Drowsiness Detection Using Federated Learning: Lessons Learnt from Dealing with Non-IID Data

Rustem Dautov
rustem.dautov@sintef.no
SINTEF Digital
Oslo, Norway

Erik Johannes Husom
erik.johannes.husom@sintef.no
SINTEF Digital
Oslo, Norway

ABSTRACT

The privacy of personal data is paramount in the realm of assisted living and digital healthcare. Federated Learning (FL), with its decentralised model training approach, has emerged as a compelling solution to reconcile the need for personalised models with the requirement to protect sensitive personal information. By allowing model training to occur locally on user devices without centralising raw data, FL is intended to strike a balance between personalisation and privacy. While the potential benefits of FL in assisted living and digital healthcare are substantial, practical implementation poses significant challenges. One of them is the non-Independently and Identically Distributed (non-IID) nature of personal data. Unlike centralised datasets, non-IID data exhibits inherent variability across different individuals, as well as their surrounding contexts. Unfortunately, many research approaches in this domain often overlook the nuances of non-IID data, potentially leading to models that lack robust generalisation across diverse healthcare scenarios. To highlight the importance of this challenge, in this paper, we report on our hands-on experience of building a FL system for drowsiness detection using non-IID data. We compare this federated setup with a traditional, centralised approach to model training by identifying and discussing the associated challenges from multiple perspectives, as well as possible solutions and recommendations for further research.

CCS CONCEPTS

• **Security and privacy** → *Privacy protections*; • **Computing methodologies** → **Neural networks**; **Distributed artificial intelligence**.

KEYWORDS

Federated Learning, Non-IID Data, Drowsiness Detection, Assisted Living, Convolutional Neural Networks, Flower, Tensorflow

ACM Reference Format:
Rustem Dautov and Erik Johannes Husom. 2024. Drowsiness Detection Using Federated Learning: Lessons Learnt from Dealing with Non-IID Data. In *The PErvasive Technologies Related to Assistive Environments (PETRA) conference (PETRA '24), June 26–28, 2024, Crete, Greece.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3652037.3652074

PETRA '24, June 26–28, 2024, Crete, Greece
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1760-4/24/06
https://doi.org/10.1145/3652037.3652074

1 INTRODUCTION

Assisted living and digital healthcare increasingly rely on Machine Learning (ML) models to deliver personalised and effective services. Federated Learning (FL) has emerged as a privacy-aligned approach, allowing model training across decentralised and privacy-sensitive data sources. Rather than centralising data, FL enables training on individual nodes (devices or servers), maintaining data locally while aggregating updates to enhance the global model's performance. This approach is particularly beneficial in applications prioritising data security and confidentiality, such as healthcare and finances.

However, the non-Independently and Identically Distributed (non-IID) nature of personal data commonly present in healthcare-related scenarios poses a significant challenge to implementing FL, potentially impacting model performance and generalisation across diverse individual data. Real-world implementation of FL faces substantial hurdles due to the non-IID nature of distributed datasets, stemming from variations in health conditions, personal habits, demographics, and medical histories. Existing research often falls short in addressing this challenge, with many studies neglecting the non-IID data challenge when reporting results. Despite the highlighted benefits, a critical examination of digital healthcare studies reveals a gap in acknowledging and mitigating the non-IID challenge, risking overestimation of model performance.

This paper advocates for a responsible approach in digital healthcare research involving FL, urging deeper consideration of the non-IID nature of personal data. Future studies should employ advanced aggregation strategies explicitly addressing the non-IID challenge. Additionally, establishing transparent reporting standards is crucial to accurately reflect the impact of dataset heterogeneity in evaluation metrics. Our conclusions are drawn from hands-on experience of building a FL system for training a drowsiness detection ML model with non-IID data, compared against the traditional, centralised setup. Throughout this paper, we will draw on examples from healthcare and assisted living, which are, on the one hand, rapidly developing thanks to the recent advances in AI and ML, but, on the other, hindered by the high sensitivity of personal data. However, the presented work's relevance extends to other application scenarios and business domains dealing with non-IID data, which can potentially benefit from adopting the FL technology.

The remainder of the paper is organised as follows. Section 2 describes the background and context of this research, emphasising the primary existing challenge. Section 3 presents an overview of existing related works. Section 4 details the two experimental evaluations we conducted, followed by a critical discussion of the results. Section 5 concludes the paper with a summary and some concluding remarks.

2 RESEARCH CONTEXT AND MOTIVATION

In this section, we brief the reader on the broader research context of this work, explaining the concepts of edge computing and AI at the edge, followed by an overview of FL and the challenges associated with non-IID data. The latter will be re-visited further down in the paper when discussing the experimental findings.

2.1 Edge AI and Federated Learning

Recent technological advances have paved the way for *ubiquitous connectivity* [34] and *pervasive computing* [14]. With the availability of network connections, devices and systems are enabled to be seamlessly connected to the Internet or other communication networks. At the same time, pervasive computing extends the concept of ubiquitous connectivity by focusing on the integration of computing capabilities into everyday objects and environments forming the so-called *computing continuum* [32]. It involves embedding intelligence into a wide range of personal devices and human-centred spaces, such as smartphones, 'wearables', household appliances, vehicles, buildings, etc. The goal is to create an environment where computing and information processing become seamlessly integrated into people's daily lives, without requiring explicit user intervention (e.g., assisted living).

