Developing a scalable and privacy-preserving deep learning model for the classification of peripheral blood cell images

Universitat Oberta de Catalunya

LIBERTAS PERFVNDET · OMNIA LVCE

UNIVERSITAT DE BARCELONA

Albert Garcia Llagostera

MU Bioinf. i Bioest.
Àrea de treball final

Advisor
Edwin Santiago Alférez Baquero

Professor/a responsable de l'assignatura
Carles Ventura Royo

18/06/2023

FITXA DEL TREBALL FINAL

Títol del treball:	*Developing a scalable and privacy-preserving deep learning model for the classification of peripheral blood cell images.*
Nom de l'autor:	*Albert Garcia Llagostera*
Nom del director/a:	*Edwin Santiago Alférez Baquero*
Nom del PRA:	*Carles Ventura Royo*
Data de lliurament (mm/aaaa):	*06/2024*
Titulació o programa:	*Màster en Bioinformática I Bioestadística*
Àrea del Treball Final:	*Bioinformàtica Estadística I Aprenentatge Automàtic*
Idioma del treball:	*anglès*
Paraules clau	*Histopathology, Federated Learning, Privacy*

Abstract

Histopathological diagnosis is a time-intensive process dependent on the expertise and interpretative criteria of pathologists. Digital pathology, employing machine learning models, offers a promising avenue to enhance diagnostic accuracy and efficiency through computer-aided diagnosis systems. Specifically, the automated counting and identification of cells from blood smears constitute 80% of the initial analyses required for detecting haematological diseases.

The intrinsic sensitivity of medical data demands robust privacy safeguards. This has focused recent investigations into the potential of collaborative learning, or Federated Learning (FL), as a scalable and inherently private training paradigm. By training data locally and subsequently aggregating parameters on a central server, the direct movement and sharing of medical data are circumvented. Nevertheless, recent studies have cast doubt on the privacy of these collaborative trainings. Moreover, collaborative learning faces the challenge of dealing with the heterogeneity of participating clients to generate an efficient model across various nodes.

This work presents a performance comparison between different training types for peripheral blood cell classification models. The findings suggest that collaborative learning, both in homogeneous (IID) and heterogeneous (non-IID) clients, could enhance the predictive capability of conventionally trained classification models. Furthermore, collaborative learning has the potential to reduce the time and resources required for model training.

Index

1. Introduction

1.1. Work's Context and Justification

This work addresses the challenge of automatic hematologic cell type recognition in peripheral blood smears using deep learning (DL) techniques, considering the prism of security for highly sensitive data (medical data).

Histopathological diagnosis is time-consuming and relies on pathologists' expertise and interpretation criteria (Fedeli et al., 2023). Digital pathology (DP) presents an opportunity to leverage computer-aided diagnosis (CAD) systems for improved diagnostic accuracy.

The inherent sensitivity of medical data needs of stringent privacy protection. Data breaches during transfer (between hospitals, patient servers, cloud storage) or during DL model training/execution can expose highly sensitive information. In the wrong hands, such data could be misused by insurance companies, employers, pharmaceutical companies, or marketing firms, with potentially detrimental consequences. The General Data Protection Regulation (GDPR) (Regulation 2016/679, Article 9) classifies health data as "special data," demanding the implementation of robust technical and organizational safeguards against unauthorized access, disclosure, alteration, or destruction.

Federated Learning

Initially described by (Brendan McMahan et al, 2017) Federated learning (FL) has emerged as a promising approach for training machine learning models while preserving data privacy (Figure 1). Briefly, FL trains a single model in a distributed manner, utilizing data from various clients (hospitals) without ever pooling the raw data itself. After every client separately trains with its own data, each client shares their fine-tunned model parameters with global server. Global server receives fine-tunned parameters and performs and aggregates them, creating new ones that are capable of scoring good scores in every client separately. These new unique parameters are shared with every node, updating local model's weights. This allows for model validation across different datasets, even if they are similar or dissimilar. Two primary variants of FL exist: Vertical Federated Learning (VFL) and Horizontal Federated Learning (HFL).

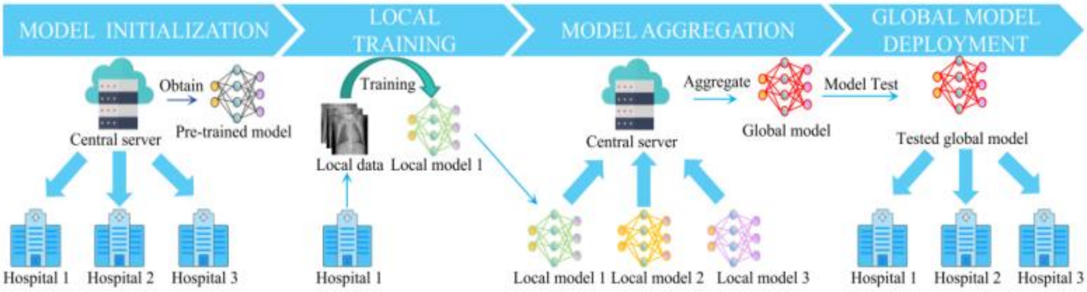


Figure 1. Basic scheme of Federated Learning. Nodes/client train with their data locally, while a central server coordinates the training cooperation among nodes/clients. (Hu & Chaddad, 2023)

HFL is particularly relevant for this work. In this approach, clients collaborate using the same set of features extracted from individual patient data. This facilitates the training of a shared model while ensuring individual data privacy. A robust HFL model would enable hospitals to jointly train a model using their combined data, effectively leveraging shared knowledge without compromising patient confidentiality.

