Fairness in Continual Federated Learning

Naima Noor

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2024

Committee:
Afra Mashhadi
Erika F. Parsons
Geethapriya Thamilarasu

Program Authorized to Offer Degree:
Computer Science & Software Engineering

University of Washington

Abstract

Fairness in Continual Federated Learning

Naima Noor

Chair of the Supervisory Committee:
Afra Mashhadi
Department of Computing & Software Systems

Continual Federated Learning (CFL) is a distributed machine learning technique that enables multiple clients to collaboratively train a shared model without sharing their data, while also adapting to new classes without forgetting previously learned ones. Currently, there are limited evaluation models and metrics for measuring fairness in CFL, and ensuring fairness over time can be challenging as the system evolves. To address this, our study explores temporal fairness in CFL, examining how the fairness of the model can be influenced by the selection and participation of clients over time.

We introduce novel fairness metrics—Delta Accuracy Fairness (DAF) and Delta Forgetting Fairness (DFF)—specifically designed to ensure temporal fairness in a CFL context. Additionally, we propose a set of client selection strategies that enhance the temporal fairness of the CFL model by addressing disparities in knowledge retention. Through comprehensive analysis, we demonstrate that while no single strategy guarantees perfect temporal fairness, the Low Participation and Low Average strategies consistently outperform others in terms of stability and equity. Furthermore, our findings underscore the adaptability of the Dynamic strategy, which shows significant promise in certain tasks. These insights pave the way for refining client selection strategies, enhancing CFL's fairness, and fostering more equitable learning environments.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

FL: Federated Learning

CL: Continual Learning

CFL: Continual Federated Learning

CIL: Class Incremental Learning

BWT: Backward Transfer

DAF: Delta Accuracy Fairness

DFF: Delta Forgetting Fairness

GAN: Generative Adversarial Networks

ACKNOWLEDGMENTS

I would like to thank my thesis advisor, Professor Afra Mashhadi. This paper would have never been accomplished without her assistance and dedicated involvement in every step of the process. I would like to thank you very much for your support and understanding over the last year.

I would also like to show gratitude to my committee, including Professor Erika F. Parsons and Professor Geethapriya Thamilarasu. I am very grateful for the technical and resource help I received that played an important part in the completion of this research work.

DEDICATION

This accomplishment is dedicated to my parents who have been there for me through my journey with chronic illness. To my family, your unwavering love and belief in me have been my pillars of strength. To my friends, thank you for your empathy, laughter, and understanding. To my medical team, your expertise and care have empowered me to not only manage my condition but also pursue my academic aspirations.

And to everyone living with chronic illnesses, this work stands as a testament to our resilience and hope. You are not alone in your struggles. This accomplishment symbolizes the triumph over my challenges and a reminder that our dreams are worth fighting for, despite the adversities we face.

Chapter 1

INTRODUCTION

Over the recent years, machine learning models have found adoption across consumer applications (such as e-commerce, streaming, social media, gig-economy) as well as mission-critical systems (such as healthcare, fraud-detection amongst others). All of these models rely on large-swaths of personal data to work effectively (better-recommendations, accurate diagnostics). Naturally, users (shoppers, business-clients, patients) often have privacy related concerns. One way to mitigate these apprehensions is build systems that can work (both for training and inference) without data leaving a customer's device or at least do not have a single-point that can be exploited. Furthermore, real-life machine learning systems often need to work in dynamically changing environments.

In this work, we investigate the fairness of the models trained in the CFL. We first propose two different ways of measuring fairness: one related to the deontological [2] ethical framework (i.e., fairness at every step which we refer to as temporal fairness) and the other one grounded on the utilitarian ethical point of view (i.e., fairness of the overall system, which we refer to as fairness-as-whole). We show that it is important for the research community to explore new definitions of fairness for tasks that are outside of traditional approaches, continual and dynamic in learning. We argue that for systems such as CFL, fairness cannot be measured at the end of the training period but has to be monitored over time and temporally.

We develop a new fairness metric based on *individual fairness* which corresponds closely to client fairness in FL, where the objective is to ensure that clients with similar contributions to training (e.g., rounds of participation, classes participated in, etc.) achieve similar outcome from the global model. We show that this metric is highly sensitive to the temporal and

continuous learning aspects of the model.

Our empirical analysis shows that state-of-the-art CFL algorithms [1] can attain fairness as a whole but not temporal fairness. We then show that by using a variety of strategies in client selection we can attain temporal fairness. The resultant algorithm is more equitable and remains robust to changes in data distribution or the model's learning trajectory. The contributions of our work are as follows:

- To the best of our knowledge, this is the first work to propose an individual fairness metric that is measured over time corresponding to the deontological ethical viewpoint. This metric includes feature representation corresponding to the participation level of clients as well as the number of incremental classes to which clients contributed to.

- We evaluate both fairness-as-whole and temporal fairness of the state-of-the-art CFL algorithm that uses generative replay and show that it fails to attain temporal fairness.

- We also propose a set of client selection strategies, including a Dynamic self-adapting algorithm, that enable the CFL model to achieve temporal fairness.

1.1 Introduction to Continual Learning (CL)

Continual Learning (CL), also known as lifelong learning, is a vital approach in machine learning where the model learns continuously, accumulates the knowledge learned from previous data, and applies it to new tasks. In the context of class incremental learning, a 'task' refers to the introduction of a new set of classes into the learning process. This method helps in overcoming the problem of catastrophic forgetting, where new learning can interfere with and overwrite previous knowledge [3].

Early studies in continual learning have been primarily focused on developing strategies that allow neural networks to retain old knowledge while adapting to new information. One of the seminal works in this area is by Kirkpatrick et al., who introduced the concept of

Elastic Weight Consolidation (EWC), which adds a regularization term in the loss function to protect the weights that are important for previous tasks [4].

Recent advancements in CL have addressed various challenges such as task-agnostic learning, where the model is not only able to learn continuously but also does not need to be explicitly informed about the task boundaries [5]. Techniques such as Gradient Episodic Memory (GEM) and Averaged Gradient Episodic Memory (AGEM) have been proposed to efficiently manage memory and computation, enhancing the scalability of continual learning models [6] [7].

Continual learning is increasingly being recognized for its potential applications across various domains including autonomous driving, healthcare, and robotics, where the ability to adapt to dynamic environments and new information is crucial [8].

The field continues to grow as researchers explore more efficient algorithms and architectures that can better mimic human-like learning, making continual learning a cornerstone for achieving true artificial intelligence [9].

1.2 Introduction to Federated Learning (FL)

Federated learning (FL) is an emerging methodology that enables model training across a multitude of devices or servers holding local data samples, without necessitating the exchange or centralization of this data. This framework provides privacy-preservation by design. In addition, the distributed nature of computations also improve the efficiency and scalability of underlying models. Consequently, FL has found applications in a variety of areas ranging from predictive-typing, disease-prediction to autonomous vehicles [10].

Early works in this area include algorithms proposed by Konečný et al. [11] and McMahan et al. [12]. McMahan et al. introduced the Federated Averaging (FedAvg) algorithm to enhance the efficiency of training deep networks by aggregating locally computed updates to improve a global model iteratively, while maintaining data privacy and security.

Recent advancements in FL have concentrated on addressing challenges related to model personalization and the handling of non-IID data (where client data distributions can be

significantly diverse). Algorithms such as Model Personalization (FedAMP) and Federated Learning of Cohorts (FLoC) have been shown to tackle these challenges by achieving more equitable and efficient learning outcomes [13].

Another area-of-focus for FL research is in the domain of ethical AI. These FL models seek to balance the benefits of AI with the imperatives of individual privacy rights [14].

1.3 Significance of Continual Learning in Federated Learning

While the de-centralized nature of FL helps to improve data privacy and security, it presents unique challenges related to handling evolving data-distributions [15, 16]. A common way to deal-with these shifting data-patterns is through continual learning. An early studies analyzing the feasibility of combining FL and continual learning was conducted by Parisi et al. [3] for the problem of "catastrophic forgetting" [3, 17].

Fairness in federated learning is another essential constraint as biased models can lead to unequal treatment or misdiagnosis in critical sectors like healthcare. This challenge is magnified by the diverse data sources inherent in federated learning. Continual Learning can also aid FL models in this regard by providing mechanisms for regular updates/refinements without losing grasp of historical data insights [13, 14].

Consider a healthcare scenario where multiple hospitals use federated learning to develop a predictive model for disease diagnosis. Each hospital's data may contain unique patterns related to patient demographics and socio-economic conditions. Suppose, hospital A primarily serves an urban population with prevalent lifestyle diseases, while Hospital B, in a rural area, reports more cases of infectious diseases. As data is pre-anonymized/trained by each hospital, traditional FL might skew the model towards the urban disease profile due to the larger data volume from Hospital A, leading to less accurate predictions for patients from Hospital B.

Our thesis aims to address this imbalance by integrating continual learning into federated learning. Continual learning enables models to adapt and learn from new data streams continuously. To illustrate, let's consider a before-and-after scenario: In addition to the

challenges in federated learning, fairness is equally crucial in the context of continual learning. Continual learning, or lifelong learning, involves models continuously adapting and learning from new data streams over time. The importance of fairness here lies in ensuring that as models evolve with new data, they do not inadvertently favor recent data or certain demographics over others [18]. This is particularly pertinent in dynamic environments where data distributions and characteristics change over time, as seen in healthcare settings where patient demographics and disease profiles may shift.

Before Continual Learning Integration: In the traditional federated learning approach, the model trained with data from Hospitals A and B might perform well initially. However, as Hospital C joins the network with a different patient profile (e.g., an aging population with chronic conditions), the model struggles to adapt to this new data, potentially becoming less accurate for Hospital C's patients.

After Continual Learning Integration: With continual learning principles applied, the model not only learns from the initial data from Hospitals A and B but also continuously adapts as Hospital C joins. This adaptability ensures that the model updates its parameters to accommodate the new data characteristics, maintaining accuracy and fairness across all patient groups [17].

By intertwining continual learning with federated learning, our research aims to pioneer new pathways in developing equitable and fair machine learning models. This involves designing algorithms that can adapt to new data while maintaining a balanced representation of all groups, and developing metrics to continually assess and correct biases as they arise. The integration aims to create models that are not only robust and adaptable but also ethically sound and socially responsible, reflecting a commitment to fairness in an ever-evolving data landscape.

1.4 Motivation

Fairness in Machine learning [19, 20] is an important aspect of advances in algorithms that can help avoid adverse societal impacts in many sectors such as healthcare [21], finance [22],

computer vision [23], and criminal justice [24], where decision-making models can significantly affect individuals' lives.

Most of the FairML literature has emerged from statistical group measures that can be translated to evaluate the outcome of the traditional machine learning algorithms, which are designed to perform well on a set of *known* tasks and classes that they have been exposed to during *centralized* training [25, 26, 27, 28].

1.4.1 Introducing Fairness in CFL

The essence of continual learning in FL is to enable models to adapt over time to new data without forgetting previously learned information. However, as datasets evolve, the risk of bias and unfairness may increase if newer data distributions reflect imbalances or if certain demographic groups are underrepresented in the data streams over time. The dynamic nature of these datasets complicates the task of maintaining fairness, as models must continuously learn from new data while ensuring equitable treatment across all groups. Furthermore, the continual adaptation of models to new data can lead to "model drift," where the performance of the model may start favoring recent data or specific demographics over others, thereby undermining fairness over time.

These challenges are magnified in federated settings where data is inherently heterogeneous and decentralized across numerous nodes. As each node may contribute data that evolves differently, ensuring that the global model remains fair to all contributions becomes a complex task. This scenario demands novel approaches to model training and updating that can account for the temporal dynamics of data and model evolution, a topic explored in the research on fairness in machine learning and federated learning systems [13, 29].

Given these limitations, a paradigm that can adapt to new private data and learn continually is highly desirable. This has led to the emergence of the Continual Federated Learning field of research. For example in the use cases related to vision tasks involving continuous learning from visual data to enhance recognition, tracking, and analysis become particularly important. In such tasks, it is required that the system updates its knowledge and

decision-making with new visual inputs continuously [30]. Recent advancements in FL have shown potential in such applications, where deployed models could be running ubiquitously on devices such as mobile phones and edge devices such as CCTV cameras [31]. As a result, underlying models need to be capable of adapting to new information in the absence of a traditional model-fitting pipeline (i.e., data collection, sharing, and centralized training). However, this newly established paradigm imposes great challenges regarding *how* and *when* to measure the fairness of the underlying model, as the existing approaches proposed in the FL community and CL community fall short.

Measuring and ensuring fairness is a challenging task that is especially difficult in distributed and decentralized settings such as federated learning (FL) where the model developer is not privy to clients' local data.