The advances in networking and computing capabilities of field-deployed devices underpin another relevant concept – *edge computing*, which is a decentralised computing paradigm that brings data processing and computation closer to the data source, instead of relying solely on centralised cloud servers [6]. Data processing at the edge can range from simple data pre-processing operations to rather advanced ML-driven AI analytics. The latter, commonly known as *Edge AI*, refers to the deployment of AI algorithms and models directly on edge devices, such as smartphones, IoT devices, edge servers, and other similar computing nodes [38]. It brings AI capabilities and decision-making closer to the data source, minimising the need for data transmission to centralised cloud servers. This enables near-real-time inference, reduces latency, saves bandwidth, enhances privacy, and enables offline functionality even in the absence of the Internet connection. All these features are especially important to the healthcare domain where physiological data collected by wearable or portable medical devices are processed either directly on those devices or on a smartphone acting as a wireless gateway [11, 15]. Similarly, the data privacy and network bandwidth constraints are usually critical aspects in various image and video recognition scenarios involving indoor or in-vehicle cameras [7, 9].

A natural next step in the Edge AI development was not only to deploy pre-trained AI models and run local inference, but also to train models at the edge. While individual edge devices are still constrained in their computing capabilities to perform heavy-weight model training, the promising solution was to combine multiple devices into an aggregated pool of computing resources and then orchestrate the iterative model training process, while keeping data locally. This ML approach, known as FL, enables training models on decentralised data without the need to transfer raw data to a central server [16]. The central idea behind FL is to enable collaborative model training while keeping the data on the local devices or servers, thereby addressing privacy and data security concerns. Here is how FL typically works:

1. *Model Initialisation*: A global ML model is initialised on a central server.
2. *Local Training*: The local nodes perform model training on their respective datasets without sharing the data itself. The training process may involve multiple iterations of training and updating the model's parameters.
3. *Model Update Aggregation*: After local training, each node sends only the model updates (i.e., changes in model parameters) to the central server.
4. *Model Aggregation*: The central server merges these model updates into the global model. This aggregation process can be done in various ways, e.g., averaging the updates.
5. *Iteration*: Steps 2 to 4 are repeated for multiple rounds, improving the global model with each iteration.

In comparison to the traditional centralised way of training ML models, FL offers several advantages:

- *Privacy Preservation*: Since raw data remains on local devices, users' privacy is better protected. The central server only sees aggregated model updates, not the actual data.
- *Data Efficiency*: FL reduces the need to transfer large datasets to a central location, which can be beneficial in scenarios with limited bandwidth or high data acquisition costs.
- *Security*: It reduces the risk of data breaches since raw data is not sent to a central location, making it harder for attackers to access sensitive information.
- *Increase Performance*: It allows for ML to be performed on edge devices, which in certain scenarios can lead to faster inference and reduced latency.

FL has applications in various privacy-critical fields, including healthcare (for medical data analysis while protecting patient privacy), IoT (for smart devices with limited processing power and connectivity), and personalised recommendations (to improve user experiences without compromising user data) [12]. Some prominent FL frameworks actively developed and used by the community include Flower,[1] Tensorflow Federated,[2] and OpenFL.[3]

2.2 Challenge: Non-Independently and Identically Distributed Data

In FL, the training data is distributed across multiple decentralised devices or nodes, and each node updates its model locally. The challenge arises when the data distribution among these nodes is non-IID, meaning that the data on different nodes is not similar in terms of statistical properties. Below are some key challenges associated with non-IID data in FL:

1. *Heterogeneous Data Distribution*: Nodes in a FL system may have different types of data or data from distinct sources, leading to variations in data distributions. This heterogeneity can result from diverse user behaviours, device types, or environmental factors. For example, nodes representing hospitals in urban areas might have a higher prevalence of

[1] https://flower.dev/

[2] https://www.tensorflow.org/federated/

[3] https://github.com/securefederatedai/openfl/

chronic diseases, while nodes from rural clinics may primarily handle cases related to agricultural injuries. This heterogeneity makes it challenging to build a global predictive model that accommodates both urban and rural healthcare needs effectively.

(2) *Ineffective Global Model*: The goal of FL is to learn a global model that generalises well across all nodes. However, non-IID data may lead to the development of a global model that performs poorly on certain nodes, affecting overall model performance. For example, if one node specialises in cardiology and another in oncology, the global healthcare model may struggle to provide accurate predictions for cardiac conditions or cancer, compromising the overall effectiveness of the healthcare model.

(3) *Biased Model Updates*: Nodes with more data or data that is more representative of the overall distribution may dominate the training process, leading to biased model updates that favour certain local characteristics over others, thus making the model being more tailored to the data on specific nodes and performing poorly on others. For example, if one node predominantly deals with pediatric data and another node focuses on geriatric patients, the global healthcare model might be biased toward the medical conditions prevalent in the pediatric population, potentially neglecting crucial aspects of geriatric care.

(4) *Algorithmic Challenges*: Standard FL algorithms may not perform optimally with non-IID data. Customised algorithms or modifications may be required to address the challenges posed by the heterogeneity of data distributions. For example, in medical imaging tasks, if nodes have data from different types of imaging devices with varying resolutions and specialties (e.g., X-rays, MRIs), traditional FL algorithms may need adjustments to handle these variations effectively in the healthcare domain.

(5) *Communication Overhead*: With non-IID data, the models may need frequent updates to adapt to the diverse local data distributions. This can result in increased communication overhead among nodes, leading to longer training times and higher resource consumption. For example, hospitals in densely populated urban areas may generate more healthcare data than those in rural regions. Frequent exchange of large model updates from urban hospitals could strain network resources, leading to increased communication costs.

Taken together, these issues represent a significant challenge slowing down the adoption of FL in real-life applications dealing with non-IID personal data. This list of outlined challenges will be further used as a reference to summarise our own hands-on experience in Section 4.3.

3 RELATED WORK

Researchers are actively working on developing techniques to mitigate the impact of non-IID data in FL. These existing works can be grouped into two categories – namely, the approaches dealing with non-IID data in centralised ML (which can potentially be applied to FL as well), and the approaches focusing on developing advanced aggregation strategies which would efficiently incorporate specifics of each individual contributor. In addition to this, in this section we also provide a brief summary of related studies focusing on other metrics (apart form the model performance) used for comparing centralised and federated approaches.