FedAvg

There exists a variety of algorithms for the deployment of Federated Learning, encompassing a diverse range of architectural frameworks and methodologies for the aggregation of parameters. The focus of this investigation will be on the *FedAvg* (Federated Averaging) algorithm (Algorithm 1). Despite its foundational simplicity, it has consistently exhibited remarkable adaptability and efficacy. The operational mechanics of *FedAvg* are delineated as follows:

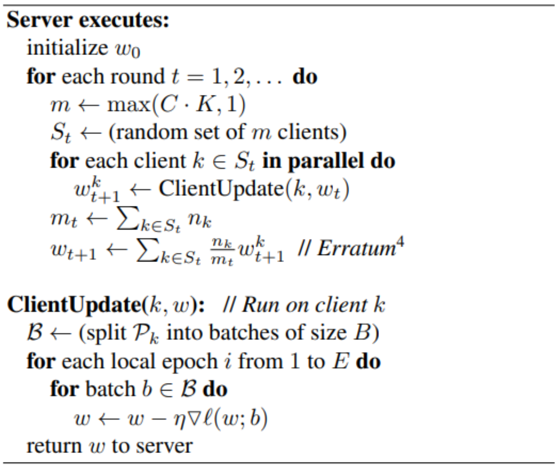
1. Initialisation: A prototypical global model undergoes initialisation within the confines of the central server.

2. Distribution of Updates: After initialisation, there is a dissemination of the global model’s weight adjustments across the client network.

3. Local Update: In an autonomous manner, each client enhances the model utilising their exclusive dataset, thereafter, transmitting the refined model to the central nexus.

4. Averaging: The central server executes a synthesis of the client-submitted enhancements, culminating in the refinement of the global model.

5. Repetition: The process is repeated several times until the desired accuracy objective is achieved.



```
Server executes:
  initialize $w_0$
  for each round $t = 1, 2, \dots$ do
    $m \leftarrow \max(C \cdot K, 1)$
    $S_t \leftarrow$ (random set of $m$ clients)
    for each client $k \in S_t$ in parallel do
      $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
    $m_t \leftarrow \sum_{k \in S_t} n_k$
    $w_{t+1} \leftarrow \sum_{k \in S_t} \frac{n_k}{m_t} w_{t+1}^k$   // Erratum[4]

ClientUpdate($k, w$):   // Run on client $k$
  $\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size $B$)
  for each local epoch $i$ from 1 to $E$ do
    for batch $b \in \mathcal{B}$ do
      $w \leftarrow w - \eta \nabla \ell(w; b)$
  return $w$ to server
```

Algorithm 1 FederatedAveraging. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate. (Brendan McMahan et al, 2017)

Federated Learning as privacy-preserving tool

FL has shown promise in the development of privacy-preserving medical applications (Andreux et al., 2020; Hosseini et al., 2023; Shen et al., 2023). Within the field of histopathology, several studies have explored the potential of FL as a privacy-aware tool for healthcare professionals, achieving promising results (details will be provided in the State of the Art section).

Identically Independent Distributed (IID) Data and non-IID Data

In the context of Federated Learning (FL), IID (Independent and Identically Distributed) and non-IID are crucial characteristics that affect the performance and robustness of the learning process. In summary, IID data assumes independence and identical distribution of data among clients, while non-IID data relaxes these assumptions and introduces more complex data structures and dependencies. In real-case scenarios, one node/client tends to not be a good representation of all the population. This is why is important to consider the study of an FL model using IID and non-IID data.

Peripheral Blood Cells

The blood periphery contains several types of cells that play crucial roles in various physiological processes. Here's an overview (Figure 2) of the different types of cells found in the blood periphery and their importance in the context of a blood smear:

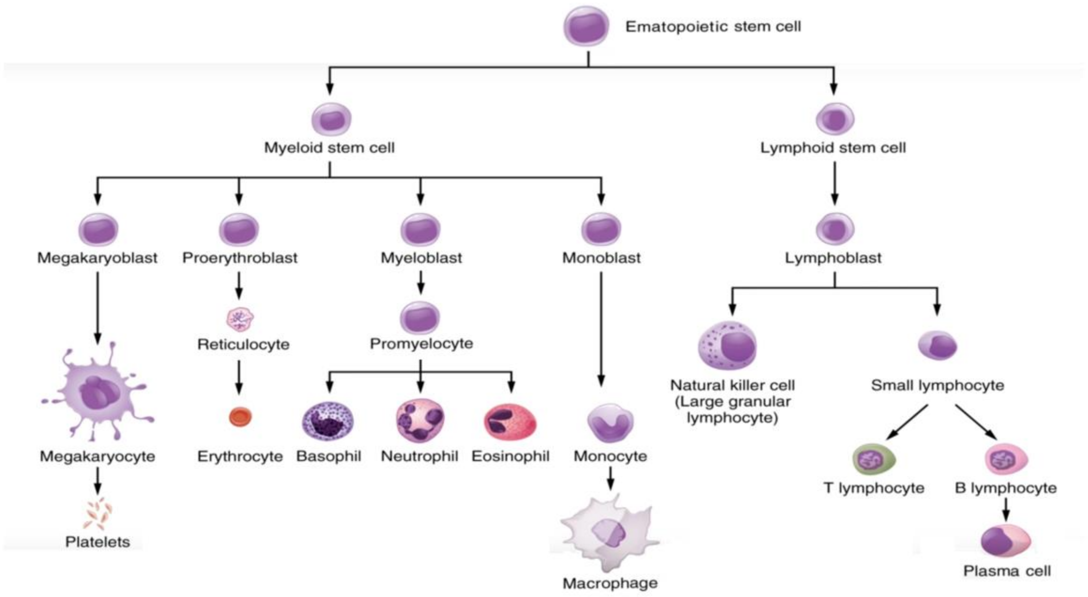


Figure 2. Ematopoietic stem cell lineages. ProfessorDaveExplains 2021

Erythrocytes (Red Blood Cells)

These cells are responsible for carrying oxygen from the lungs to the body's tissues. They are typically biconcave and have a diameter of approximately 7 micrometers. Erythrocytes are not being considered for model training.