1.4.2 Algorithmic Fairness Over Time

The CFL paradigm introduces new dimensions to fairness challenges, given its dynamic nature where models learn and adapt over time without centralized data aggregation. The temporal aspect of fairness becomes increasingly complex as models must ensure fair treatment across a continually updating dataset and a potentially changing population of clients. This continuous learning process can inadvertently lead to temporal biases if earlier data disproportionately influences the model compared to later data, or if certain client updates are prioritized over others.

In this vein, temporal fairness emerges as a critical component in the continuum of CFL. Traditional measures of fairness may not suffice in a landscape where models must not only be just at a single point but maintain that equity through time and across evolving data streams. This ongoing assessment is crucial in applications where the timing of decisions is as consequential as the decisions themselves—for example, in adaptive traffic control systems that learn from a constant influx of data to optimize flow while ensuring equitable mobility for all parts of a city, or in dynamic content recommendation engines that must continually adjust to user preferences without biasing exposure to new content. The var-

ied participation rates of clients, influenced by their technical capabilities, data availability, and operational conditions, also demand that fairness metrics accommodate the temporal dimension to prevent biases against less frequently participating clients, as shown in Figure 1.1

Hence, addressing the fairness challenges of CFL requires a framework that not only understands the evolving data distributions but also respects the temporal sequences of data contributions and participation levels from clients.

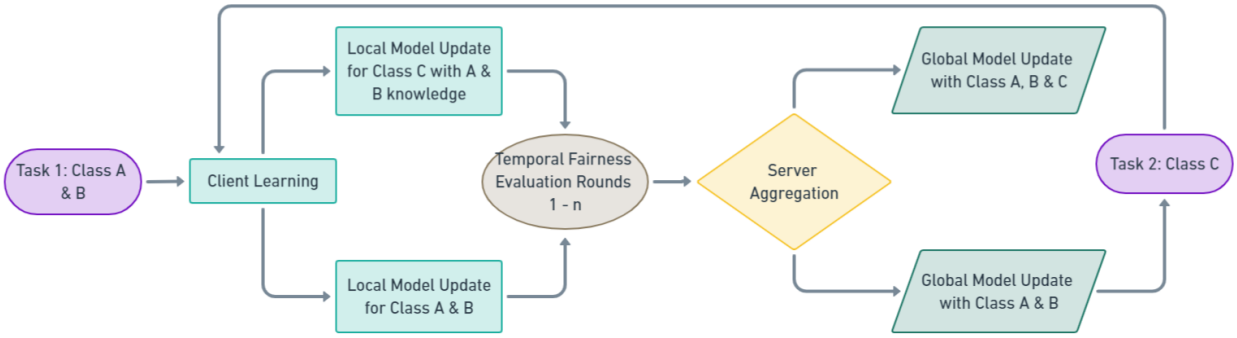

Figure 1.1: Temporal Fairness within CFL. Task represents sequential learning of a new class in a class incremental learning scenario.

1.5 Proposed Solution

By intertwining continual learning with federated learning, our research stands at the confluence of federated learning and continual learning, aiming not just to navigate but to innovate the pathways towards developing machine learning models that epitomize equity and fairness. Central to our approach is the design of adaptive algorithms capable of learning from new data streams while ensuring a representationally balanced perspective across all participant groups. A pivotal element of our methodology involves the development of dynamic metrics specifically designed to monitor and assess fairness over time. This initiative is motivated

by the recognition that fairness is not a static attribute but a dynamic quality that must be vigilantly maintained as data landscapes evolve and models continually adapt.

The core of our proposed solution is the development of a fairness metric that is sensitive to the temporal aspects of model training in a federated learning environment integrated with continual learning principles. This metric will enable the ongoing assessment of model fairness, identifying potential biases as they emerge and providing feedback for model adjustment in real-time. By continuously monitoring fairness, we aim to ensure that models remain equitable across different groups, irrespective of changes in data distribution or the model's learning trajectory.

In this way, the integration of continual learning into federated learning is not just a technical enhancement but a step towards more ethically sound and socially responsible AI systems. By ensuring that our models are not only intelligent but also fair and just, this research promises to make a significant contribution to the field of machine learning, particularly in contexts where data is a continuously evolving phenomenon.

In this study, we aim to investigate following research questions:

RQ1. Maintain fairness when new clients join or leave the federation over time.

RQ2. Explore fairness-aware continual learning algorithms for use in federated learning.

1.6 Outline

The thesis is organized as follows: Chapter 2 outlines the related works in the field of fairness in ML, FL and CL. Chapter 3 dives deep into the proposed definition of fairness in CFL. Chapter 4 dives deep into the proposed fairness metrics and methodology. Chapter 5 focuses on the different experiments, and their results. Chapter 6 focuses on the proposed work's overall discussion, limitations, implications, and future aspects.

Chapter 2

RELATED WORK

In this chapter, we summarize recent works related to fairness in federated learning and continual learning.

2.1 Fairness in Federated Learning (FL)

There has been an increase in the number of FL algorithms specifically designed to tackle fairness in the past years [32, 33, 34, 35, 36]. Most of these works have emerged at the intersection of group fairness and FL, where some rely on sharing statistics and sensitive attributes with the FL server to enforce demographic parity [37, 38]. Others aim to achieve group fairness by applying debiasing on each client locally [39, 36, 40] and with sporadic global debiasing [32].

The concept of *individual* fairness in Federated Learning has received less attention in recent years. To this end, a handful of studies [41, 42] have proposed different models and techniques that try to address data-level and client-level fairness. For instance, Zeng et al. [32] proposed a theoretical framework suggesting that federated learning can inherently improve model fairness compared to centralized approaches. They introduce FEDFB, a privacy-preserving fair learning algorithm, which demonstrates competitive performance against existing methods [32]. However, other studies have shown that models trained with federated learning can often exhibit higher bias and become more susceptible to attacks due to variations in local data distributions across clients [43].

In [42], Shi et al. proposed a taxonomy termed as "Fairness-aware federated learning" (FACL) which covered different stages like client selection, optimization, contribution evaluation, and incentive distribution. This work aimed to develop algorithms that could ensure

similar model performance across clients fairness. In [44], Lyu et al. presented a framework called "Collaborative Fair Federated Learning" (CFFL) which leveraged reputations to encourage participants to converge towards models promoting fairness and accuracy. However, they also considered fairness only as a measurement of the final snapshot of the learning process.

2.2 Fairness in Continual Learning (CL)

Fairness in continual learning (CL) adds another layer of complexity, ensuring that the model performs consistently well for different data distributions or subgroups present across the learning sequence. As models evolve to accommodate new information, maintaining a balance to prevent catastrophic forgetting is crucial [45].

Innovative approaches have been developed to adjust training sequences for fairness, minimizing biases that might favor recent data [30]. Additionally, robust methods such as integrating data augmentation with regularization strategies are employed to enhance training fairness [46].

Zhao et al. [47] presented a class incremental learning (IL) method that utilized fine-tuning alongside a dual memory system to mitigate the adverse effects of catastrophic forgetting in image recognition. Their approach included a two-phase method involving knowledge distillation to maintain the model's discrimination capabilities across previously learned classes and introduced Weight Aligning (WA) to adjust biases in the final layer, thus promoting fairer treatment across all classes.

In a different vein, Troung et al. [48] introduced a fairness-oriented continual learning approach tailored to the semantic segmentation problem. They proposed a Prototypical Contrastive Clustering loss to address the challenges posed by catastrophic forgetting and background shifts, demonstrating a novel way to integrate fairness directly into the loss function of learning algorithms.

Further advancing the discussion on domain-specific applications, the study by Churamani et al. [49] focused on Domain-Incremental Continual Learning (Domain-IL) to enhance

fairness in Facial Expression Recognition (FER) systems. This work emphasized the comparative advantage of CL methods over traditional non-CL approaches, highlighting significant improvements in accuracy and fairness metrics, particularly in addressing biased data distributions.

Despite these advancements, significant limitations remain in the field of CL. Most approaches still grapple with catastrophic forgetting, even with sophisticated methods like knowledge distillation and fine-tuning. Furthermore, while efforts to minimize biases through tailored training sequences and loss functions are noteworthy, they often require complex adjustments and may not be effective universally across different data types or scenarios. A critical gap in current research is the lack of emphasis on temporal fairness, especially crucial as models continuously learn and evolve over time. This oversight points to the need for methodologies that can assess and ensure fairness dynamically as learning progresses.

Chapter 3

FAIRNESS DEFINITION IN CONTINUAL FEDERATED LEARNING

3.1 Philosophical Frameworks of Ethics

In the evaluation of machine learning models, the ethical framework of utilitarianism, as introduced by philosopher Jeremy Bentham [50], is currently prioritized over other ethical paradigms such as deontology. This preference is largely due to the fact that most fairness metrics are designed to measure allocative harms. As a result, utilitarianism, with its quantitative nature, has found significant application in fields such as economics and public policy. Therefore, it is essential to understand the implications of this utilitarian approach when assessing the fairness of machine learning models.

Similarly, the machine learning community has been thinking about fairness as a balance between the trade-off that is between fairness and accuracy. ML models are debiased to achieve an equilibrium between predictive performance metrics like accuracy, f1-score, mean squared errors, etc. with other fairness-related metrics such as demographic parity, equalized odds, etc. This view of the discipline has been guided by utilitarian interpretations of fairness which focus on a so-called summum bonum, or the highest order of the good.

In the same vein, we are beginning to observe similar narrow utilitarian considerations of fairness in the CFL community. For instance, in [1] authors measure the disparity between the client's accuracy at the end of the training for each task as fairness, with no consideration for the client's participation, and whether during the lifetime of the system the fairness was achieved. Only recently a handful of works including a recent article by Mougan et al. [51] have raised concerns regarding the alignment problem of the utilitarian framework for advanced ML and AI applications. We also believe that by adopting a utilitarian view and

measuring fairness at the end of the CFL system, we run into the risk of using metrics that are intrinsically misaligned with the paradigm.

In contrast, Kant's moral theory known as Groundwork [52] is grounded on two main principles: "Act only on a maxim that you can also will to become a universal law (Groundwork, 4:421)", and "Act in such a way that you treat humanity, whether in your own person or anyone else's, never merely as a means, but also always as an end (4:429)". Using this deontological perspective in this work, we posit the following hypothesizes:

3.1.1 Definitions

To ground our work on the discussed deontological principles, we use the following definitions.

Definition 1: Individual Fairness posits that a model $f(\cdot)$ achieves individual fairness w.r.t. two individual samples x and x' [53] if:

$$d(f(x), f(x')) \leq L \cdot d(x, x') \quad \forall\, x, x' \in \mathcal{X} \tag{3.1}$$

where $L > 0$ is a constant and $d(\cdot, \cdot)$ represents a distance metric on the set $\mathcal{X}$.

Put simply, for two clients with similar contributions to training a model, we expect a similar treatment (i.e., accuracy) of the model.

Definition 2: We define temporal fairness in CFL as the *consistent treatment of clients over time*, ensuring that fairness is maintained not just after the end of training but throughout the *continual* learning process. We can thus rewrite the above definition for continuous temporal time t as:

$$d(f_t(x_t), f_t(x'_t)) \leq L_t \cdot d(x, x') \quad \forall\, x_t, x'_t \in \mathcal{X}_t \tag{3.2}$$

In the next section, we detail how we expand on these two definitions and propose concrete metrics to measure temporal fairness in CFL.

3.2 Proposing Novel Metrics for Fairness in CFL

Fairness in CFL refers to the equitable opportunity for each client to contribute to the learning process and benefit from it, as defined in eq. 3.1. This involves ensuring that the model updates do not induce or perpetuate bias and that the learning process is transparent and accountable. In this section, we propose a methodology to achieve this task. To this end, let us start by considering various client-specific features.

3.2.1 Feature Set for client Similarity Calculation

An expanded set of features is essential for dynamically assessing the similarity (individual fairness (eq. 3.1) among clients in terms of engagement and contribution to the learning process. Let $\mathcal{C}$ be the set of all clients and let $m = |\mathcal{C}|$ be the number of observations (number of clients). To this end, for each client i, we create a set $\mathcal{Q}_i$, which includes frequency f_i, regularity r_i, trend t_i, participation count p_i, average interval a_i, and number of classes c_i such that $\mathcal{Q}_i = \{f_i, r_i, t_i, p_i, a_i, c_i\}$.

If the objective is to measure fairness as a whole, one could argue that similarity can be assessed by looking at the overall participation of clients. This approach provides a global view of fairness, taking into account the overall contribution of each client to the learning process. However, this procedure averages out the temporal aspects of the fairness measures (eq. 3.2). Therefore, a more granular approach is required for evaluating temporal fairness.