3.1 Addressing Non-IID Data in Centralised ML

Addressing non-IID data in ML requires thoughtful strategies to ensure model generalisation across diverse datasets. Non-IID data arises when the distribution of the training data is not consistent across different samples or subsets. One effective approach involves careful data pre-processing, where techniques like stratified sampling are employed during the split into training and testing sets [19]. This helps maintain a consistent class distribution in each subset, preventing one from having a significantly different distribution than the others. Data augmentation is another relevant technique, which relies on introducing variations to create new samples, preserving the underlying data distribution [22]. Transfer learning leverages pre-trained models on large, diverse datasets [39]. Fine-tuning these models on the target non-IID data allows them to adapt to specific characteristics while benefiting from the knowledge gained from broader datasets. Ensemble methods involve training multiple models using different subsets of the non-IID data or employing different algorithms [44]. Combining their predictions through techniques like bagging or boosting can enhance overall model performance. Domain adaptation becomes crucial when the source and target domains have different distributions [17]. Methods in this category aim to align feature distributions between domains to improve model generalisation. Meta-learning involves training models on various tasks, enabling them to adapt quickly to new tasks [37]. This also proves beneficial when dealing with non-IID data distributions, allowing models to generalise across different distributions. Relevant approaches also include reweighting and dynamic learning rate adjustment during training based on the model's performance on specific subsets or classes.

Collectively, these strategies provide a comprehensive toolkit for addressing the challenges posed by non-IID data, allowing ML models to adapt and generalise effectively across diverse datasets. The choice of specific methods often depends on factors such as the nature of the data, available computational resources, and the objectives of the ML application. In FL, however, the challenge of non-IID data is taken to a whole new level due to the distributed and isolated nature of the datasets, making these existing techniques less effective or completely inapplicable.

3.2 Addressing Non-IID Data in FL using Aggregation Algorithms

FL addresses the challenge of non-IID data through various aggregation algorithms [26, 27], each designed to facilitate the combination of model updates from decentralised nodes while managing the impact of heterogeneous local data distributions. These aggregation methods ensure that the global model can learn effectively from the diverse datasets distributed across multiple nodes.

One fundamental aggregation algorithm in FL is *Federated Averaging* (FedAvg) [33]. It calculates the average of model updates received from participating nodes, treating each node's contribution equally. However, in the presence of non-IID data, FedAvg may encounter challenges when nodes have significantly different amounts or types of data, potentially leading to biased models. To this end, extensions and variations of FedAvg have been introduced. *Weighted Federated Averaging* (Weighted Average) [3], for instance, assigns weights to each node's contribution during aggregation, offering a more nuanced approach by giving higher influence to nodes with more representative or diverse data. This weighting mechanism aims to address the non-IID challenge by better aligning the contributions with the overall learning objective. Another approach, *FL with Momentum* (FedProx), introduces a proximal term to the optimisation objective [18]. This term penalises large changes in model parameters between consecutive rounds, stabilising the learning process and preventing large unexpected deviations. By doing so, FedProx helps mitigate the impact of nodes with non-IID data that might introduce noisy updates, contributing to more stable convergence. *FL with Adaptive Learning Rates* (FedAdapt) adjusts learning rates for each node based on its historical performance [40]. Nodes with more consistent and accurate updates receive higher learning rates, adapting to the varying data characteristics of different nodes. This adaptability allows nodes with non-IID data to contribute effectively without disrupting the learning process. *FL with Personalisation* (FedPer), introduces personalisation factors to the aggregation process, enabling nodes to have personalised contributions to the global model [2]. These personalisation factors are learned based on the local data distribution, allowing nodes with non-IID data to contribute unique information without being overshadowed by the majority. Additionally, a clustering approach involves grouping nodes based on the similarity of their data distributions. *Clustered FL* aggregates updates within each cluster separately before combining cluster-level updates to obtain the global model update [29]. This clustering strategy helps address non-IID challenges by creating subgroups of nodes with similar data characteristics, facilitating more effective aggregation within clusters.

All these aggregation algorithms exemplify the diverse strategies employed in FL to handle non-IID data, offering adaptability and customisation based on the specific characteristics of the decentralised learning scenario. The specifics of non-IID datasets and application scenarios, however, often pose unique challenges, which cannot be easily addressed by none of these algorithms. In particular, in this work we will report on our attempt of using the widely adopted Weighted Average algorithm.

3.3 Other Comparative Studies

While in this paper we will primarily focus on the FL challenges stemming from the the non-IID nature of the training data and evaluate the performance of the trained models, the comparison of FL against the traditional centralised approaches may span across multiple other dimensions. FL and centralised training setups have been extensively compared across various metrics beyond just the performance of the trained model. Privacy preservation, network overheads, latency, computational intensity, scalability, fault tolerance and other related metrics play crucial roles in evaluating the efficiency and feasibility of both approaches. Overall, while FL introduces challenges such as increased network overhead and latency, it offers compelling benefits in privacy preservation, scalability, fault tolerance, and other key dimensions of distributed model training. Evaluating these metrics comprehensively helps researchers and practitioners understand the trade-offs and potential benefits of adopting FL in various application domains.

FL excels in *preserving data privacy* by keeping user data decentralised and local to individual devices. This distributed model reduces the risk of data breaches and unauthorised access, enhancing user trust and compliance with privacy regulations, such as GDPR and CCPA. By aggregating model updates instead of raw data, FL enables collaborative learning across distributed devices without compromising sensitive information. Being one of the core benefits of FL, has been extensively explored and reviewed by the research community resulting in multiple survey papers [23, 36, 42].