Leukocytes (White Blood Cells)

Granulocytes
Neutrophils: These cells are the most abundant type of leukocyte in the blood and are responsible for fighting bacterial infections. They can be segmented or band shaped.
Eosinophils: These cells are involved in the immune response and play a role in fighting parasitic infections.
Basophils: These cells are involved in allergic reactions and play a role in the inflammatory response.

Agranulocytes
Monocytes: These cells mature into macrophages, which are responsible for phagocytosing foreign particles and cellular debris.

Lymphocytes: These cells are involved in the immune response and can be further classified into B cells and T cells.

Platelets (Thrombocytes)

Platelets: These cells are small, irregularly shaped fragments of megakaryocytes and are essential for blood clotting.

Counting cells in a blood smear is crucial for diagnosing and monitoring treatment of various blood disorders and diseases. Some examples:
Infection Diagnosis: Counting leukocytes and their morphology can help diagnose infections, such as bacterial or parasitic infections.
Blood Clotting Disorders: Counting platelets can help diagnose blood clotting disorders, such as thrombocytopenia (low platelet count) or thrombocytosis (high platelet count).
Cancer Diagnosis: Identifying too many immature white blood cells in a blood smear can be an indicator for cancer disease.

Importance of this work

The ability to rapidly and accurately diagnose pathologies at an early stage is critical in modern medicine. Early detection, as demonstrated by various studies on cancer survival rates (Crosby et al., 2022). Such timely diagnosis not only saves lives but also reduces the burden of morbidity and healthcare resources. Several factors hinder us from achieving optimal diagnostic outcomes for these pathologies, but it is hard to ignore the potential impact of two key factors: a larger database and an automated cell type detection system. Two key factors hinder optimal diagnostic outcomes: limited data availability and a lack of automated cell detection systems.

Blood sample collection and microscopic imaging are relatively simple and inexpensive procedures. Moreover, these data are routinely collected in healthcare facilities worldwide. Therefore, advancements in automated cell type detection using readily available data hold significant promise for healthcare systems globally. However, the sensitive nature of this data presents a significant challenge to traditional model training approaches.

A thorough literature search revealed no existing studies that specifically apply federated learning to peripheral blood smear images. This work aims to bridge this gap by exploring and developing initial steps towards a robust FL-based system for hematologic cell type recognition.

This master's thesis will systematically evaluate the effectiveness of various supervised deep learning models within an HFL framework. The primary focus will be on classifying cell types in microscopic images of peripheral blood smears. Additionally, the model's robustness against potential privacy attacks will be investigated.

The expected outcome is a privacy-preserving FL model capable of accurately classifying cell types from peripheral blood smears. Such a model could serve as a valuable support tool for histopathology, enhancing diagnostic efficiency and accuracy.

1.2. Work's Objectives

a. General objectives

To develop a scalable and privacy-preserving deep learning model for the classification of normal peripheral blood cell images.

b. Specific Objectives

c.

i. **To develop** and evaluate a Federated Deep Learning tool for the detection and classification of cellular types on peripheral blood samples.

ii. The developed tool should be **scalable and suitable** (low requirements) for hospital's computers and connection.

iii. To conceive a **data privacy study/blueprint** for the developed AI tool.

1.3. Socio-ethical and Diversity Impacts

ODS 3 3 GOOD HEALTH AND WELL-BEING	This work aims for a faster and automatized detection of different pathologies via the automated classification of cellular types within peripheral-blood tissue samples. Doing so, we pursue for an improvement on citizen's health from different perspectives. First, a tool for automated pathology detection may become helpful to diagnose those patients who are being checked for any condition in which blood tissue image samples are likely to play a role. Second, an automated tool for pathology detection, could facilitate the early detection of different blood-related pathologies on patients, who firstly, were not necessarily being checked for that specific condition. Third, this tool may become useful to world areas and population that don't have an easy access to medicine. These are the reasons for whose, with this work, we aspire to align with 3rd ODS's statement, "Good Health and Well-being".
Data privacy	This project aims to propose a machine learning model that safeguards the privacy of data used for learning. In this way, it can facilitate collaboration among hospitals to train a joint model with more samples and likely greater diversity. It is anticipated that this collaboration will enhance the reliability and applicability of the models. Beyond hospital collaboration, the project also seeks to ensure the privacy of patients' personal medical data, preventing its inappropriate use.
Concerns about laboral intrusion	AI or automatic tools that perform histological tasks such as the ones studied in this work may generate concerns about laboral intrusion or replacement. This is far from reality. Automated tools for histological tasks are not intended to replace doctor's tasks of diagnosis. These tools aim for an assistance for professionals, to help them on their routine tasks such as blood cell counting or WSI analysing. A nice example of this, which is being performed on Catalonia is the DigiPatICS project (Temprana-Salvador et al., 2022).

1.4. Methodology

The nature of this project is scientific, that's why identifying and reading literature regarding similar issues is the first step forward to do. After doing so, taking care of the more specifical Machine Learning nature of the project, it is crucial to decide which or whose dataset/s are going to be used.