This is particularly relevant in a CL setting where the model is updated progressively during multiple rounds of training. Therefore, in CFL settings, it is necessary to calculate these features in each round of training. This way, the resultant metric can provide each client an equitable opportunity to contribute to the model training and benefit from the learning process. In this work, we consider the following client attributed features:

- **Frequency (n_i)**: Represents the total number of participations by client i within a specific timeframe, where $n_i \in \mathbb{N}$. This metric is indicative of the client's engagement level within the CFL environment.

- **Regularity** (r_i): Quantifies the consistency of participation over time for client i, measured as the standard deviation of time intervals between participations, where $r_i \in \mathbb{R}_{\geq 0}$. This metric helps to quantify the stability and predictability of a client's participation.

- **Trend** (t_i): Measures the directional change in client i's participation over time, reflecting whether that client's engagement is increasing, decreasing, or stable, where $t_i \in \mathbb{Z}$. It helps in anticipating future participation patterns.

- **Participation Count** (p_i): The total number of specific tasks in which client i has engaged, where $p_i \in \mathbb{N}$. This metric helps to track a client's involvement across diverse learning tasks.

- **Average Interval** (a_i): Measures the average time between consecutive participations for client i, where $a_i \in \mathbb{R}_{\geq 0}$, offering a complementary perspective to regularity by highlighting the temporal dynamics of engagement.

- **Number of Classes** (c_i): Denotes the number of unique classes encountered by client i within the learning process, where $c_i \in \mathbb{N}$, encoding the complexity and breadth of each client's engagement in class-incremental learning. In an i.i.d (Independent and Identically Distributed) setting, each client is likely to encounter a similar distribution of classes, which simplifies the analysis of fairness as the learning experiences and challenges are uniform across clients. However, in a non-i.i.d (non-Independent and Identically Distributed) scenario, clients may be exposed to different subsets of classes, reflecting a more complex and heterogeneous learning environment. This heterogeneity can lead to variability in the learning experience and model performance, which must be accounted for when assessing fairness. By tracking c_i for each client and considering the i.i.d or non-i.i.d nature of class distribution, we can, more accurately, adjust our fairness metrics to ensure that clients with similar number of class exposures receive

similar treatment, thereby upholding the principle of individual fairness in the CFL ecosystem.

These attributes $\mathcal{Q}_i$ form the basis for the client feature matrix, $\mathbf{S}_{\text{features}} \in \mathbb{R}^{m \times |\mathcal{Q}|}$ ($|\cdot|$ represents the cardinality of a set), defined as:

$$\mathbf{S}_{\text{features}} = \begin{bmatrix} n_1 & r_1 & t_1 & p_1 & \mathrm{a}_1 & c_1 \\ n_2 & r_2 & t_2 & p_2 & \mathrm{a}_2 & c_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ n_m & r_m & t_m & p_m & \mathrm{a}_m & c_m \end{bmatrix}, \tag{3.3}$$

where each row represents the unique engagement profile of a client i across the defined metrics, and m is the number of clients.

Also, this matrix $\mathbf{S}_{\text{features}}$ serves as a quantitative foundation for comparing client behavior, helping us to compute similarity across different client pairs. Through this analytical framework, we aim to enhance our understanding of client engagement dynamics and contribute to the development of more effective, personalized learning environments.

The selected features are particularly pertinent to CFL, as they capture various dimensions of client engagement that are essential to the iterative learning process. In a CFL setup, where the model is updated continuously, it is critical to track not just the quantity of data contributed by a client, but also the regularity and variability of their contributions. These features were carefully chosen to ensure that the temporal and distributional aspects of client engagement are adequately represented, allowing for a nuanced analysis of fairness. By considering both the engagement patterns (frequency, regularity, trend, and average interval) and the diversity of contributions (participation count and number of classes), we ensure a holistic evaluation of client similarity and fairness in the continually evolving model landscape of CFL. The choice of these features is further substantiated by their relevance in capturing individual client behavior, which directly influences the training and performance of the federated model.

The dynamic nature of CFL necessitates the recalibration of these metrics in each learning round. This ensures that the system's fairness measures remain adaptive and responsive to:

1. The most current client engagement and contribution patterns.

2. Changes introduced by new classes and learning tasks.

3. Evolving client behavior, participation trends, and learning challenges.

3.2.2 Delta Output Matrix for Fairness Assessment

We introduce two novel fairness measures tailored for the CFL environment: Delta Accuracy and Delta Forgetting. These metrics are designed to capture the nuances of model performance and learning consistency across different clients, further refining our assessment of fairness. In our notation, the large Δ signifies the difference operation used in the matrices, while the small δ specifically refers to the individual matrices assessing different aspects of fairness—accuracy and forgetting.

Delta Accuracy Fairness (DAF) Matrix

For assessing fairness in accuracy improvements across clients, we utilize the Delta Accuracy Fairness (DAF) matrix, denoted as $\Delta_{\delta\text{AF}} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|}$. Each element $\Delta_{\delta\text{AF}_{i,j}}$ represents the absolute difference in accuracy changes between clients i and j from one training round to the next. The matrix is defined as:

$$\Delta_{\delta\text{AF}_{i,j}} = \left|\Delta\text{Acc}_i - \Delta\text{Acc}_j\right|, \tag{3.4}$$

where ΔAcc_i and ΔAcc_j are the accuracy changes for clients i and j, respectively. Lower values in $\Delta_{\delta\text{AF}}$ indicate minimal disparity in learning improvements, highlighting fairness in terms of accuracy changes.

Forgetting Measure

Before constructing the Delta Forgetting Fairness (DFF) matrix, it is crucial to define the forgetting measure, which quantifies the loss of previously learned information as new knowledge is acquired. This is particularly relevant in scenarios such as Class-Incremental Learning (CIL) where continuous learning is critical. The forgetting measure for client i is calculated as follows:

$$\Delta\text{Forg}_i = \frac{1}{|\mathcal{K}_i|} \sum_{k \in \mathcal{K}_i} \max\left(0, \frac{A_{i,k}^{\text{initial}} - A_{i,k}^{\text{current}}}{A_{i,k}^{\text{initial}}}\right), \tag{3.5}$$

where $\mathcal{K}_i$ represents the classes encountered by client i, with $A_{i,k}^{\text{initial}}$ and $A_{i,k}^{\text{current}}$ indicating the initial and current accuracies on class k.

Delta Forgetting Fairness (DFF) Matrix

To assess fairness in knowledge retention, the Delta Forgetting Fairness (DFF) matrix, $\Delta_{\delta\text{FF}} \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|}$, is used. Each element $\Delta_{\delta\text{FF}_{i,j}}$ measures the absolute difference in forgetting rates between clients i and j as:

$$\Delta_{\delta\text{FF}_{i,j}} = \left|\Delta\text{Forg}_i - \Delta\text{Forg}_j\right|, \tag{3.6}$$

where ΔForg_i and ΔForg_j are the forgetting measures for clients i and j, respectively. Smaller values in $\Delta_{\delta\text{FF}}$ reflect a more equitable distribution of knowledge retention, thus supporting fairness in the learning process.

By integrating these matrices, $\Delta_{\delta\text{AF}}$ and $\Delta_{\delta\text{FF}}$, into our fairness assessment, we provide a comprehensive analysis of how individual fairness is maintained in both the learning and forgetting processes within the CFL environment.

3.2.3 Computation of Fairness Matrices for Delta Accuracy and Delta Forgetting

In our CFL framework, fairness is quantified through two distinct matrices: the Delta Accuracy Fairness (DAF) matrix, as applied to federated learning, and the Delta Forgetting

Fairness (DFF) matrix for continual learning. These are denoted as $\mathbf{F}_{\delta\text{AF}}$ and $\mathbf{F}_{\delta\text{AF}}$, respectively, and are constructed in $\mathbb{R}^{|\mathcal{I}|\times|\mathcal{J}|}$. Each matrix element is derived from a similarity score between clients and the corresponding changes in accuracy (DAF) or forgetting measures (DFF).

For Delta Accuracy Fairness (DAF), tailored to federated learning, the matrix element $F_{\delta\text{AF}_{i,j}}$ is defined as:

$$F_{\delta\text{AF}_{i,j}} = S_{i,j} \times \Delta_{\delta\text{AF}_{i,j}}, \tag{3.7}$$

where $\delta\text{AF}_{i,j}$ quantifies the difference in accuracy changes between clients i and j. A lower $F_{\delta\text{AF}_{i,j}}$ value indicates a fairer distribution of learning benefits, reflecting more uniform changes in accuracy across clients.

For Delta Forgetting Fairness, specific to continual learning, the matrix element $F_{\delta\text{FF}_{i,j}}$ is computed as:

$$F_{\delta\text{FF}_{i,j}} = S_{i,j} \times \Delta_{\delta\text{FF}_{i,j}}, \tag{3.8}$$

where $\delta\text{FF}_{i,j}$ measures the difference in forgetting rates between clients i and j. A lower $F_{\delta\text{FF}_{i,j}}$ value signals a more equitable distribution of the forgetting effect, thus promoting fairness in knowledge retention.

These matrices, $F_{\delta\text{AF}_{i,j}}$ and $F_{\delta\text{FF}_{i,j}}$, serve as fundamental tools for evaluating temporal fairness, where low values are indicative of fairness, ensuring that clients experiencing similar conditions undergo comparable changes in accuracy and forgetting. An ideal value of $F_{\delta\text{AF}_{i,j}} = 0$ or $F_{\delta\text{FF}_{i,j}} = 0$ represents perfect fairness, where similarly engaged clients have identical outcomes. Higher values, however, highlight potential disparities and emphasize the need for enhanced fairness mechanisms in client treatment.

3.2.4 Fairness Ratio

After computing the fairness matrix $\mathbf{F}$, we define the *fairness ratio*, denoted by ρ, which quantifies the proportion of fair decisions relative to the total number of decisions. The *fairness ratio* ρ is expressed by the formula:

$$\rho = \frac{\text{Fair Counts}}{\text{Fair Counts} + \text{Unfair Counts}}. \tag{3.9}$$

Here, a *fair count* is indicated by a lower $F_{i,j}$ value, suggesting similar outcomes for comparable clients, thus reflecting algorithmic fairness. Conversely, an *unfair count* is marked by a higher $F_{i,j}$ value, indicating disparities in the treatment of similar clients, which suggests potential unfairness.

The *fairness ratio* ρ serves as a quantitative indicator of the algorithm's performance in terms of fairness, with a higher ρ indicating a greater prevalence of fair decisions within the algorithm's operations.

Chapter 4

METHODOLOGY

4.1 Background

Continual Learning (CL) In this work, we concentrate on Class Incremental Learning (CIL), a paradigm essential for models to learn new classes sequentially without forgetting previously learned one [54], as shown in figure . In CIL, we denote the set of classes at time t as C_t, with the condition $C_{t-1} \subset C_t$, indicating a progressive expansion of knowledge. The goal is to optimize:

$$\min_{\theta} L(T_t, \theta) + \lambda \sum_{i=1}^{t-1} L(T_i, \theta), \tag{4.1}$$

where θ represents model parameters, and λ balances learning new tasks and remembering old ones, addressing the challenge of catastrophic forgetting.

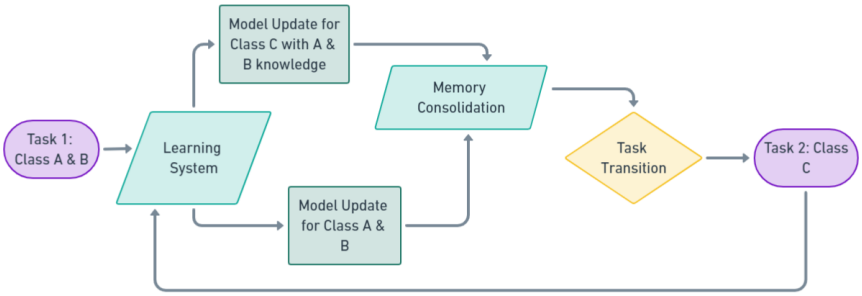

Figure 4.1: Continual Learning

Continual Federated Learning (CFL) aims to equip distributed models with the ability to learn new classes over time without losing previously acquired knowledge, critical for lifelong learning in decentralized networks, as shown in Figure 4.2. We denote the evolving

class set at time t in any node as C_t, where $C_{t-1} \subset C_t$, indicating incremental class expansion. The optimization goal for CFL is succinctly captured as:

$$\min_{\theta} L(T_t, \theta) + \lambda \sum_{i=1}^{t-1} L(T_i, \theta), \tag{4.2}$$

with θ representing the decentralized model parameters and λ a balance between new learning and memory retention across the federated network.