FL also demonstrates *scalability* advantages by leveraging the computational resources of participating devices, enabling distributed model training at scale [4]. Unlike centralised approaches, FL can accommodate a large number of devices without overwhelming central servers or imposing significant communication overhead. This scalability makes FL well-suited for applications with diverse and dynamic user populations, such as mobile devices, IoT devices, and edge computing environments. This distributed nature of FL, however, leads to the increased *computational intensity*. FL distributes computation across local devices, potentially alleviating the burden on central servers. However, this decentralisation introduces challenges in coordinating model updates and aggregating gradients efficiently. As a result, the computational intensity of FL may vary depending on factors such as the number of participating devices, the complexity of the model, and the communication protocol used for synchronisation. Almanifi *et al.* [1] review the existing approaches dealing with this challenge.

FL's distributed nature also incurs higher *network overheads and latency* compared to centralised training due to the frequent exchange of model updates between devices and the central server. These communication costs can impact network bandwidth and overall system performance, especially in scenarios with limited network capacity or high latency connections. The increased network communication in FL can also lead to higher latency compared to centralised training approaches, primarily due to the reliance on communication between devices and the central server for model synchronisation. This latency can affect the responsiveness of FL systems, particularly in real-time applications or scenarios where timely model updates are critical. As reviewed in [20, 35], mitigating latency in FL often involves optimising communication protocols, minimising data transfer, and leveraging edge computing resources to perform local computations.

In addition to these core metrics, studies have also examined FL's performance in terms of *fault tolerance* and *resilience* [10, 31], *energy efficiency* [30, 41], *model convergence speed* [5, 13, 25], and *algorithmic fairness* [21, 43]. albeit beyond the scope of this paper, considering all these metrics provide additional insights into the practical feasibility and advantages of adopting FL for distributed ML tasks in real-world application scenarios.

4 EMPIRICAL EVALUATION

We now proceed with an empirical evaluation of a FL scenario dealing with non-IID data in the context of drowsiness detection. We first explain the scenario highlighting its relevance to the problem at hand, and then proceed with an explanation and comparison of the two types of experiments that we conducted. One of our main intentions in this described research work was to demonstrate how the introduction of non-IID training data commonly present in real-world application scenarios affects the performance of the trained ML model. To this end, our conscious design decision was to keep the two evaluated setups as similar as possible to ensure fair comparison.

4.1 Drowsiness Detection Using Convolutional Neural Networks

Drowsiness poses a significant challenge in various professions, where its impact can have severe consequences. Defined by feelings of sleepiness, fatigue, and reduced alertness, drowsiness compromises the ability to maintain focus, make quick decisions, and respond rapidly – all critical aspects of safe driving [28]. In driving, drowsiness is a crucial concern due to its direct correlation with an increased risk of accidents. Insufficient or poor-quality sleep, long working hours, night shifts, and monotonous driving conditions contribute to drowsiness among drivers. Certain professions are at a higher risk of drowsiness-related incidents, particularly those involving extended hours on the road. Long-haul truck drivers, delivery professionals, and emergency service providers working irregular hours are in the risk zone. The implications of drowsy driving extend beyond individual performance, emphasising the need for effective strategies and technology-driven solutions to protect drivers and the broader public on the road.

Mitigation strategies in professions involving driving include implementing clear policies on rest breaks, promoting sufficient sleep, and providing education on the risks of drowsy driving. Technology-based solutions, such as driver monitoring systems in vehicles, analyse facial expressions and movements to detect signs of drowsiness, issuing alerts to encourage breaks and prevent accidents. Facial expression recognition using ML plays a crucial role in addressing drowsiness detection, particularly in the context of driver monitoring systems. By leveraging computer vision algorithms and deep learning models, facial features and expressions indicative of drowsiness can be analysed in real-time. ML models are trained to recognise specific visual cues such as drooping eyelids, changes in facial expressions, and other signs associated with drowsiness. These algorithms can process live video feeds from in-vehicle cameras, continuously monitoring the driver's face for subtle changes. Upon detection of potential drowsiness, these systems can trigger timely alerts or interventions, such as audio warnings or suggestions for the driver to take a break. The non-intrusive nature of facial expression recognition makes it an effective and scalable solution for drowsiness detection, enhancing road safety by addressing this critical challenge in a proactive manner [8].

To this end, using Convolutional Neural Networks (CNNs) for driver drowsiness detection is a powerful application of deep learning. CNNs excel in image recognition tasks by automatically learning hierarchical features from visual data. Among other tools, TensorFlow[4] in combination with Keras[5] provide a robust framework for implementing such systems. In this context, facial images captured by in-vehicle cameras serve as input. TensorFlow, an open-source ML library, seamlessly integrates with Keras, a high-level neural networks API, simplifying the implementation of CNN architectures. By structuring layers with convolutional and pooling operations, a CNN can effectively extract intricate patterns and features from facial expressions, eye movements, and other indicators of drowsiness. Training the model on labelled datasets allows it to learn and generalise patterns associated with drowsiness. Once trained, the CNN can be deployed in real-time driver monitoring systems, providing accurate and rapid detection of drowsiness, thereby contributing to enhanced road safety.

4.2 Experiments: Centralised vs Federated

In our empirical evaluation, we will use the driver drowsiness dataset[6] of 41,790 RGB images containing extracted and cropped faces of drivers. The dataset contains 28 distinct subjects represented by non-IID chunks of varying size (smallest dataset: 415 images, largest dataset: 2892 images). All images are split between two labelled classes, namely *Drowsy* and *Non Drowsy*. The dataset was originally prepared and used in the context of the research work reported in [24]. In addition to the sensitive nature of the images, one specific reason for choosing this drowsiness detection dataset for the experiments is the ability to split it into 28 smaller chunks each belonging to a distinct individual. Each chunk essentially represents a non-IID subset, which is different in terms of its size, as well as the unique facial features of the represented subject. This way, it provides a solid foundation for model training in a federated setup using truly non-IID data, which can then be compared to the original centralised setup.