Once the State of the Art and the dataset points are solved, there are multiple ways to approach the experimental research of the Deep Learning and privacy/security solution to the main issue; To first search for the best of the models, fine tune it and then address the privacy topic. It could also be possible first look for the optimal privacy preserving solution and the deal with the deep learning part. Nevertheless, prioritizing the importance of the temporal resources required for this work, as well as the goal of drawing relevant conclusions, and with the guidance of the supervisor, it has been logical to opt for the following methodology.

Data preprocessing and feature selection: To enhance the scalability and adaptability of the model to the computational and connectivity capabilities of hospitals, we will conduct an initial study on the significance of different features within the dataset for label inference. Also, if needed, data normalization and scaling, like min-max scaling of pixels, will be performed.

Artificial Neuronal Network (ANN) model selection and global fine-tunning: After an initial reading phase, 2 to 3 pre-existing Artificial Neural Network (ANN) models will be selected, based on the performance on similar tasks, computational efficiency and literature about them. These models will be fine-tuned using a dataset of peripheral blood images. The initial fine-tuned models will serve as reference points for the validation metrics that subsequent models aim to approximate.

Data Distribution: Data will be distributed both in Independent and Identically Distributed (IID) and non-IID settings. In the case of non-IID data, various configurations will likely emerge to better understand the model's weak and strong points.

Federated Learning and Validation: If we successfully develop federated models with acceptable validation performance, considering the aforementioned reference, we will proceed to the next phase.

Understandability of the model: Aiming for the computational requirements reduction of the models, in this phase, the requirements and fastness of the different models will be studied. Also, if possible, model understanding methods will be applied to check if the images can be pre-processed before entering the model, so computational cost can be cut down.

Robustness Assessment: The subsequent phase involves studying the robustness of the different models against feature inference attacks (Privacy Attack) and Denial of Service Attack (DoS, Security Attack). We will attempt various relevant attacks based on existing literature (such as inference attacks, poisoning attacks, and denial-of-service (DoS) attacks). The goal is to evaluate the models' resistance to these attacks. This phase will likely be the final part of the experimental work, during which we aim to enhance the model's robustness against the most critical and plausible attacks in the context of medical data.

Model's Robustness and Performance trade-offs: When considering a Federated Learning architecture with non-IID data distribution, it is common to expect a poorer performance comparing to non-FL scenarios (Zou et al., 2023). The trade-offs between privacy and performance of the model are expected to appear while performing this study. The aim of this study is not to decide which is the correct balance between this to approaches, but to provide a first study on this topic on peripheral blood images.

As the experimental phase unfolds, the chosen ANN models will be compared, considering their validation metrics, computational requirements, and robustness against privacy and security attacks. As validation metrics, when classifying normal cellular types, the bet metrics may be Accuracy and F1-Score. This is because there is still no need of minimizing false positives (FP) or false negatives (FN), so we are looking for a balanced model, capable of correctly classifying different cellular types. If an approach to abnormal cellular types is finally performed, considering "normal" class as "0/Negative" and "abnormal" class as "1/Positive", relevance will go with minimizing the false negatives. When looking for abnormal cells, it is preferable to incorrectly classify a tissue as abnormal confirm the classification with other procedures. A tissue incorrectly classified as healthy may avoid posterior analysis, which may resume in bad consequences for the patient. This is why in the classification of abnormal/normal cellular types, a high Recall metric is essential.

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FN)}$$

Formula 1. Formula Recall evaluation metric calculation.

Formula 1. Formula Recall evaluation metric calculation.

Finally, based on the entire body of work, we will formulate the concluding remarks. If deemed useful, we will also highlight important takeaways.

1.5. Working Plan

1.5.1 Tasks

Every specific objective is split into several realizable and evaluable tasks. Also, a short tag, in bold, is added to every task to identify it on the calendar.

To develop a Federated Deep Learning tool for the detection and classification of cellular types on peripheral blood samples.
Reading literature about previous applications of Federated Learning on medical data. **FL-Read (ANN-1)**
Study of the peripheral blood dataset. **Dataset-Study (ANN-2)**
Data normalization and feature selection of the dataset. **Data-Prep (ANN-3)**
Selecting 3-4 already developed classification models for images. **Model-selection (ANN-4)**
Selection of the validation metrics and graphs that are going to be used for comparison of the models. **Validation-Selection (ANN-5)**

Fine tuning the selected models on a non-federated configuration. **Fine-Tunning (ANN-6)**
Distribution of the data on an IID configuration. **IID-Distribution (ANN-7)**
Distribution of the data on a non-IID configuration. **Non-IID-Distribution (ANN-8)**
Federated architecture design and code. **Federated-Architecture (ANN-9)**
Fine tuning of the federated models. Taking account of federated learning specific parameters: Number of local rounds, number of clients that train each round. **Fine-tuning-FL (ANN-10)**
Validation comparison between models. **Validation-Comparison (ANN-11)**
Model understandability test. **White-Box (ANN-12)**
Final comparison and conclusions.

Table 2. Specific Objective 1 Tasks.

The developed tool should be scalable and suitable (low requirements) for hospital's computers and connection.
Reading literature of this topic within Machine Learning and Federated Learning. **Scalable-Read (SCAL-1)**
Learning about the standard computation capacities of hospitals. **Hospital-Requirements (SCAL-2)**
Recording fastness and computational hardware use of every model and training. **Scalable-Metrics (SCAL-3)**
Training the models using CPU and GPU and comparing the output. **GPU-CPU (SCAL-4)**
Avoid over-fitting of the models. **Overfitting (SCAL-4)**
Study the minimum local-epochs necessary that are necessary to effectively train each round. **Local-Epochs (SCAL-5)**
Comparison of computational and scalability potential of models. **Computational-Comparison (SCAL-6)**
Final comparison and conclusions.