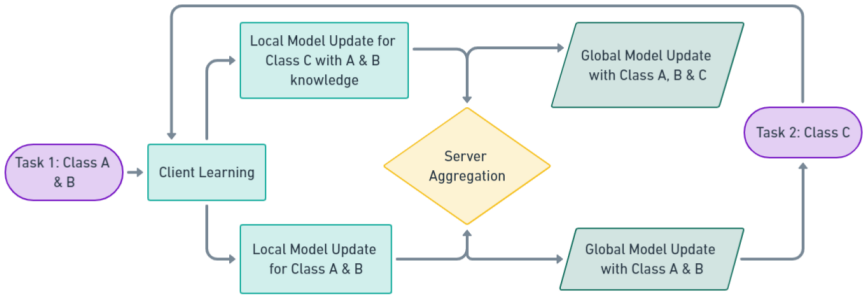

Figure 4.2: Continual Federated Learning

4.2 Continual Federated Learning (CFL)

This section introduces an adapted methodology for Continual Federated Learning, influenced by [1], which incorporates generative replay mechanisms within a federated learning context to address issues of catastrophic forgetting. The approach emphasizes the integration of server-side model consolidation and client-side consistency enforcement as shown in Figure 4.3, taking into account the evolving nature of client data and class information.

Server-Side Model Consolidation: The server initiates each training cycle by synthesizing a global model from the combined parameters of all participating clients $(\Theta_1, \Theta_2, \ldots)$. As new class data emerges from client interactions, the server model is strategically enhanced through a consolidation process utilizing synthetic data produced by client-specific

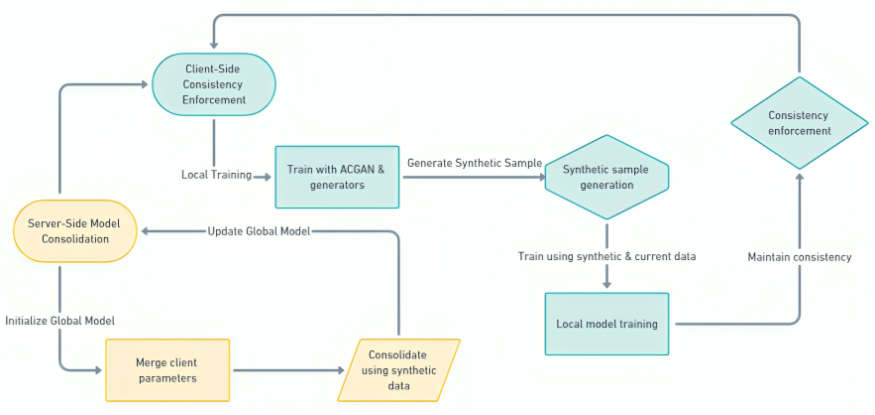

Figure 4.3: Continual Federated Learning Framework (Adapted from [1])

GANs. This consolidation is managed by applying the ACGAN loss function $L_{\text{server}} = L_{\text{acgan}}(\Theta_{\text{global}}, X_g)$, where X_g represents the synthetic samples generated by clients, and Θ_{global} are the server model parameters. This process not only updates the global model with new class knowledge but also ensures a cohesive integration of diverse client insights, enhancing model adaptability and robustness.

Client-Side Consistency Enforcement: During local training, each client employs an ACGAN architecture enhanced by two generators: one from the preceding task ($G^{t-1}_{\Theta_G}$) and the global model generator (G^{g}_{global}). These generators fabricate synthetic samples X^{g}_{t-1} reflective of both past and present tasks, which are incorporated into the current training set X. The local model is trained using the ACGAN loss function $L_{\text{local}} = L_{\text{acgan}}(\Theta, X^{g}_{t-1} \cup X)$. Consistency across varied data types is maintained through specific loss functions (L_{c1}, L_{c2}, L_{c3}), culminating in the total client loss $L_{\text{client}} = L_{\text{local}} + L_c$. This method ensures that each client's model remains stable and effective across changing data landscapes and tasks, thereby minimizing forgetting and promoting consistent performance throughout the federated network.

4.3 Client Similarity and Fairness Evaluation

Next, we leverage the individual fairness definition to examine the relationship between client similarity and model outcomes. Prior to computing the cosine similarity, we apply min-max normalization to the features in S_{features} to ensure that each feature contributes equally to the similarity measure. The cosine similarity between two clients i and j is calculated using the formula:

$$\text{sim}(i, j) = \frac{\mathbf{S}_i \cdot \mathbf{S}_j}{\|\mathbf{S}_i\|\|\mathbf{S}_j\|} \tag{4.3}$$

where $\mathbf{S}_i$ and $\mathbf{S}_j$ are the feature vectors of clients i and j respectively, and $\|\mathbf{S}_i\|$ and $\|\mathbf{S}_j\|$ are the magnitudes of these vectors. This normalization process scales all the features to a fixed range, typically [0, 1], which is crucial for preventing any single feature from disproportionately influencing the similarity scores. We then derive a client-to-client similarity matrix $\mathbf{S}$ using cosine similarity measures on the normalized S_{features} as introduced in Section 3.2.1. The matrix $\mathbf{S}$, where each element $S_{i,j}$ indicates the similarity score between clients i and j, provides a quantifiable measure of similarity that is integral to our fairness assessment. This assessment is conducted using the fairness matrix $\mathbf{F}$ defined in Equations 3.7 and 3.8, and further evaluating fairness through the fairness ratio ρ in Equation 3.9.

4.4 Client Selection in Federated Learning

Client selection is a process in Federated Learning (FL) that decides which clients are chosen for training in each round. This is a crucial step because not all clients may be available for training at all times due to reasons such as network connectivity, device availability, etc [55]. Moreover, including all clients in every round of training can be computationally expensive $O(n)$ and time-consuming.

An effective FL client selection scheme can significantly improve model accuracy, enhance fairness, strengthen robustness, and reduce training overheads. One commonly used technique for Client Selection is **Random Selection** [56]. In random selection, the server

randomly chooses all clients or a subset of clients from the total clients for the training.

However, random selection of clients in Federated Learning, while straightforward, has several disadvantages [57]. One of the primary issues is the risk of over-representation of data. When clients with over-represented data are selected randomly, the model can become biased towards this data, affecting its performance on under-represented data [56]. Furthermore, random sampling of clients in each training round may not fully exploit the local updates from heterogeneous clients, which can result in lower model accuracy. This random selection can also lead to a slower convergence rate of the model. Lastly, the fairness of the model can be degraded as random selection might not ensure that every client gets a fair chance to contribute to the model training.

Opting for subsets of clients rather than the entire client base ameliorates multiple operational challenges. It addresses scalability issues as it curbs the computational and network demands, especially as the client pool expands. This method also respects the varying client availability and reliability, which can be contingent on geographic or technical variables. By training with subsets, models can incorporate a wider data spectrum, bolstering generalization capabilities and curtailing bias toward over-represented client data [14].

Such selection strategies hold significant value in sectors reliant on nuanced data-driven insights. In environmental conservation, FL harnesses selective sensor data to develop predictive models for ecological shifts, balancing developmental needs with environmental preservation [58].

In the automotive sector, particularly for autonomous vehicles, FL's subset selection integrates disparate data, addressing diverse traffic conditions and sensor discrepancies. This refines algorithmic responses to localized contexts, augmenting safety and vehicular autonomy [59].

The retail and logistics domains similarly benefit from FL. Selective data integration from various supply chain components enables sophisticated demand predictions and logistical efficiencies. This strategy adapts to regional market trends and sidesteps data biases, ensuring models are well-fitted and equitable [60].

In order to address the challenges associated with random selection in Federated Learning, we have devised a set of unique strategies. Our proposed strategies for client selection in Federated Learning encompass the following key aspects:

- **Participation-Based Selection (p_i):** This strategy prioritizes clients that had lower participation in previous rounds. The rationale behind this approach is to ensure that all clients have an equitable opportunity to contribute to the model training process. The participation of a client i can be quantified as $p_i = \frac{n_i}{N}$, where n_i is the number of times client i has participated and N is the total number of rounds. We refer to this selection strategy as *Low Participation*.

- **Accuracy-Based Selection (a_i):** This strategy involves selecting clients based on their performance in previous rounds. Specifically, clients that demonstrated lower accuracy in previous rounds are given priority. This approach aims to improve the overall performance of the federated learning model by focusing on clients that could benefit from additional training. We refer to this selection strategy as *Low Accuracy*.

- **Feature Set Average-Based Selection:** This strategy involves selecting clients based on the average of their feature set S_{features}. Clients with a lower average are given priority, with the aim of ensuring a balanced and comprehensive representation of the feature space in the global model. This strategy is particularly beneficial in the context of continual federated learning, where the learning process is an ongoing cycle and the model is continually updated with new data. By prioritizing clients with a lower average feature set, we can ensure that the model stays up-to-date and relevant to all clients, thereby enhancing the overall performance and fairness of the federated learning system. We refer to this selection strategy as *Low Average*.

These strategies are designed to address the challenges associated with random client selection, thereby enhancing the efficiency, fairness, and performance of Continual Federated Learning systems.

4.5 Dynamic Self-Adaptive Client Selection in Federated Learning

Building upon our client selection strategies, we introduce a dynamic self-adaptive algorithm, denoted as $\mathcal{A}_{\text{DSA}}$, for client selection in Federated Learning (FL). This algorithm, $\mathcal{A}_{\text{DSA}}$, intelligently selects the most appropriate client selection technique, based on real-time performance metrics, with a focus on optimizing fairness.We refer to this selection strategy as *Dynamic*.

4.5.1 Algorithm Overview

Let $\mathcal{C} = \{\textbf{Participation-Based}, \textbf{Accuracy-Based}, \textbf{Feature Set Average-Based}\}$ be the set of predefined client selection strategies. The operation of the algorithm $\mathcal{A}_{\text{DSA}}$ is guided by the fairness ratio ρ, as delineated in Equation 3.9.

The algorithmic process for $\mathcal{A}_{\text{DSA}}$ is detailed as follows:

By dynamically shifting between client selection strategies based on the fairness ratio, $\mathcal{A}_{\text{DSA}}$ ensures that the FL system adaptively optimizes for fairness. This approach not only addresses the variability in client contributions and system performance but also maintains the principle of fairness, allowing each client a proportionate opportunity to influence and benefit from the federated model. The strategic adaptability embedded in $\mathcal{A}_{\text{DSA}}$ substantially elevates the resilience and fairness of the CFL ecosystem.

The implementation of the algorithms discussed above, along with the fairness metrics, is available on GitHub at https://github.com/noornaima/Fairness-in-Continual-Federated-Learning. This ensures that the results can be easily replicated and further explored by other researchers.

Algorithm 1 Dynamic Self-Adaptive Client Selection ($\mathcal{A}_{\text{DSA}}$)

Require: Set of client selection strategies $\mathcal{C}$, fairness ratio ρ, fairness threshold θ_{fair}.

Ensure: Fair client selection in FL training rounds.

Initialize $\mathcal{C}_{\text{current}}$ to an initial strategy from $\mathcal{C}$
for each FL training round **do**
    Compute fairness ratio ρ for the current model state
    if $\rho \geq \theta_{\text{fair}}$ **then**
        Continue with $\mathcal{C}_{\text{current}}$
    else
        Select $\mathcal{C}_{\text{next}}$ from $\mathcal{C}$ with the potential to increase ρ
        Set $\mathcal{C}_{\text{current}} = \mathcal{C}_{\text{next}}$
    end if
    Apply $\mathcal{C}_{\text{current}}$ for client selection
    Re-evaluate ρ after client selection
    Update $\mathcal{C}_{\text{current}}$ if necessary to ensure $\rho \geq \theta_{\text{fair}}$
end for
return The updated set of selected clients for the next training round

4.6 Evaluation Metrics

Average Accuracy (AA) represents the overall effectiveness of the learning algorithm across the continual learning process. It is defined as:

$$AA = \frac{1}{N} \sum_{i=1}^{N} A_{i,N},$$

where $A_{i,N}$ is the accuracy on task i after the model has been trained on all N tasks. This measure provides insight into the global learning capability and generalization performance of the model across a diverse set of tasks.

Backward Transfer (BWT) quantifies the impact of learning new tasks on the performance of previously learned tasks. It is calculated as:

$$BWT = \frac{1}{N-1} \sum_{i=1}^{N-1} (A_{i,N} - A_{i,i}),$$

where $A_{i,j}$ denotes the accuracy on task i after training on task j, with N being the total number of tasks. A positive BWT indicates beneficial transfer, whereas a negative BWT indicates detrimental forgetting.