To make the comparison of the two approaches fair, we aimed to keep the federated setup as similar to the centralised on as possible. More specifically, both setups use exactly the same CNN, optimiser, data loading, pre-processing and augmentation parameters. The main difference comes from the Weighted Average aggregation strategy that we applied, as well as the number of training iterations. The main comparison metrics for the two approaches were accuracy, precision, and recall – well-established metrics for evaluating model performance in ML. Please note that the focus of these experiments was not to train a highly accurate classification model for drowsiness detection (in fact, as we describe below, the performance of the centralised model has further room for improvement). Rather, the main goal was to compare the two approaches and highlight the challenges associated with non-IID data in the federated setup. The source code for the experiments is available in a public repository.[7]

4.2.1 Centralised Training. In the centralised setup, the entire dataset of 41,790 images is accessible in one location, and the model is trained on this comprehensive dataset. The CNN iteratively learns to recognise patterns associated with drowsiness by adjusting its parameters based on the entire dataset. This centralised approach is straightforward, as it involves a single training process on the

[4] https://www.tensorflow.org/

[5] https://keras.io/

[6] https://www.kaggle.com/datasets/ismailnasri20/driver-drowsiness-dataset-ddd/data

[7] https://github.com/SINTEF-9012/fl-ddd

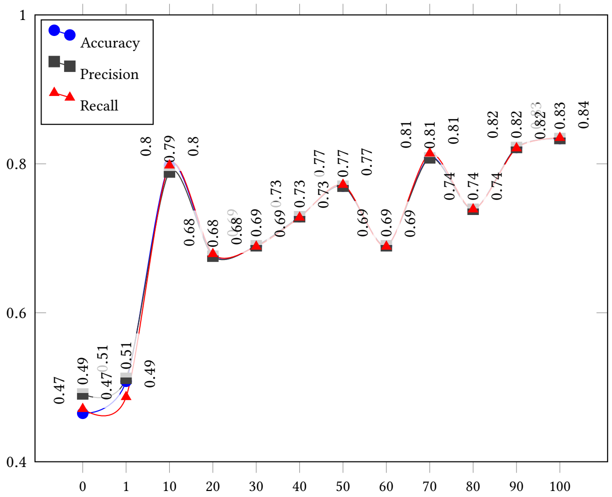

Figure 1: Performance of the centralised model over a dataset of 41,793 images (after 100 epochs).

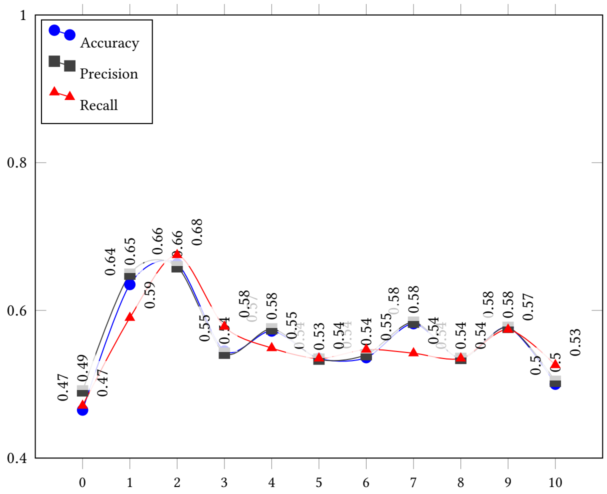

Figure 2: Performance of the trained model in a federated setup (after 10 training rounds of 10 epochs each).

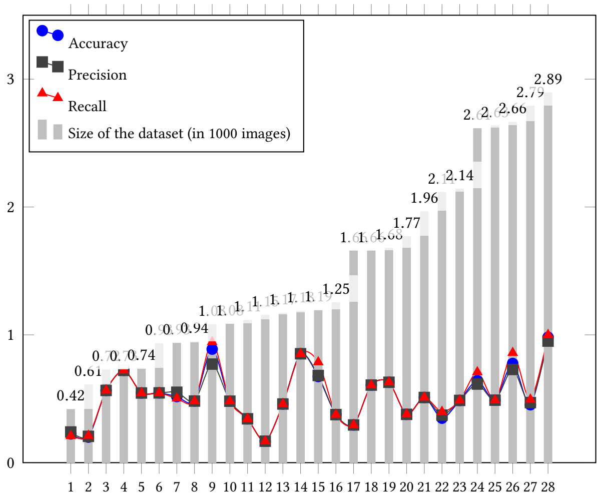

Figure 3: Performance of the aggregated model on individual datasets and its potential correlation with the dataset sizes.

complete dataset. Fig. 1 depicts how the performance of the resulting model changed over 100 epochs. For clarity, the diagram only includes the performance of the base model before the training started (epoch 0), followed by 10 more training checkpoints corresponding to 10, 20, 30 epochs, and so on.

4.2.2 Federated Learning. The centralised setup raises concerns related to data privacy, since the images contain identifiable personal information. Additionally, the centralised model may not generalise well to diverse driving conditions or individual differences in facial expressions, as it is trained on a uniform dataset. In the FL setup where the entire dataset is split into 28 separate parts, each representing a specific subject, the focus extends beyond individual model training to the aggregation of these decentralised model updates. After each local model is trained on its respective subset, the model updates are shared and aggregated to create a global model that captures the collective knowledge learned from all subjects. As the underlying FL framework, we used Flower, which we coupled with the centralised Tensorflow/Keras implementation. To align with the centralised setup, we aimed for 10 training rounds of 10 epochs each, thus resulting in 100 epochs in total.[8]

The non-IID nature of the distributed datasets suggested the Weighted Average algorithm as an effective aggregation strategy. After local training, the model updates are weighted based on the size of their corresponding training datasets and averaged to produce a global update. These global updates are then applied to the initial global model. In theory, the global model was expected to adapt to the diverse facial features specific to each subject and the Weighted Average mechanism would ensure that contributions from subjects with larger or more representative datasets received appropriate emphasis during the aggregation process. The performance of the resulting aggregated model after each round is depicted in Fig. 2 (round 0 corresponds to the very initial model, which has not been trained on the target data yet).