Table 3. Specific Objective 2 Tasks.

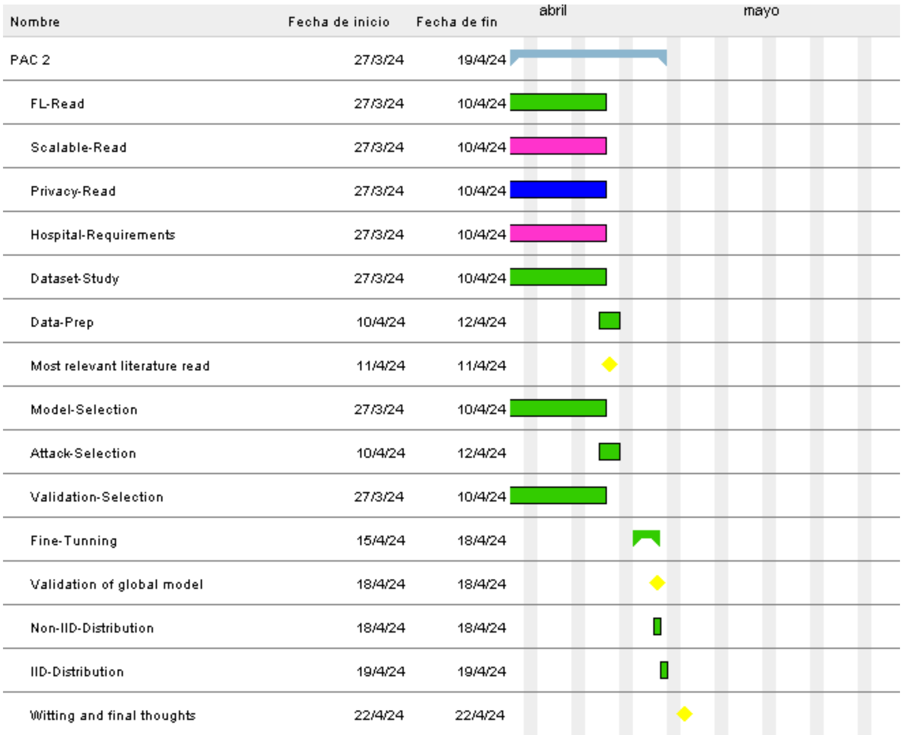
To conceive a data privacy study/blueprint for the developed AI tool.
Reading literature regarding privacy and security topics within federated learning and ANN. **Privacy-Read (ATTACK-1)**
Selection of the attacks that will be performed to the models. **Attack-Selection (ATTACK-2)**
Performing of the attacks. **Attack-Performing (ATTACK-3)**
Evaluation of the robustness of the models to attacks. **Attack-Eval (ATTACK-4)**
Comparison of robustness of the models. **Attack-Comparison (ATTACK-5)**
Explain, if possible, the vulnerability causes of every model. **Attack-explain (ATTACK-6)**
Improve, if possible, the robustness of every model. Strategy proposal. **Robustness-Improve (ATTACK-7)**

Explain the consequences of the vulnerability of every model. **Attack-Consequences (ATTACK-8)**
Final comparison and conclusions.

Table 4. Specific Objective 3 Tasks.

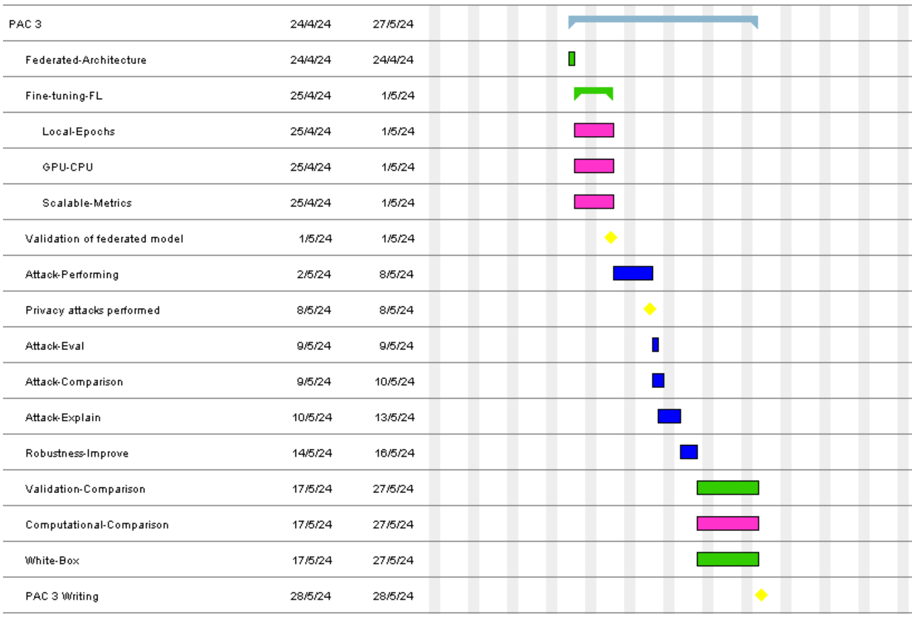
1.5.2 Gant Chart

In order to optimize task allocation within the available time frame and considering the credit requirements for the Master’s Thesis, a comprehensive evaluation has been conducted using a Gantt chart created (Table 5, 6, 7) with the Gantt Project software. It is important to note that certain items may appear more as considerations or methods rather than discrete tasks. However, for clarity and ease of visualization alongside relevant tasks, these items have been categorized as tasks.