Temporal accuracy (TA) is assessed at the conclusion of each task within the continual learning sequence. It provides insights into the model's knowledge retention and acquisition efficacy at various stages of learning. For a sequence of N tasks, the temporal accuracy after task j is given by the formula:

$$TA_j = \frac{1}{j} \sum_{i=1}^{j} A_{i,j},$$

where $A_{i,j}$ is the accuracy of the model on task i after training up to and including task j. This metric emphasizes the evolving nature of the learning process and highlights the model's adaptability to new information over time.

4.7 Experimental Setup

We performed our evaluation using the EMNIST-Balanced dataset [61], organizing the study into five distinct tasks for each client, with each task encompassing two classes. Our setup

included a total of five clients, but to streamline the analysis within the federated learning (FL) framework and to specifically address the complexities inherent in continual learning (CL), we selected a subset of three out of the five clients for participation in all experiments. This approach simplifies the analysis by focusing on a consistent group of clients. For the local training, we set the number of iterations to $T = 400$ and global communication round $R = 200$ for all models. With five tasks allocated per client, this configuration results in an effective 40 rounds dedicated to each task. To ensure the reliability of our findings, each experiment was repeated four times, and we report the mean of these results.

Chapter 5

RESULTS

This section analyzes the effectiveness of client selection strategies for fairness in continual federated learning. We evaluate how different strategies impact both individual fairness and model performance over time. A comparative analysis is provided, focusing on the variability and stability of these strategies and their influence on learning and fairness within a dynamic data environment.

5.1 Fairness using Delta Accuracy Fairness (DAF) matrix

Fairness-as-whole: In evaluating the individual fairness of the resulting system, as expressed by the Individual Fairness criterion (eq. 3.1), our analysis depicted in Figure 5.1, uncovers varied patterns of fairness across different participation and strategy frameworks. Utilizing the Delta Accuracy Fairness metric, as formulated in equation (**??**), we note that strategies marked by low participation and Low Average tend to demonstrate a convergence in the accuracy improvements among clients. This convergence is indicative of a fairness approach that aims to equalize outcomes for clients with similar levels of engagement, striving to meet the fairness criterion established by Individual Fairness.

In stark contrast, the randomness inherent to the Random selection strategy is associated with significant discrepancies in client participation and, consequently, a broad range of outcomes in terms of knowledge retention. This lack of consistency in the treatment of clients, particularly when clients have equivalent levels of participation, signifies a divergence from the ideal of individual fairness.

Meanwhile, the Low Accuracy strategy, despite fluctuations in both client accuracy and participation, seems to maintain a closer adherence to fairness principles. The relatively

consistent treatment of similarly situated clients aligns more closely with the individual fairness standard.

Adding to this, the Dynamic Algorithm strategy shows an interesting pattern where clients with comparable engagement levels tend to achieve similar accuracy scores. This suggests an underlying mechanism within the Dynamic Algorithm that recognizes and possibly compensates for the varying degrees of client engagement, thus striving for individual fairness. While there is some variability, it does not overshadow the strategy's potential to maintain parity among clients who share similar engagement levels, which is a positive indicator of the strategy's alignment with the individual fairness criterion.

Thus, our findings—bolstered by the data represented in Figure 5.1 —suggest that the Low Participation and Low Average strategies support the realization of individual fairness to varying extents within the system. Moreover, the Dynamic Algorithm strategy, through its adaptive nature, indicates a promising avenue for achieving fairness among similarly engaged clients, thus contributing an important dimension to the fairness discourse.

Temporal Fairness: In our analysis of temporal fairness in CFL (eq. 3.2), we began by computing a fairness matrix F (eq. 3.7) to assess the model's fairness over time, setting a fairness threshold at less than 0.1. We then computed the fairness ratio ρ (eq. 3.9), considering instances with a ratio of 0.8 or higher as fair, allowing us to quantify fair instances as a percentage and establish a clear metric of fairness per round. The values were calculated by taking their percentage to provide a clear representation of fair instances across different rounds.

Data analysis from Table 5.1 revealed significant variability in temporal fairness across different client selection strategies. The Random strategy showed notable inconsistency and lower temporal fairness over time, attributed to its random client selection process. This led to substantial fluctuations in its fairness ratio, underscoring its unpredictability and inefficiency in maintaining fairness.

Conversely, the Low Participation and Low Average strategies exhibited greater temporal

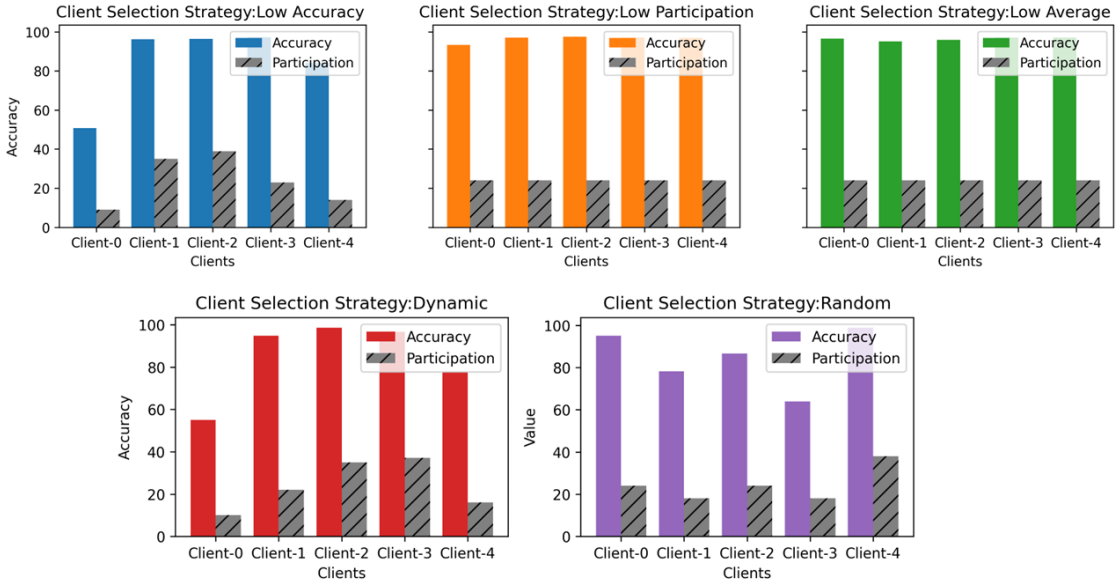

Figure 5.1: Comparative Analysis of Client Performance and Engagement across Client Selection Strategies

stability and alignment with the fairness threshold, indicating their effectiveness in promoting equitable client participation and outcomes. Particularly, the Low Participation strategy showed a strong commitment to equal contribution across clients, aligning closely with our fairness threshold.

The Low Accuracy strategy, while performing better than the Random strategy, displayed some variance, suggesting less stability in fairness. However, it maintained a relatively fair average, indicating its potential to achieve moderate fairness over time.

Interestingly, the Dynamic Algorithm strategy presented a mixed picture. While it generally demonstrated a trend towards fair outcomes, especially in Task 5, the results indicate a higher variance in earlier tasks. This variance, however, lessened in later tasks, aligning more closely with our threshold for temporal fairness. This suggests that the Dynamic Algorithm may require a period of adaptation before achieving a more consistent level of fairness,

particularly for tasks that necessitate ongoing client engagement.

Our findings indicate that the **Low Participation** and **Low Average** strategies are notably effective in fostering temporal fairness in CFL. The Dynamic Algorithm strategy, given its adaptive nature, also holds potential for promoting fairness, especially as it appears to adjust over time to better accommodate fair outcomes across clients. This underscores the critical role of strategic client selection and adaptation mechanisms in improving fairness, potentially influencing fairness trajectories positively over time, which is vital for the development of equitable continual learning systems.

Table 5.1: Temporal Fairness Across Tasks for Different Client Selection Strategies on the basis of Delta Accuracy Fairness (DAF) matrix, highlighting top significant performance per task in bold. Each task (Task 1 to Task 5) represents sequential learning of a new class in a class incremental learning scenario. Values before and after ',' denote mean and standard deviation, respectively, based on four runs, illustrating each strategy's temporal fairness consistency. Results significance computed in t-test comparison with Random are demonstrated with $***$ for p-value < 0.001, $**$ for p-value < 0.005, and $*$ for p-value < 0.05.

Strategy	Task 1	Task 2	Task 3	Task 4	Task 5
Random	$0.22, 0.02$	$0.13, 0.06$	$0.35, 0.18$	$0.36, 0.35$	$0.29, 0.15$
Low Accuracy	$0.3, 0.06$	$0.3, 0.09$	$0.45^{***}, 0.13$	$0.80^{***}, 0.08$	$0.92^{***}, 0.06$
Low Participation	$\mathbf{0.25}^{*}, \mathbf{0.06}$	$0.21, 0.02$	$0.23, 0.03$	$\mathbf{0.91}^{***}, \mathbf{0.038}$	$0.74^{***}, 0.11$
Low Average	$0.28, 0.03$	$0.32, 0.17$	$\mathbf{0.84}^{***}, \mathbf{0.03}$	$0.47^{***}, 0.05$	$\mathbf{0.93}^{***}, \mathbf{0.02}$
Dynamic	$\mathbf{0.30}, \mathbf{0.30}$	$\mathbf{0.31}^{*}, \mathbf{0.07}$	$0.30, 0.15$	$\mathbf{0.68}^{***}, \mathbf{0.23}$	$\mathbf{0.90}^{***}, \mathbf{0.09}$

Expanding our evaluation from fixed thresholds to a continuum, we assess the sustainability of fairness across an entire range of possible fairness ratio thresholds, from 0 to 1. This comprehensive approach is depicted in Figure 5.2, which showcases a dynamic evaluation of fairness under varied threshold conditions for different strategies.

Within this framework, the fairness ratio at any given threshold is indicative of how closely a strategy adheres to the desired fairness standard, with higher ratios signaling a better approximation to the ideal fairness level.

The 'Low Accuracy' strategy, as observed in Figure 5.2, upholds higher fairness ratios at more stringent thresholds across most tasks, suggesting its robustness in maintaining fairness when the model aims to achieve an optimal fairness standard. Conversely, the 'Low Participation' and 'Low Average' strategies demonstrate variability in performance at different thresholds, which hints at their relative strengths in specific task conditions. The 'Random' strategy typically falls behind the more methodical approaches, with this trend most apparent in Tasks 3 and 4.

Adding to this, the 'Dynamic' strategy showcases a consistent performance across all thresholds. It particularly stands out in Task 2, maintaining high fairness ratios even at stricter thresholds. This indicates the strategy's adeptness in adapting to varying fairness requirements across different tasks.

Task 5 is noteworthy for illustrating the efficacy of the 'Low Average' strategy, which excels particularly at the 0.6 threshold, signifying its substantial contribution to fairness under defined operational constraints.

The performance variations across tasks and thresholds captured in Figure 5.2 reveal the intrinsic complexities of CFL systems. These variations suggest that no single strategy may consistently realize fairness across every context. Instead, the choice and implementation of client selection strategies ought to be tailored to the specific demands of each task and the desired fairness target. The dynamic shifts in fairness ratios across thresholds also highlight the importance of developing adaptive algorithms that are capable of sustaining fairness in the face of a continuously evolving learning environment.