4.3 Comparing the Two Approaches: Lessons Learnt

The main observation drawn from this experimental evaluation was the fact that converting a centralised ML training setup into a federated one by using default, off-the-shelf techniques is not a trivial task given the non-IID nature of the training data. To be more specific, we now revisit the five challenges associated with non-IID data in FL outlined in Section 2.2, and summarise our experience based on the conducted experimental evaluation.

[8] In FL, training rounds do not align precisely with epochs in centralised training but share a similar iterative concept. A training round encompasses the collaborative learning process across local nodes, where each node contributes to the global model. This iterative collaboration over multiple rounds serves a role similar to epochs in centralised training, albeit with the distinction that FL involves decentralised updates from individual nodes.

4.3.1 Heterogeneous Data Distribution. This is essentially the very nature of non-IID data. The heterogeneity of the datasets used in our experiments emerged as the main factor influencing the performance disparities observed between centralised and federated models. The datasets varied significantly in size, ranging from small to large, reflecting the diverse nature of data available across decentralised nodes. Notably, all datasets comprised facial features from distinct individuals, introducing a level of complexity in capturing diverse facial expressions and characteristics. This diversity, while enriching the dataset, also underscored the challenges associated with achieving uniform model performance across federated nodes.

4.3.2 Ineffective Global Model. In our comparative study of centralised and FL, we observed notable differences in model performance. The centrally trained model consistently outperformed its federated counterpart in terms of accuracy, precision and recall. Despite the promise of FL in preserving data privacy and enabling decentralised training, our results indicate that the federated model's performance was not on par with the centralised model. More specifically, as indicated by Figg. 1 and 2, the accuracy, precision and recall of the centralised model reached approximately 83%, whereas the federated model's best performance was about 66% and remained around 50% on average.

4.3.3 Biased Model Updates. Furthermore, our experiments also revealed that the FL model exhibited even worse performance not only in comparison to the centralised model but also when assessed on the datasets of individual local nodes. The federated approach, while inherently advantageous for privacy preservation, faced challenges in effectively leveraging the diversity of the local non-IID datasets. This is reflected in Fig. 3, which contains average values for accuracy, precision and recall for each of the 28 federated node. Performance on some of the nodes dropped to critical 16%. With the model performance being so low, it is hard to draw any conclusions about potential bias in the resulting model towards some of the nodes, albeit the significant performance variations (up to 80%) among the nodes clearly indicates that the model is not generalisable enough.

4.3.4 Algorithmic Challenges. The default logic of the Weighted Average algorithm based on the dataset size arguably failed to capture the specifics of individual nodes. The grey bars in Fig. 3 represent the sizes of individual non-IID datasets on each of the 28 nodes. Contrary to the expectations associated with the Weighted Average approach, our experiments demonstrated a lack of clear correlation between the size of individual datasets and the performance of the federated model on local nodes. This challenges the effectiveness of simple size-based weighting as a reliable metric for aggregation, suggesting the need for more nuanced and adaptive strategies in FL to better accommodate the intricacies of diverse imagery datasets across decentralised nodes.

4.3.5 Communication Overhead. The implementation of experiments in the FL setup (even with just 10 rounds) resulted in a substantially extended training duration compared to the centralised approach. This prolonged training period was attributed to the iterative nature of FL, involving communication rounds between local nodes and the central server. The increased computational demand and extended training times translated to a higher consumption of computing resources. Furthermore, the reliance on communication between nodes placed an additional strain on network bandwidth, emphasising the resource-intensive nature of FL implementations.

5 CONCLUSION

As healthcare increasingly integrates technological advancements, striking a balance between personalisation and privacy becomes imperative. While the challenges of implementing FL in healthcare are acknowledged, future research endeavours must prioritise addressing the non-IID nature of personal data to unlock FL's full potential in personalised and privacy-preserving healthcare applications.

Our experiments demonstrated that neglecting the heterogeneity in personal data can lead to sub-optimal models incapable of capturing the nuanced features of individual profiles. As we argued, default solutions for converting a centralised setup into a federated one are not immediately applicable and require thoughtful adjustment to a specific use-case and training data. Consequently, research efforts should focus on developing FL systems that explicitly account for and mitigate the impact of non-IID data from the initial design phase, ensuring the generated models are accurate and generalisable across diverse application scenarios.

In light of these findings, we advocate for increased attention within the research community to the non-IID nature of personal datasets when employing FL. Future studies should explore more sophisticated aggregation strategies beyond the Weighted Average aggregation. Additionally, establishing transparent reporting standards is crucial to accurately reflect the impact of dataset heterogeneity in evaluation metrics.

ACKNOWLEDGMENTS

This work has received funding from: the European Union's Horizon Europe research and innovation programme under grant agreements No. 101135576 (INTEND), No. 101095634 (ENTRUST), and No. 101120657 (ENFIELD), and the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101020416 (ERATOSTHENES).