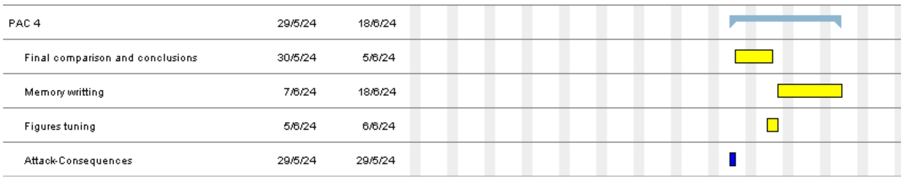
Nombre	Fecha de inicio	Fecha de fin
PAC 2	27/3/24	19/4/24
FL-Read	27/3/24	10/4/24
Scalable-Read	27/3/24	10/4/24
Privacy-Read	27/3/24	10/4/24
Hospital-Requirements	27/3/24	10/4/24
Dataset-Study	27/3/24	10/4/24
Data-Prep	10/4/24	12/4/24
Most relevant literature read	11/4/24	11/4/24
Model-Selection	27/3/24	10/4/24
Attack-Selection	10/4/24	12/4/24
Validation-Selection	27/3/24	10/4/24
Fine-Tunning	15/4/24	18/4/24
Validation of global model	18/4/24	18/4/24
Non-IID-Distribution	18/4/24	18/4/24
IID-Distribution	19/4/24	19/4/24
Witting and final thoughts	22/4/24	22/4/24

Table 5. PAC 2 Gantt Chart



PAC 3	24/4/24	27/5/24
Federated-Architecture	24/4/24	24/4/24
Fine-tuning-FL	25/4/24	1/5/24
Local-Epochs	25/4/24	1/5/24
GPU-CPU	25/4/24	1/5/24
Scalable-Metrics	25/4/24	1/5/24
Validation of federated model	1/5/24	1/5/24
Attack-Performing	2/5/24	8/5/24
Privacy attacks performed	8/5/24	8/5/24
Attack-Eval	9/5/24	9/5/24
Attack-Comparison	9/5/24	10/5/24
Attack-Explain	10/5/24	13/5/24
Robustness-Improve	14/5/24	16/5/24
Validation-Comparison	17/5/24	27/5/24
Computational-Comparison	17/5/24	27/5/24
White-Box	17/5/24	27/5/24
PAC 3 Writing	28/5/24	28/5/24

Table 6. PAC 3 Gantt Chart



PAC 4	29/5/24	18/6/24
Final comparison and conclusions	30/5/24	5/6/24
Memory writting	7/6/24	18/6/24
Figures tuning	5/6/24	6/6/24
Attack-Consequences	29/5/24	29/5/24

Table 7. PAC 4 Gantt Chart

1.6. Obtained Products

Result	Description
Work's Document	The project involves crafting a detailed report that includes an introduction, a survey of existing knowledge, the methods used, the results obtained, and the final conclusions. A key goal is to assess the performance, efficiency, and security of different federated learning models.
Classification Models	Several models have been successfully developed that accurately predict the cell type of peripheral blood cells. The text will be revised to ensure clarity and formality while maintaining a level of accessibility appropriate for a broad academic audience.

Table 8. Obtained product descriptions.

2. State of the art

Digital pathology has been a transformative force in medical imaging, facilitating the analysis of medical images for disease diagnosis and classification. The advent of computer-aided diagnostic systems necessitates an effective classification model to support medical professionals in their decision-making processes. The DigiPatICS project (Temprana-Salvador et al., 2022) successfully developed a Deep Neural Network (DNN) model that aids medical experts (Figure 3) throughout Catalonia in analysing Whole Slide Images (WSIs) of stained breast tissue samples. These images encompass tens of thousands of cells, yet according to the World Health Organization (WHO) protocol, only about 1,000 are typically counted during routine analysis, focusing on five areas of the WSI deemed pertinent by the histopathologist. The DigiPatICS tool enables the automated enumeration and categorisation of all cells within the WSI, thus providing substantial support to healthcare professionals.

Figure 3. Residents diagnosing using a 55-inch 4K UHD monitor (55UH5F-B) to analyse WSI of breast tissue sample. (Temprana-Salvador et al., 2022)

In hematology, numerous studies have corroborated the proficiency of deep learning models in identifying various cell types, including both white and red blood cells, within tissue or peripheral blood samples. (Kohsasih et al., 2022) conducted a comprehensive evaluation of renowned classifiers such as VGG16, VGG19, ResNet50, and AlexNet, assessing their ability to accurately identify lymphocytes, eosinophils, monocytes, and neutrophils in single-cell images. Among these, ResNet50 was distinguished by its exceptional performance, achieving an accuracy rate of 99%.

Substantial advancements have been achieved in the deployment and integration of deep learning paradigms within the medical sector. Nevertheless, these sophisticated models necessitate a plethora of diverse datasets throughout the training phase. Should the volume and integrity of data prove inadequate, the resultant models may exhibit deficient generalisation capabilities, thereby undermining their efficacy in practical scenarios. Moreover, there is a need for iterative retraining with novel datasets to ensure the models remain attuned to evolving real-world conditions. A salient resolution to the quandary of data scarcity is the establishment of collaborative networks amongst medical institutions. (Hu & Chaddad, 2023) elucidate the merits of a nascent machine learning paradigm, Federated Learning (FL), as postulated by (Brendan McMahan et al, 2017) in facilitating collaboration between hospitals FL facilitates the localised training of Deep Neural Networks (DNNs) across disparate nodes, culminating in the centralised amalgamation of their parameters through successive iterations, engendering a model proficient across all participating nodes.