Statistical Analysis To gauge the effectiveness of our client selection strategies in promoting temporal fairness, as defined by the Delta Accuracy Fairness (DAF) matrix (eq 3.7), pairwise t-tests were conducted comparing each strategy against a baseline random strategy

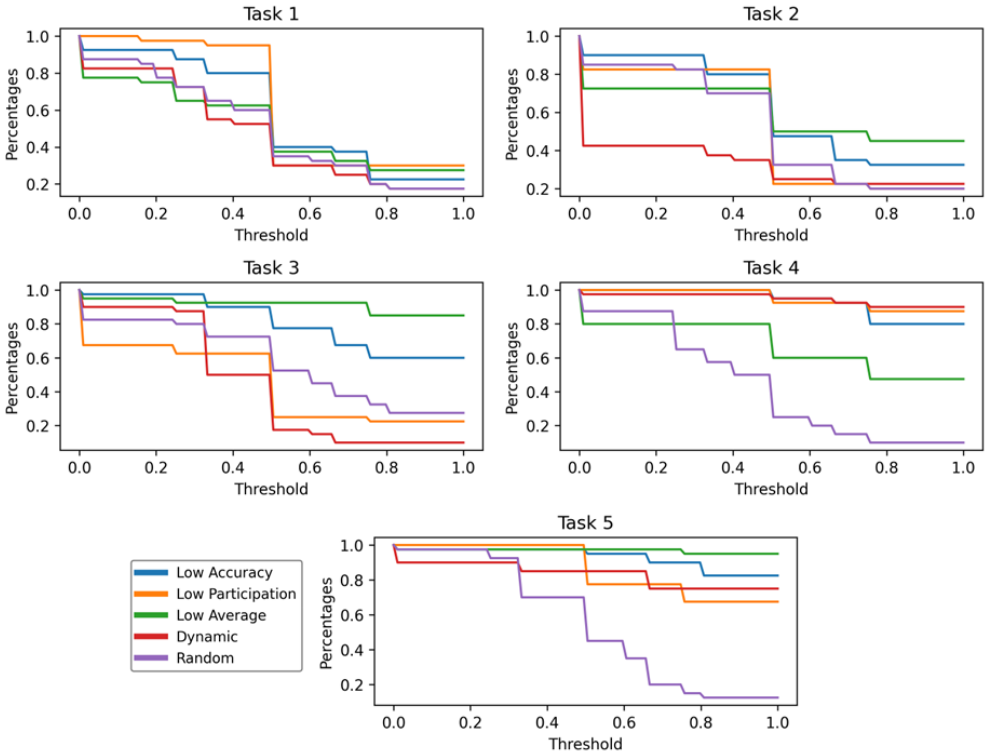

Figure 5.2: Temporal Fairness Across Tasks for Different Client Selection Strategies on the basis of Delta Accuracy Fairness (DAF) matrix using different thresholds. The y-axis represents percentages, providing a clear representation of fair instances across different rounds.

(Table 5.1). These tests, essential for assessing the statistical validity of temporal fairness achievements, are denoted by asterisks in Table 5.1, indicating significant differences from the random baseline.

At the outset in Task 1, no statistically significant difference is noted between the random and low accuracy strategies, suggesting equivalent performance in temporal fairness. However, the low participation strategy demonstrated a statistically significant difference, flagged by an asterisk, pointing to potential areas for enhancing our fairness matrix.

As we proceed beyond the initial tasks, a pattern emerges where all strategies, notably

from Task 3 onward, begin to deviate significantly from the random strategy, suggesting improved temporal fairness. The Dynamic strategy starts to show its strengths as early as Task 2, marked by a single asterisk, signaling a significant advantage over the random approach. This significant improvement in the Dynamic strategy is sustained through Tasks 4 and 5, where it reaches a high level of significance, as denoted by three asterisks.

The observed fluctuations in statistical significance across tasks are indicative of the complexities endemic to continual federated learning environments. These complexities are characterized by changing data distributions and concept drift, which can challenge the maintenance of temporal fairness. The relationship between evolving environmental conditions and the efficacy of client selection strategies manifests in the variable significance of our t-test results, emphasizing the dynamic nature of achieving fairness.

5.2 Fairness using Delta Forgetting Fairness (DFF) matrix

Fairness-as-whole: In our examination of CFL through the lens of the Individual Fairness criterion, detailed in equation 3.1, Figure 5.3 serves as a visual inquiry into the interplay between client participation and forgetting measures across differing client selection strategies. The principle of Individual Fairness posits that clients with equivalent levels of participation should, ideally, experience comparable rates of knowledge retention. Yet, as Figure 5.3 illustrates, this parity proves challenging to achieve.

The data gleaned from the Random strategy particularly underscores this challenge, revealing pronounced discrepancies where similar levels of participation do not equate to similar forgetting rates. This is exemplified by Client-1 and Client-3, whose parallel stripes of participation are contrasted by disparate forgetting measures, suggesting an infringement of the Individual Fairness criterion.

While the Low Participation and Low Average strategies appear to more closely mirror the expectations of Individual Fairness, as seen in the more uniform distribution of forgetting measures relative to participation levels, anomalies persist. For example, within the Low Accuracy strategy, Client-0 and Client-2 show the same striped pattern for participation

yet diverge in forgetting measures, highlighting an inherent inconsistency in achieving fair outcomes.

The Dynamic strategy, while embodying flexibility in client selection, reflects a similar complexity. Instances of equivalent participation, like those observed for Client-1 and Client-3, do not consistently result in matching forgetting measures. This inconsistency accentuates the intricate challenge of calibrating a dynamic system to fulfill the stringent requirements of Individual Fairness.

From these observations, it becomes evident that ensuring fairness in forgetting measures solely based on participation level is a formidable task, as demonstrated by the variability across all strategies in the image. It suggests that while participation is a significant factor, achieving Individual Fairness in practice is a multifaceted issue that may not be entirely resolved by aligning participation metrics.

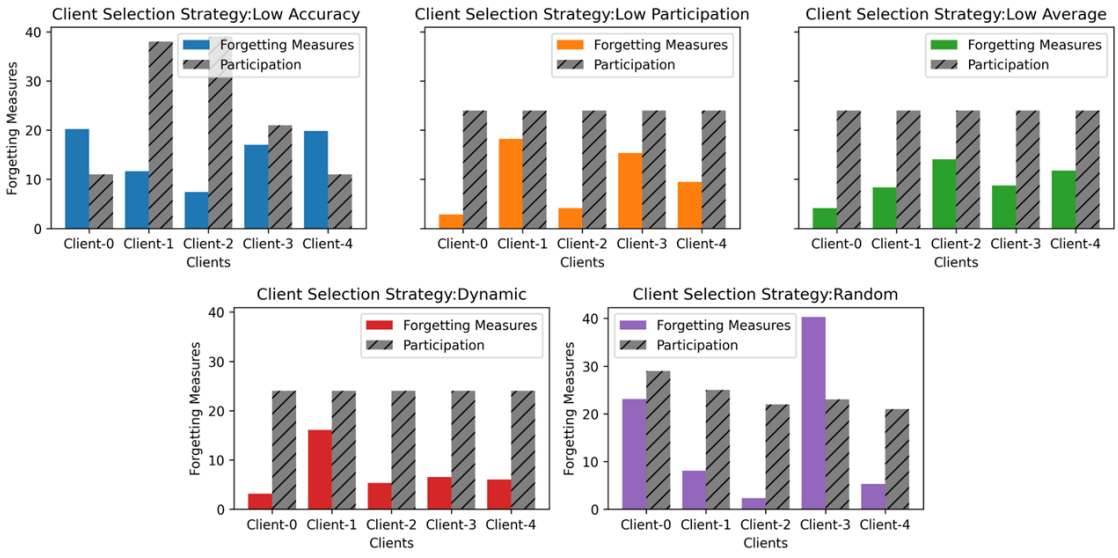

Figure 5.3: Comparative Analysis of Client Forgetting Rates and Engagement across Client Selection Strategies

Temporal Fairness: In the examination of temporal fairness within CFL, detailed in equation 3.2, the initial step involved calculating a fairness matrix F (refer to equation 3.8) to evaluate the model's temporal fairness. A fairness threshold was established at less than 0.1. Subsequently, the fairness ratio ρ was computed as delineated in equation 3.9, with ratios of 0.8 or greater deemed as fair. This approach facilitated the quantification of fair instances as a percentage, thereby defining a precise metric of fairness for each round. The values were expressed as percentages to clearly represent the fair instances across different rounds.

As demonstrated in Table 5.2, the Random strategy is characterized by its high variance, which underscores the unpredictability inherent in its design. This strategy, due to its stochastic nature, often results in a wide range of outcomes, affecting its reliability in fostering equitable client participation over time. While it can occasionally achieve reasonable fairness, its inconsistency across tasks raises concerns about its reliability in maintaining temporal fairness over time.

In contrast, the Low Accuracy strategy exhibits a more consistent performance and reduced variability, showing a more stable and dependable approach in achieving temporal fairness. Although it does not always attain the highest mean fairness, its stability across runs highlights its reliability, especially in tasks where it outperforms other strategies. However, it presents mixed results in maintaining fairness across different tasks, indicating a potential trade-off between its general consistency and achieving optimal fairness levels.

The Low Participation strategy is distinguished by its exceptionally high fairness levels in specific tasks, demonstrating a strong ability to maintain equitable participation with low variance, which indicates its reliability. Its consistent performance across various iterations underscores its potential as an effective method for ensuring fairness. However, the effectiveness of the Low Participation strategy is not uniform across all tasks; while it achieves high fairness scores and shows a strong ability in certain situations, its performance variability suggests that it may not be universally reliable.

Meanwhile, the Low Average strategy maintains a strong performance across various

tasks, indicating a robust and stable approach to temporal fairness. This strategy's ability to achieve high levels of fairness with lower variability makes it a preferable option for sustained equitable client selection.

The Dynamic strategy also demonstrates commendable performance across a variety of tasks, showing particular strength in Task 2 with the highest mean fairness but maintaining robust results in other tasks as well (e.g., Task 5 with a mean of 0.655 and a standard deviation of 0.271). Its adaptability and effectiveness in achieving higher levels of fairness across different learning scenarios suggest it might be particularly valuable in environments requiring flexible responses to changing dynamics.

The analysis underscores the complexity of achieving temporal fairness in CFL, where no single strategy universally excels across all tasks. However, **the Low Participation** and **Low Average** strategies emerge as particularly effective, offering a balance of high fairness and consistency. This insight is crucial for designing client selection mechanisms that not only promote fairness but also adapt effectively to the temporal dynamics of CFL, thus contributing to the development of more equitable and resilient learning systems.

Upon closer examination of the limitations of each strategy, we observe that: the Random strategy's high variance reflects its unpredictability and inability to systematically address disparities in forgetting rates among users, as measured by the DFF metric. The Low Accuracy strategy's focus on lower-performing clients doesn't always translate into improved fairness, highlighting a potential misalignment with the DFF's goal of equalizing knowledge retention. The Low Participation strategy, although effective in certain contexts, does not consistently address the fairness requirements across all tasks, suggesting its limited applicability. Meanwhile, the Dynamic strategy, while adaptable and generally effective, may still fall short in situations where rapid shifts in learning paradigms occur, challenging its ability to maintain fairness under all conditions. Lastly, the Low Average strategy's general approach might miss subtle fairness issues, indicating that even well-performing strategies can have limitations in fully aligning with the fairness objectives outlined by the DFF metric.

Transitioning from a singular, fixed threshold, our analysis adopts a fluid approach by

Table 5.2: Temporal Fairness Across Tasks for Different Client Selection Strategies on the basis of Delta Forgetting Fairness (DFF) matrix, highlighting top performers in bold. Each task (Task 2 to Task 5) represents sequential learning of new classes in a class incremental learning scenario, focusing on the effectiveness of each strategy in minimizing the forgetting of previously learned classes. The examination of fairness begins with Task 2, since there is no previously learned knowledge in Task 1 against which to measure forgetting. Values before and after ',', denote mean and standard deviation, based on four runs, illustrating each strategy's temporal fairness consistency. Results significance computed in t-test comparison with Random are demonstrated with $***$ for p-value < 0.001, $**$ for p-value < 0.005, and $*$ for p-value < 0.05.

Strategy	Task 2	Task 3	Task 4	Task 5
Random	$0.47, 0.37$	$0.16, 0.26$	$0.46, 0.179$	$0.27, 0.11$
Low Accuracy	$\mathbf{0.35}^{*}, \mathbf{0.047}$	$\mathbf{0.51}^{***}, \mathbf{0.12}$	$0.46, 0.12$	$\mathbf{0.65}^{***}, \mathbf{0.24}$
Low Participation	$\mathbf{0.48}^{***}, \mathbf{0.07}$	$\mathbf{0.18}^{***}, \mathbf{0.01}$	$\mathbf{0.91}^{***}, \mathbf{0.03}$	$\mathbf{0.29}^{*}, \mathbf{0.04}$
Low Average	$0.51, 0.042$	$\mathbf{0.34}^{***}, \mathbf{0.21}$	$\mathbf{0.54}^{*}, \mathbf{0.16}$	$\mathbf{0.86}^{***}, \mathbf{0.26}$
Dynamic	$\mathbf{0.59}^{***}, \mathbf{0.09}$	$0.46^{***}, 0.13$	$0.4, 0.21$	$0.65, 0.27$

evaluating a full spectrum of fairness ratio thresholds from 0 to 1. This methodology enables us to scrutinize the strategy's enforcement of fairness at every juncture within this range, offering a holistic view on the versatility of fairness measures under varying conditions, as shown in Figure **??**. The 'Low Accuracy' strategy upholds a consistent level of fairness at the lower thresholds across Tasks 2 to 5, suggesting its suitability for settings where strict fairness is paramount. Nevertheless, as the threshold becomes more lenient, this strategy's performance begins to diverge, which may indicate sensitivity to the strictness of the fairness criteria. The 'Low Participation' strategy displays notable resilience in Task 4, with a high fairness ratio across the entire threshold range, signifying its adaptability and effectiveness

in varied CFL scenarios.