REFERENCES

[1] Omair Rashed Abdulwareth Almanifi, Chee-Onn Chow, Mau-Luen Tham, Joon Huang Chuah, and Jeevan Kanesan. 2023. Communication and computation efficiency in federated learning: A survey. *Internet of Things* 22 (2023), 100742. https://doi.org/10.1016/j.iot.2023.100742

[2] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. 2019. Federated learning with personalization layers. (2019). https://doi.org/10.48550/arXiv.1912.00818

[3] Martin Beaussart, Felix Grimberg, Mary-Anne Hartley, and Martin Jaggi. 2021. Waffle: Weighted averaging for personalized federated learning. (2021). https://doi.org/10.48550/arXiv.2110.06978

[4] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, Brendan McMahan, et al. 2019. Towards federated learning at scale: System design. *Proceedings of machine learning and systems* 1 (2019), 374–388. https://doi.org/10.48550/arXiv.1902.01046

[5] Mingzhe Chen, H Vincent Poor, Walid Saad, and Shuguang Cui. 2020. Convergence time optimization for federated learning over wireless networks. *IEEE Transactions on Wireless Communications* 20, 4 (2020), 2457–2471. https://doi.org/10.1109/TWC.2020.3042530

[6] Rustem Dautov, Salvatore Distefano, Dario Bruneo, Francesco Longo, Giovani Merlino, and Antonio Puliafito. 2017. Pushing intelligence to the edge with a stream processing architecture. In *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart*

Data (SmartData). IEEE, 792–799. https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2017.121

[7] Rustem Dautov, Salvatore Distefano, Dario Bruneo, Francesco Longo, Giovanni Merlino, Antonio Puliafito, and Rajkumar Buyya. 2018. Metropolitan intelligent surveillance systems for urban areas by harnessing IoT and edge computing paradigms. *Software: Practice and experience* 48, 8 (2018), 1475–1492. https://doi.org/10.1002/spe.2586

[8] Rustem Dautov, Salvatore Distefano, and Rajkumaar Buyya. 2019. Hierarchical data fusion for smart healthcare. *Journal of Big Data* 6, 1 (2019), 1–23. https://doi.org/10.1186/s40537-019-0183-6

[9] Rustem Dautov, Salvatore Distefano, Giovani Merlino, Dario Bruneo, Francesco Longo, and Antonio Puliafito. 2017. Towards a Global Intelligent Surveillance System. In *Proceedings of the 11th International Conference on Distributed Smart Cameras*. ACM New York, NY, USA, 119–124. https://doi.org/10.1145/3131885.3131918

[10] Rustem Dautov and Erik Johannes Husom. 2024. Raft Protocol for Fault Tolerance and Self-Recovery in Federated Learning. In *Proceedings of the 19th IEEE/ACM Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS 2024)*. IEEE. https://doi.org/10.1145/3643915.3644093

[11] Rustem Dautov, Erik Johannes Husom, Fotis Gonidis, Spyridon Papatzelos, and Nikolaos Malamas. 2022. Bridging the Gap Between Java and Python in Mobile Software Development to Enable MLOps. In *2022 18th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 363–368. https://doi.org/10.1109/WiMob55322.2022.9941679

[12] Rustem Dautov, Erik Johannes Husom, Sagar Sen, and Hui Song. 2023. Towards Community-Driven Generative AI. In *Position Papers of the 18th Conference on Computer Science and Intelligence Systems, Annals of Computer Science and Information Systems*, Vol. 36. 43–50. https://doi.org/10.15439/2023F5494

[13] Canh T Dinh, Nguyen H Tran, Minh NH Nguyen, Choong Seon Hong, Wei Bao, Albert Y Zomaya, and Vincent Gramoli. 2020. Federated learning over wireless networks: Convergence analysis and resource allocation. *IEEE/ACM Transactions on Networking* 29, 1 (2020), 398–409. https://doi.org/10.1109/TNET.2020.3035770

[14] Maria R Ebling. 2016. Pervasive Computing and the Internet of Things. *IEEE Pervasive Computing* 15, 1 (2016), 2–4. https://doi.org/10.1109/MPRV.2016.7

[15] Erik Johannes Husom, Rustem Dautov, Adela Nedisan Videsjorden, Fotis Gonidis, Spyridon Papatzelos, and Nikolaos Malamas. 2022. Machine Learning for Fatigue Detection using Fitbit Fitness Trackers. In *Proceedings of the 10th International Conference on Sport Sciences Research and Technology Support - icSPORTS*. INSTICC, SciTePress, 41–52. https://doi.org/10.5220/0011527500003321

[16] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated learning: Strategies for improving communication efficiency. In *29th Conference on Neural Information Processing Systems (NIPS2016)*. 1–5. https://doi.org/10.48550/arXiv.1610.05492

[17] Wouter M Kouw and Marco Loog. 2018. *An introduction to domain adaptation and transfer learning*. Technical Report. Delft University of Technology. https://doi.org/10.48550/arXiv.1812.11806

[18] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine* 37, 3 (2020), 50–60. https://doi.org/10.1109/MSP.2020.2975749

[19] Edo Liberty, Kevin Lang, and Konstantin Shmakov. 2016. Stratified sampling meets machine learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16)*. JMLR.org, 2320–2329. https://doi.org/10.5555/3045390.3045635

[20] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. 2020. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials* 22, 3 (2020), 2031–2063. https://doi.org/10.1109/COMST.2020.2986024

[21] Lingjuan Lyu, Xinyi Xu, Qian Wang, and Han Yu. 2020. Collaborative fairness in federated learning. *Federated Learning: Privacy and Incentive* (2020), 189–204. https://doi.org/10.1007/978-3-030-63076-8_14

[22] Agnieszka Mikołajczyk and Michał Grochowski. 2018. Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)*. IEEE, 117–122. https://doi.org/10.1109/IIPHDW.2018.8388338

[23] Viraaji Mothukuri, Reza M Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, and Gautam Srivastava. 2021. A survey on security and privacy of federated learning. *Future Generation Computer Systems* 115 (2021), 619–640. https://doi.org/10.1016/j.future.2020.10.007