The heightened sensitivity associated with medical imagery has accentuated its significance in the evolution of machine learning apparatuses for analytical purposes. FL propounds a framework wherein disparate healthcare establishments can collectively refine models, without necessitating the transfer of images beyond their originating repositories.

In the context of discerning anomalous lymphocytes within oncogenic tissues, as investigated by (Baid et al., 2022) FL has been instrumental in deriving robust models without extensive sample availability. Furthermore, FL's intrinsic privacy-preserving attribute is underscored. The distribution of data amongst clients during FL model training is pivotal. An Independent and Identically Distributed (IID) data configuration implies homogeneity across all nodes/clients. Conversely, a non-IID distribution, which mirrors real-world conditions more closely, entails heterogeneous data distributions among clients (Brendan McMahan et al, 2017). Substantial advancements have been achieved in the deployment and integration of deep learning paradigms within the medical sector. Nevertheless, these sophisticated models necessitate a plethora of diverse datasets throughout the training phase. Should the volume and integrity of data prove inadequate, the resultant models may exhibit deficient generalisation capabilities, thereby undermining their efficacy in practical scenarios. Moreover, there is a necessity for iterative retraining with novel datasets to ensure the models remain attuned to evolving real-world conditions. A salient resolution to the quandary of data scarcity is the establishment of collaborative networks amongst medical institutions.

Pertaining to privacy concerns, (Zhu & Han, 2020) have demonstrated the feasibility of reconstructing training images from the gradients of updated model parameters (weights and biases). This revelation, alongside other potential privacy breaches inherent to FL, has catalysed a plethora of studies dedicated to the development of FL-trained models fortified with robust privacy safeguards. In the research conducted by (Adnan et al., 2022) , the Differential Privacy (DP) framework (Figure 5), initially conceptualised by (Dwork, 2006) was employed as a privacy bulwark in the federated training involving Whole Slide Images (WSIs) of stained tissues. DP, heralded as a potentially universal privacy safeguard, entails the obfuscation of sensitive data—herein, the training model parameters—through the application of various techniques, such as Stochastic Gradient Descent (SGD).

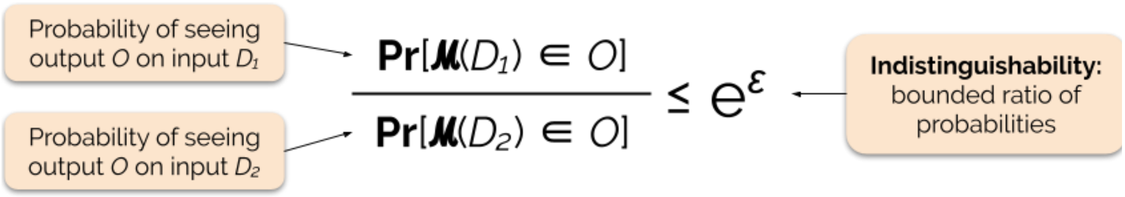

Probability of seeing output O on input D_1
Probability of seeing output O on input D_2
$\frac{\Pr[\mathcal{M}(D_1) \in O]}{\Pr[\mathcal{M}(D_2) \in O]} \leq e^{\varepsilon}$
Indistinguishability: bounded ratio of probabilities

Figure 5. Definition of Differential Privacy (DP). Function (M), Output(O), Dataset-1(D1), Dataset-2(D2), ε (Indistinguishability parameter). In an optimal case, D1 and D2 only differs from a sample.

The equation in the figure indicates that, given the same random function (M) applied to two different datasets, the difference between the two outputs should be as small as possible. Factor ε plays a crucial role; the smaller the ε, the greater the indistinguishability of the datasets based on their output.

In this study, we aim to compare the performance and efficiency of different image classification models, with and without Differential Privacy (DP), for the particular case of automatic classification of white blood cell images.

3. Materials and methods

3.1 The framework – Tensorflow and cloud computing machine

The frameworks selected to carry out the work have been TensorFlow (Abadi, Agarwal, et al., 2016) and Keras (Chollet, F., 2015). Both are known open-source machine learning frameworks for Python and were firstly chosen because their wealth of documentations, examples and can also integrate Tensorflow Federated, which facilitates Federated Learning training of Tensorflow models.

It was also an initial determination factor that Tensorflow can use the module Tensorflow Privacy (Abadi, Chu, et al., 2016), which has different privacy related functions for machine learning work, such as optimizers with Differential Privacy and privacy metrics calculators. Moreover, newest version Tensorflow Privacy 0.9.0, ensures compatibility with Keras models.

To train the Machine Learning models, jarvislabs.ai was chosen as a cloud platform to have access to high-end GPUs like the NVIDIA RTX 5000.

Code used can be found at: https://github.com/agarciall/TFM

3.2 Classification Models

ResNet50

ResNet50 is a widely used deep convolutional neural network (CNN) architecture for image classification tasks. Developed by (He et al., 2015), ResNet50 has become a benchmark model for large-scale classification, achieving state-of-the-art performance on the ImageNet dataset.

ResNet50 has an input size of 224, 224, 3 and is composed of 50 convolutional layers, interspersed with residual units. These residual units (Figure 6) are the key element of the ResNet50 architecture and are responsible for its success. A residual unit consists of two convolutional layers followed by an element-wise sum. This sum allows the output of the first convolutional layer to be passed directly to the output of the residual unit, without going through the second convolutional layer. This creates a "shortcut" that allows information to flow directly through the network, mitigating the vanishing gradient problem that can arise in deep networks, in which earlier layers struggle to learn from errors during backpropagation.