The 'Dynamic' strategy shows a particularly compelling performance in Task 2, where it maintains high fairness ratios across a broad range of thresholds, reflecting its capability to dynamically adjust to different fairness demands. It illustrates the strategy's adeptness in responding to changing conditions within the CFL environment, suggesting a well-tuned balance between fairness and participation rates.

Task 5 is distinctive, revealing the 'Low Average' strategy's pronounced proficiency at more relaxed thresholds, highlighting its efficiency in conditions that do not require stringent fairness protocols. In contrast, the 'Random' strategy, often serving as a baseline for comparison, displays a diminishing fairness value as thresholds ascend, notably in Tasks 3 and 5. This observation suggests its relative ineffectiveness in maintaining fairness when the criteria are less demanding.

The patterns discerned from the various strategies and their respective performances as thresholds fluctuate, as depicted in Figure **??**, reinforce the notion that temporal fairness in CFL systems is a complex objective, without a one-size-fits-all solution. Each strategy's effectiveness is influenced by the specific task and the selected fairness threshold. The shifts in performance with changing thresholds accentuate the necessity for adaptive strategies that are capable of satisfying diverse fairness requirements, thus ensuring just and equitable learning opportunities for all participants in the CFL ecosystem.

Statistical Analysis To discern the effectiveness of our client selection strategies in attaining temporal fairness, as operationalized by the Delta Forgetting Fairness (DFF) matrix, we conducted pairwise t-tests in comparison to a baseline random strategy (Table 5.2). These significance tests, indicated by asterisks in the table, assess the statistical distinction between the examined strategies and the random baseline. This assessment aids in identifying which strategies successfully mitigate forgetting, a vital component of maintaining fairness in CFL systems where clients may experience divergent rates of knowledge decay.

The analysis illuminated substantial variations in the effectiveness of the strategies across

different tasks. The Low Accuracy strategy demonstrated significant gains in Task 2, signaling an effective reduction in forgetting when compared to the baseline. This tendency toward significant improvement persisted in Task 3, where the strategy achieved highly significant results, implying a robust approach to preserving fairness in knowledge retention. The Low Participation strategy also showed formidable significance in Task 3, validating its effectiveness in that particular context. However, the Low Accuracy strategy did not sustain its performance into Task 4, where it failed to show a significant difference from the random strategy. On the other hand, the Low Participation strategy continued to show a significant reduction in forgetting rates in Task 4, reflecting its adaptability and consistent effectiveness across various settings. Task 5 underscored the strength of the Low Average strategy, which exhibited highly significant results, suggesting that it can significantly enhance fairness by effectively managing forgetting rates among clients, even in a continuous learning environment.

The Dynamic strategy also stood out, particularly showing highly significant improvements in Task 2. This pattern suggests that the Dynamic strategy, with its adaptability and potential to fine-tune client selection, can be effective in reducing forgetting rates in certain contexts. Nonetheless, this strategy did not demonstrate consistent significance across all tasks, suggesting that while it has the potential for high performance, it may require further tuning to ensure consistency across different learning scenarios.

Confronted with the inherent obstacles of non-stationary data distributions and concept drift, which influence model performance and the stability of the learning process, achieving temporal fairness is a formidable challenge. Variations in data quality and volume among clients further compound these issues. Our analyses, underpinned by pairwise t-tests, demonstrate that specific strategies, including the Dynamic strategy, significantly attenuate forgetting disparities, thus enhancing temporal fairness. These strategies are instrumental in ensuring that no client is disproportionately affected by the natural process of forgetting, a critical factor for the ethical sustainability of CFL systems. By moderating the learning dynamics and accommodating the complexities of knowledge retention, these strategies bolster

the overall robustness and fairness of the CFL ecosystem.

5.3 Comparative Analysis of different Client Selection Strategies

In our CFL setup, we assess various client selection strategies for their impact on model accuracy. Figure 5.4 provides a visual comparison of these strategies over multiple rounds. The Random strategy exhibits considerable volatility in performance, characterized by its frequent convergence with the lower confidence bounds of other strategies, signaling its relative unpredictability and potential suboptimality.

The **Low Participation** strategy demonstrates superior performance, maintaining higher accuracy levels across all tasks. This consistency suggests that selecting clients based on their participation rate can substantially improve the learning efficacy.

Conversely, the Dynamic strategy, while not as consistently high-performing as the Low Participation strategy, shows adaptability across tasks. Its performance is notably competitive, particularly in later tasks, indicating its potential as a viable strategy in environments where model performance can benefit from dynamically adjusting client selection based on evolving learning contexts.

Backward Transfer (BWT) The BWT compares the client selection strategies in terms of their impact on previously learned tasks. BWT measures the influence that learning new tasks has on the performance of old tasks, where negative values indicate forgetting.

The **Backward Transfer (BWT)** compares the client selection strategies in terms of their impact on the retention of previously learned tasks. BWT measures the influence that learning new tasks has on the performance of old tasks, with negative values representing a loss of previous knowledge, commonly referred to as forgetting.

As shown in Figure 5.5, all client selection strategies tested demonstrate negative BWT, suggesting that the introduction of new tasks results in some forgetting across the board. Notably, the Low Participation and Low Average strategies are characterized by less negative BWT, highlighting their relative effectiveness in mitigating the impact of learning new

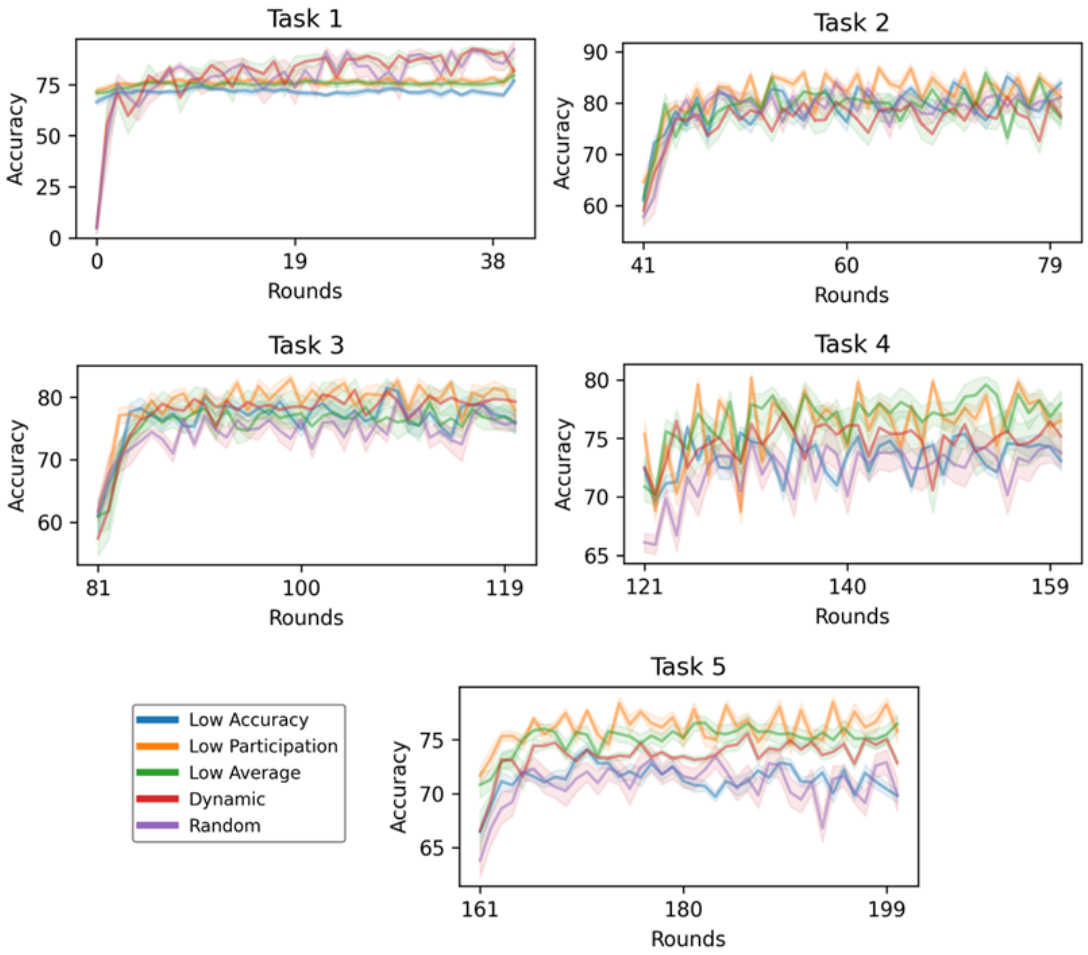

Figure 5.4: Task Temporal Accuracy for Different Client Selection Strategies

information on previously acquired knowledge.

In comparison, the Dynamic and Random strategies show greater and similar levels of negative BWT. This outcome suggests that these strategies, while capable of adapting to or covering a broad range of learning experiences, do not necessarily translate into better retention of past knowledge. The variability in the Dynamic strategy's BWT might reflect its adaptive learning process, which does not consistently outperform more stable strategies like **Low Participation** and **Low Average** in terms of retaining old information.

These results underscore the importance of strategy selection in continual learning sce-

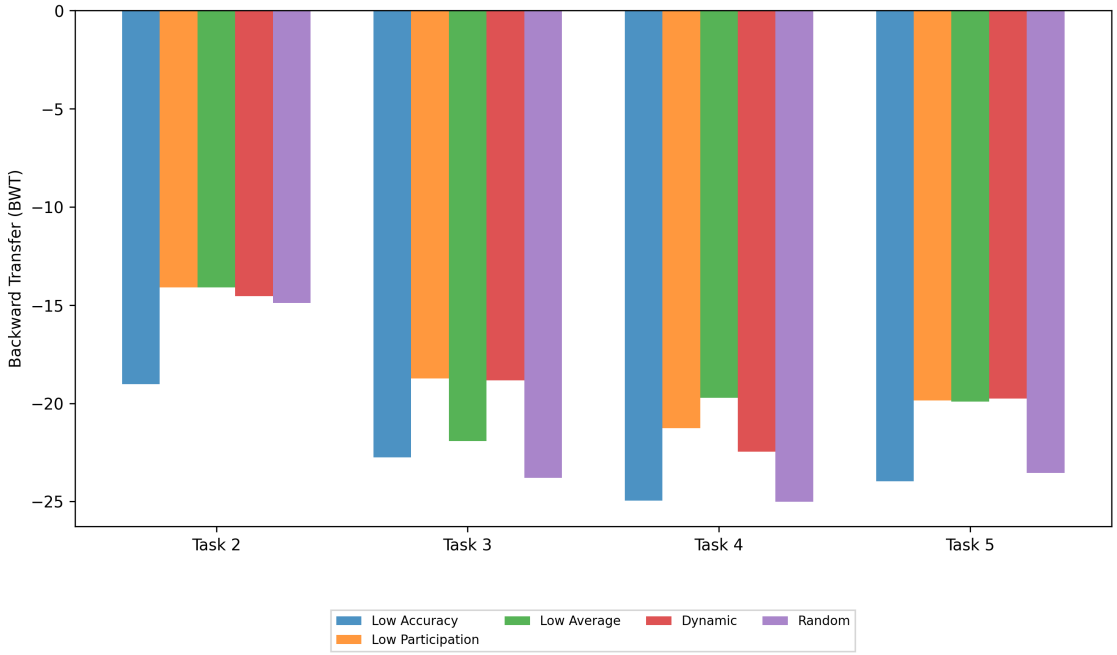

Figure 5.5: BWT across Tasks for different Client Selection Strategies (Note: Starting from Task 2 due to no previous knowledge in Task 1)

narios, particularly when the preservation of past knowledge is crucial and must be carefully considered against the demands of assimilating new information.

5.4 System Wide Performance Analysis

In the context of our overall system performance assessment, the client selection strategies are critical in the continual federated learning framework.

The **Low Participation** strategy exhibits strong performance over the *Random* baseline, with a higher mean accuracy ($M = 84.57$) and lower standard deviation ($SD = 1.97$), underscoring its robustness. Notably, the **Low Average** approach excels, achieving the highest mean accuracy ($M = 84.93$) and presenting the least variability ($SD = 0.67$), proving to be the most consistent and reliable strategy.

