: A Modern Approach*. South-Western Cengage Learning, 2013. ISBN 9781111534394. URL https://books.google.fr/books?id=4TZnpwAACAAJ.

Wu et al. Evaluating recommender systems. *International Conference on Digital Information Management*, 2012. doi: 10.1109/ICDIM.2012.6360092.

G. Wu, S. Gong, and P. Li. Striking a balance between stability and plasticity for class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1124–1133, 2021a.

T.-Y. Wu, G. Swaminathan, Z. Li, A. Ravichandran, N. Vasconcelos, R. Bhotika, and S. Soatto. Class-incremental learning with strong pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9601–9610, June 2022.

Y. Wu, Y. Chen, L. Wang, Y. Ye, Z. Liu, Y. Guo, and Y. Fu. Large scale incremental learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 374–382, 2019.

Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2021b. doi: 10.1109/TNNLS.2020.2978386.

G. Xu, T. D. Duong, Q. Li, S. Liu, and X. Wang. Causality learning: A new perspective for interpretable machine learning, 2021. URL https://arxiv.org/abs/2006.16789.

A. Yang, P. M. Esperança, and F. M. Carlucci. Nas evaluation is frustratingly hard, 2020. URL https://arxiv.org/abs/1912.12522.

L. Yang and A. Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, Nov. 2020. ISSN 0925-2312. doi: 10.1016/j.neucom.2020.07.061. URL http://dx.doi.org/10.1016/j.neucom.2020.07.061.

Z. Yang, X. Zhu, C. Jiang, W. Liu, and L. Shen. Ramface: Race adaptive margin based face recognition for racial bias mitigation. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8. IEEE, 2021.

D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks?, 2014. URL https://arxiv.org/abs/1411.1792.

Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, and C.-J. Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Syx4wnEtvH.

Z. You, J. Ye, K. Li, Z. Xu, and P. Wang. Adversarial noise layer: Regularize neural network by adding noise. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 909–913. IEEE, 2019.

L. Yu, B. Twardowski, X. Liu, L. Herranz, K. Wang, Y. Cheng, S. Jui, and J. v. d. Weijer. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

S. Yucer, F. Tektas, N. A. Moubayed, and T. P. Breckon. Measuring hidden bias within face recognition via racial phenotypes. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3202–3211, 2022. doi: 10.1109/WACV51458.2022.00326.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/zemel13.html.

F. Zenke, B. Poole, and S. Ganguli. Continual learning through synaptic intelligence. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3987–3995, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/zenke17a.html.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, Feb. 2021. ISSN 0001-0782. doi: 10.1145/3446776. URL https://doi.org/10.1145/3446776.

H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization, 2018. URL https://arxiv.org/abs/1710.09412.

S. Zhang, L. Yao, A. Sun, and Y. Tay. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38, 2019.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert, 2020a. URL https://arxiv.org/abs/1904.09675.

X. Zhang, J. P. Ono, H. Song, L. Gou, K.-L. Ma, and L. Ren. Sliceteller: A data slice-driven approach for machine learning model validation. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):842–852, 2023. doi: 10.1109/TVCG.2022.3209465.

Z. Zhang, P. Cui, and W. Zhu. Deep learning on graphs: A survey, 2020b. URL https://arxiv.org/abs/1812.04202.

B. Zhao, X. Xiao, G. Gan, B. Zhang, and S. Xia. Maintaining discrimination and fairness in class incremental learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 13205–13214. IEEE, 2020.

J. Zhao, L. Xiong, P. Karlekar Jayashree, J. Li, F. Zhao, Z. Wang, P. Sugiri Pranata, P. Shengmei Shen, S. Yan, and J. Feng. Dual-agent gans for photorealistic and identity preserving profile face synthesis. *Advances in neural information processing systems*, 30, 2017.

J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, et al. Towards pose invariant face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2207–2216, 2018.

W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, Dec. 2003. ISSN 0360-0300. doi: 10.1145/954339.954342. URL https://doi.org/10.1145/954339.954342.

S. Zheng, Y. Song, T. Leung, and I. Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 4480–4488, 2016.

T. Zheng and W. Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5(7), 2018.

T. Zheng, W. Deng, and J. Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.

Z. Zheng, Y. Cai, and Y. Li. Oversampling method for imbalanced classification. *Comput. Informatics*, 34:1017–1037, 2015. URL https://api.semanticscholar.org/CorpusID:7053926.

B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

F. Zhu, Z. Cheng, X.-y. Zhang, and C.-l. Liu. Class-incremental learning via dual augmentation. *Advances in Neural Information Processing Systems*, 34, 2021a.

F. Zhu, X.-Y. Zhang, C. Wang, F. Yin, and C.-L. Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5871–5880, 2021b.

K. Zhu, W. Zhai, Y. Cao, J. Luo, and Z.-J. Zha. Self-sustaining representation expansion for non-exemplar class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9296–9305, 2022.

Z. Zhu et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021c.

C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32, 2005.

B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le. Learning transferable architectures for scalable image recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8697–8710, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society. doi: 10.1109/CVPR.2018.00907. URL `https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00907`.