[24] Ismail Nasri, Mohammed Karrouchi, Hajar Snoussi, Kamal Kassmi, and Abdelhafid Messaoudi. 2022. Detection and prediction of driver drowsiness for the prevention of road accidents using deep neural networks techniques. In *WITS 2020: Proceedings of the 6th International Conference on Wireless Technologies, Embedded, and Intelligent Systems*. Springer, 57–64. https://doi.org/10.1007/978-981-33-6893-4_6

[25] Hung T Nguyen, Vikash Sehwag, Seyyedali Hosseinalipour, Christopher G Brinton, Mung Chiang, and H Vincent Poor. 2020. Fast-convergent federated learning. *IEEE Journal on Selected Areas in Communications* 39, 1 (2020), 201–218. https://doi.org/10.1109/JSAC.2020.3036952

[26] Adrian Nilsson, Simon Smith, Gregor Ulm, Emil Gustavsson, and Mats Jirstrand. 2018. A performance evaluation of federated learning algorithms. In *Proceedings of the second workshop on distributed infrastructures for deep learning*. ACM, 1–8. https://doi.org/10.1145/3286490.3286559

[27] Pian Qi, Diletta Chiaro, Antonella Guzzo, Michele Ianni, Giancarlo Fortino, and Francesco Piccialli. 2024. Model aggregation techniques in federated learning: A comprehensive survey. *Future Generation Computer Systems* 150 (2024), 272–293. https://doi.org/10.1016/j.future.2023.09.008

[28] Muhammad Ramzan, Hikmat Ullah Khan, Shahid Mahmood Awan, Amina Ismail, Mahwish Ilyas, and Ahsan Mahmood. 2019. A survey on state-of-the-art drowsiness detection techniques. *IEEE Access* 7 (2019), 61904–61919. https://doi.org/10.1109/ACCESS.2019.2914373

[29] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. 2020. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems* 32, 8 (2020), 3710–3722. https://doi.org/10.1109/TNNLS.2020.3015958

[30] Dian Shi, Liang Li, Rui Chen, Pavana Prakash, Miao Pan, and Yuguang Fang. 2022. Toward energy-efficient federated learning over 5g+ mobile devices. *IEEE Wireless Communications* 29, 5 (2022), 44–51. https://doi.org/10.1109/MWC.003.2100028

[31] Jinhyun So, Başak Güler, and A Salman Avestimehr. 2020. Byzantine-resilient secure federated learning. *IEEE Journal on Selected Areas in Communications* 39, 7 (2020), 2168–2181. https://doi.org/10.1109/JSAC.2020.3041404

[32] Hui Song, Rustem Dautov, Nicolas Ferry, Arnor Solberg, and Franck Fleurey. 2022. Model-based fleet deployment in the IoT–edge–cloud continuum. *Software and Systems Modeling* 21, 5 (2022), 1931–1956. https://doi.org/10.1007/s10270-022-01006-z

[33] Tao Sun, Dongsheng Li, and Bao Wang. 2022. Decentralized federated averaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4289–4301. https://doi.org/10.1109/TPAMI.2022.3196503

[34] Vamsi Talla, Mehrdad Hessar, Bryce Kellogg, Ali Najafi, Joshua R Smith, and Shyamnath Gollakota. 2017. LoRa Backscatter: Enabling The Vision of Ubiquitous Connectivity. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1, 3 (2017), 1–24. https://doi.org/10.1145/3130970

[35] Nguyen H Tran, Wei Bao, Albert Zomaya, Minh NH Nguyen, and Choong Seon Hong. 2019. Federated learning over wireless networks: Optimization model design and analysis. In *IEEE INFOCOM 2019-IEEE conference on computer communications*. IEEE, 1387–1395. https://doi.org/10.1109/INFOCOM.2019.8737464

[36] Nguyen Truong, Kai Sun, Siyao Wang, Florian Guitton, and YiKe Guo. 2021. Privacy preservation in federated learning: An insightful survey from the GDPR perspective. *Computers & Security* 110 (2021), 102402. https://doi.org/10.1016/j.cose.2021.102402

[37] Joaquin Vanschoren. 2019. Meta-learning. *Automated machine learning: methods, systems, challenges* (2019), 35–61. https://doi.org/10.1007/978-3-030-05318-5_2

[38] Xiaofei Wang, Yiwen Han, Victor CM Leung, Dusit Niyato, Xueqiang Yan, and Xu Chen. 2020. *Edge AI: Convergence of edge computing and artificial intelligence*. Springer. https://doi.org/10.1007/978-981-15-6186-3

[39] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data* 3, 1 (2016), 1–40. https://doi.org/10.1186/s40537-016-0043-6

[40] Di Wu, Rehmat Ullah, Paul Harvey, Peter Kilpatrick, Ivor Spence, and Blesson Varghese. 2022. Fedadapt: Adaptive offloading for iot devices in federated learning. *IEEE Internet of Things Journal* 9, 21 (2022), 20889–20901. https://doi.org/10.1109/JIOT.2022.3176469

[41] Zhaohui Yang, Mingzhe Chen, Walid Saad, Choong Seon Hong, and Mohammad Shikh-Bahaei. 2020. Energy efficient federated learning over wireless communication networks. *IEEE Transactions on Wireless Communications* 20, 3 (2020), 1935–1949. https://doi.org/10.1109/TWC.2020.3037554

[42] Xuefei Yin, Yanming Zhu, and Jiankun Hu. 2021. A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–36. https://doi.org/10.1145/3460427

[43] Zirui Zhou, Lingyang Chu, Changxin Liu, Lanjun Wang, Jian Pei, and Yong Zhang. 2021. Towards fair federated learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. ACM, 4100–4101. https://doi.org/10.1145/3447548.3470814

[44] Zhi-Hua Zhou. 2012. *Ensemble Methods: Foundations and Algorithms* (1st ed.). Chapman & Hall/CRC. https://doi.org/10.1201/b12207