The **Dynamic** strategy also performs well, with a mean accuracy ($M = 81.22$) that

Table 5.3: Average Accuracy of Different Client Selection Strategies. Bold values represent strategies with superior performance. SD denotes the standard deviation measured based on four runs, illustrating each strategy's overall accuracy

Strategy	Mean Accuracy (%) , SD
Random	$76.86, 2.03$
Low Accuracy	$75.44, 4.95$
Low Participation	$\mathbf{84.57, 1.97}$
Low Average	$\mathbf{84.93, 0.67}$
Dynamic	$81.22, 3.10$

exceeds the *Random* baseline and exhibits a reasonable level of variability ($SD = 3.10$). This performance indicates that the **Dynamic** strategy is a strong contender that offers a balance between accuracy and consistency.

In contrast, the *Random* strategy achieves a moderate mean accuracy ($M = 76.86$) with increased fluctuations ($SD = 2.03$), indicating less stability. The *Low Accuracy* technique has the lowest mean accuracy ($M = 75.44$) and the highest standard deviation ($SD = 4.95$), which may point to its inconsistent and potentially less optimal performance. These findings are detailed in Table 5.3.

Chapter 6

DISCUSSION

In this work, we introduced two novel fairness matrices tailored specifically for the distinct contexts of Federated Learning (eq 3.7) and Continual Learning (eq 3.8) within CFL systems. This approach ensures that clients are treated equitably over time, with each matrix designed to address unique aspects of fairness specific to either federated or continual learning scenarios. Our fairness metrics comprehensively account for client behavior throughout the learning process, underscoring our commitment to fostering fairness as a dynamic and continuous priority within these evolving educational environments.

Our findings from multiple tasks, presented in Tables 5.1 and 5.2, indicate that no strategy perfectly ensures fairness across all conditions. However, the **Low Participation** and **Low Average** strategies consistently outperform others, particularly the Random selection method, by offering more stable and equitable outcomes.

The Random strategy's variability underscores its unpredictability, while the Low Accuracy strategy shows mixed results, indicating its limitations in different tasks. Conversely, the Low Participation strategy excels in promoting equitable participation, and the Low Average strategy demonstrates robustness across various tasks, proving effective in maintaining fairness.

In addition to these strategies, our exploration of the Dynamic strategy has yielded insightful results. It has demonstrated a strong capability for adaptability, evidenced by its significant outperformance of the Random baseline in some taks and its consistency across multiple thresholds. As shown in Tables 5.1 and 5.2, the Dynamic strategy dynamically adjusts its parameters in response to evolving data characteristics, maintaining accuracy and fairness across all groups. Its potential to achieve high levels of fairness, suggests its

utility as a valuable addition to the CFL system. This adaptability is crucial as it allows the system to evolve alongside the clients it serves, ensuring that fairness remains a continuous priority in an ever-changing educational landscape.

The DAF and DFF matrices provide a solid framework for assessing and promoting fairness in CFL. The inclusion of the Dynamic strategy, with its adaptive nature, adds an important dimension to our work. Continual refinement of these strategies, coupled with the exploration of adaptive methods like the Dynamic strategy, will enhance the efficacy and fairness of CFL systems, leading to more equitable learning environments.

Practical Implications: Our empirical findings underscore the crucial role of strategic client selection in creating a fairer learning environment. Our findings emphasize the importance of considering temporal aspects of fairness in CFL. By implementing the different client selection strategies, CFL models can be more consistently fair over time. This is a step forward in making sure that as learning models evolve, they do so in a way that is just and equitable for all participants consistently over time.

Theoretical Implications: Our work has a main theoretical implication on creating a forum and being the first step towards reconsidering how fairness should be calculated for advanced AI systems such as CFL. We believe as these technologies are advancing at a great pace, there is a lack of consideration on how the research community should be measuring fairness and plan for strategies to attain the newly proposed metrics. As our work has demonstrated, it is important to revisit some of the traditional metrics of fairness.

Future Work: Future work should explore the application of reinforcement learning for client selection strategies to enhance fairness in CFL systems. This approach could dynamically adapt to changing conditions and data distributions, optimizing the balance between knowledge retention and fairness. Integrating continual learning measures reflective of the evolving environment, and deploying reinforcement learning algorithms could provide a

robust framework for ensuring equitable knowledge distribution among clients. Investigating these sophisticated techniques will be critical for advancing adaptable and just CFL systems that respond effectively to the dynamic nature of real-world data.

Furthermore, it will be imperative to apply and evaluate our fairness metrics in conjunction with a variety of continual learning methods. This would ensure that the balance between knowledge retention and fairness is not specific to one method but is broadly applicable and effective across different continual learning paradigms. Integrating and testing these metrics with diverse continual learning algorithms will provide a comprehensive understanding of their efficacy and robustness, contributing to the development of CFL systems that are both adaptable and equitable. Additionally, these metrics should be tested with other datasets to validate their effectiveness and generalizability across various scenarios.

While this study primarily focused on individual fairness, future research should also consider group fairness. Evaluating and ensuring fairness across different groups of clients will be crucial for creating more inclusive CFL systems. Moreover, as this work utilized only five clients, which is relatively few, future studies should involve a larger number of clients to better understand and address scalability issues and to ensure the robustness of the fairness metrics.

Additionally, the fairness matrix introduced in this work does not take into consideration the non-IID (non-Independent and Identically Distributed) nature of federated learning. Future research should address this limitation by developing and incorporating fairness measures that account for the heterogeneity of data distributions across clients.

As we endeavor to create just CFL systems, it will be crucial to investigate these sophisticated techniques, ensuring they can robustly handle the complexities and dynamic nature of real-world data and learning scenarios.

BIBLIOGRAPHY

[1] Daiqing Qi, Handong Zhao, and Sheng Li. Better generative replay for continual federated learning. *arXiv preprint arXiv:2302.13001*, 2023.

[2] Keith J Holyoak and Derek Powell. Deontological coherence: A framework for commonsense moral reasoning. *Psychological Bulletin*, 142(11):1179, 2016.

[3] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

[4] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[5] Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11254–11263, 2019.

[6] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

[7] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–547, 2018.

[8] Timothée Lesort, Vincenzo Lomonaco, Andrei Stoian, Davide Maltoni, David Filliat, and Natalia Díaz-Rodríguez. Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. *Information fusion*, 58:52–68, 2020.

[9] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

[10] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

[11] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[12] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[13] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

[14] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

[15] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017.

[16] Paul Vanhaesebrouck, Aurélien Bellet, and Marc Tommasi. Decentralized collaborative learning of personalized models over networks. In *Artificial Intelligence and Statistics*, pages 509–517. PMLR, 2017.

[17] Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, 24(12):1028–1040, 2020.

[18] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *arXiv preprint arXiv:2302.00487*, 2023.

[19] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[20] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application*, 8:141–163, 2021.

[21] Melissa D McCradden, Shalmali Joshi, Mjaye Mazwi, and James A Anderson. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, 2(5):e221–e223, 2020.

[22] Michelle Seng Ah Lee and Luciano Floridi. Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. *Minds and Machines*, 31(1):165–191, 2021.

[23] Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. Facet: Fairness in computer vision evaluation benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20370–20382, 2023.

[24] Deborah Hellman. Measuring algorithmic fairness. *Virginia Law Review*, 106(4):811–866, 2020.

[25] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

[26] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[27] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

[28] Jenny Yang, Andrew A. S. Soltan, David W. Eyre, Yang Yang, and David A. Clifton. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *npj Digital Medicine*, 6(1):1–10, March 2023. Number: 1 Publisher: Nature Publishing Group.

[29] Octavian Suciu, Connor Nelson, Zhuoer Lyu, Tiffany Bao, and Tudor Dumitraș. Expected exploitability: Predicting the development of functional vulnerability exploits. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 377–394, 2022.

[30] Vincenzo Lomonaco et al. Cvpr 2020 continual learning in computer vision competition: Approaches, results, current challenges and future directions. *Artificial Intelligence*, 2022.

[31] Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10143–10153, 2022.

[32] Yuchen Zeng, Hongxu Chen, and Kangwook Lee. Improving fairness via federated learning. *arXiv preprint arXiv:2110.15545*, 2021.

[33] Cong Su, Guoxian Yu, Jun Wang, Hui Li, Qingzhong Li, and Han Yu. Multi-dimensional fair federated learning. *arXiv preprint arXiv:2312.05551*, 2023.

[34] Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A Salman Avestimehr. Fairfed: Enabling group fairness in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7494–7502, 2023.

[35] Zeou Hu, Kiarash Shaloudegi, Guojun Zhang, and Yaoliang Yu. Federated learning meets multi-objective optimization. *IEEE Transactions on Network Science and Engineering*, 9(4):2039–2051, 2022.

[36] Borja Rodríguez Gálvez, Filip Granqvist, Rogier van Dalen, and Matt Seigel. Enforcing fairness in private federated learning via the modified method of differential multipliers. In *NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021.

[37] Daniel Yue Zhang, Ziyi Kou, and Dong Wang. Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1051–1060. IEEE, 2020.

[38] Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 181–189. SIAM, 2021.

[39] Annie Abay, Yi Zhou, Nathalie Baracaldo, Shashank Rajamoni, Ebube Chuba, and Heiko Ludwig. Mitigating bias in federated learning. *arXiv preprint arXiv:2012.02447*, 2020.

[40] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. *arXiv preprint arXiv:2012.01696*, 2020.

[41] Taki Hasan Rafi, Faiza Anan Noor, Tahmid Hussain, and Dong-Kyu Chae. Fairness and privacy preserving in federated learning: A survey. *Information Fusion*, 105:102198, 2024.

[42] Yuxin Shi, Han Yu, and Cyril Leung. Towards fairness-aware federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[43] Mustafa Safa Ozdayi and Murat Kantarcioglu. The impact of data distribution on fairness and robustness in federated learning. In *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pages 191–196. IEEE, 2021.

[44] Lingjuan Lyu, Xinyi Xu, Qian Wang, and Han Yu. Collaborative fairness in federated learning. *Federated Learning: Privacy and Incentive*, pages 189–204, 2020.

[45] Jaehong Yoon et al. Scalable and order-robust continual learning with additive parameter decomposition. *arXiv preprint arXiv:1902.09432*, 2019.

[46] Pietro Buzzega et al. Dark experience for general continual learning: A strong, simple baseline. *Advances in Neural Information Processing Systems*, 2020.

[47] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13208–13217, 2020.

[48] Thanh-Dat Truong, Hoang-Quan Nguyen, Bhiksha Raj, and Khoa Luu. Fairness continual learning approach to semantic scene understanding in open-world environments. *Advances in Neural Information Processing Systems*, 36, 2024.

[49] Nikhil Churamani, Ozgur Kara, and Hatice Gunes. Domain-incremental continual learning for mitigating bias in facial expression and action unit recognition. *IEEE Transactions on Affective Computing*, 2022.

[50] Jeremy Bentham. *The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation*. Clarendon Press, 1996.

[51] Carlos Mougan and Joshua Brand. Kantian deontology meets ai alignment: Towards morally robust fairness metrics. *arXiv preprint arXiv:2311.05227*, 2023.

[52] Immanuel Kant and Jerome B Schneewind. *Groundwork for the Metaphysics of Morals*. Cambridge University Press, 1785.

[53] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[54] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2022.

[55] Lei Fu, Huanle Zhang, Ge Gao, Mi Zhang, and Xin Liu. Client selection in federated learning: Principles, challenges, and opportunities. *IEEE Internet of Things Journal*, 2023.

[56] Ala Gouissem, Zina Chkirbene, and Ridha Hamila. A comprehensive survey on client selections in federated learning. *arXiv preprint arXiv:2311.06801*, 2023.

[57] Carl Smestad and Jingyue Li. A systematic literature review on client selection in federated learning. In *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, pages 2–11, 2023.

[58] Mazhar Ali, Ankit Kumar Singh, Ajit Kumar, Syed Saqib Ali, and Bong Jun Choi. Comparative analysis of data-driven algorithms for building energy planning via federated learning. *Energies*, 16(18):6517, 2023.

[59] Konstantinos D Stergiou, Konstantinos E Psannis, Vasileios Vitsas, and Yutaka Ishibashi. A federated learning approach for enhancing autonomous vehicles image recognition. In *2022 4th International Conference on Computer Communication and the Internet (ICCCI)*, pages 87–90. IEEE, 2022.

[60] ZhiHui Wang, DeQian Fu, and Jiawei Zhang. Logistics data sharing method based on federated learning. In *Fifth International Conference on Computer Information Science and Artificial Intelligence (CISAI 2022)*, volume 12566, pages 380–385. SPIE, 2023.

[61] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

[62] Q Li, Z Wen, and B He. Federated learning systems: Vision, hype and reality for data privacy and protection. arxiv 2019. *arXiv preprint arXiv:1907.09693*.

[63] Viraaji Mothukuri, Reza M Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, and Gautam Srivastava. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619–640, 2021.

[64] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.

[65] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[66] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.