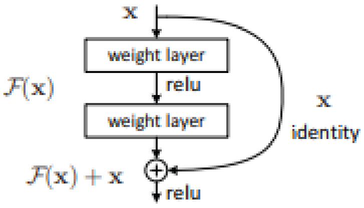


Figure 6. Residual Learning building block. (He et al., 2015)

VGG16

VGG16, is a deep learning model designed for image recognition tasks. It was developed by (Simonyan & Zisserman, 2014) of the Visual Geometry Group (VGG) at the University of Oxford in 2014. VGG16 achieved impressive results on image classification benchmarks, making it a popular choice for computer vision tasks.

VGG16 consists of 16 convolutional layers, followed by 3 fully-connected layers for classification (Figure 7). This stacked architecture allows the network to learn complex features from the input images. VGG16 utilizes 3x3 filters throughout the convolutional layers. While these small filters may seem less powerful than larger ones, stacking multiple layers of 3x3 filters allows the network to capture a wider range of features effectively. Max pooling layers are strategically placed throughout the network to reduce the feature maps and reduce computational cost.

The final part of the network comprises three fully connected layers. These layers transform the high-level features extracted by the convolutional layers into class probabilities.

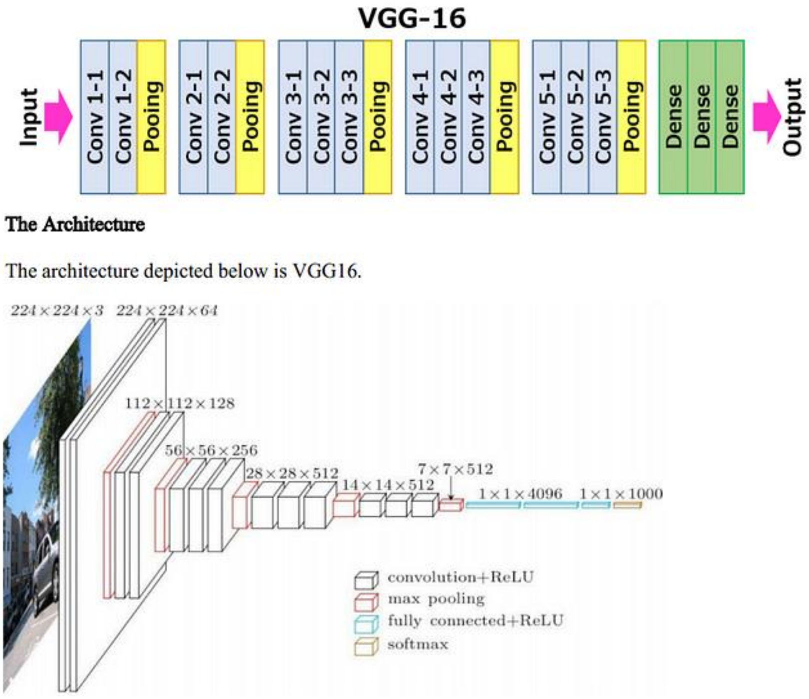


Figure 7. Layer composition of VGG16. Thaker, Nerd for Tech.

Head Model and Trainable Layers

“Head model” refers to the upper segment of the model tasked with the specific classification challenge we aim to address. This Head Model (Figure 8) is appended to the culmination of the “backbone” models, which have been pre-trained with images from the ImageNet dataset. It is stipulated that 30% of the layers in the final model will be employed for training, whilst the remainder will remain static.

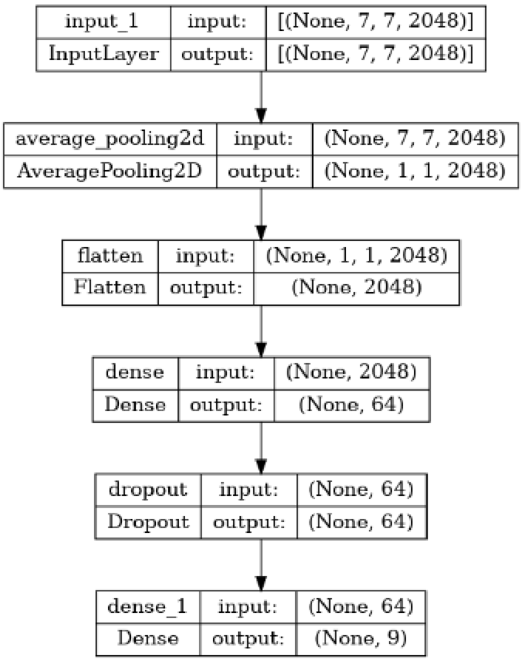

input_1 | input: | [(None, 7, 7, 2048)]
InputLayer | output: | [(None, 7, 7, 2048)]
average_pooling2d | input: | (None, 7, 7, 2048)
AveragePooling2D | output: | (None, 1, 1, 2048)
flatten | input: | (None, 1, 1, 2048)
Flatten | output: | (None, 2048)
dense | input: | (None, 2048)
Dense | output: | (None, 64)
dropout | input: | (None, 64)
Dropout | output: | (None, 64)
dense_1 | input: | (None, 64)
Dense | output: | (None, 9)

Figure 8. Head Model shaped to ResNet50 output.

3.3 Datasets

For the training of models using collaborative learning, Federated Learning, it is necessary to distribute data across clients or nodes. Each client will represent a medical organisation. In the case of Independent & Identically Distributed (IID) data distribution, the data will be separated into 5 different clients, all following the same probabilistic distribution. Essentially, the IID distribution of the data leaves 5 clients with apparently identical characteristics regarding data composition. However, when distributing data in a non-IID manner, we encounter the following question: How should we distribute the data to approximate a real-case scenario?

(Adnan et al., 2022) distributed the data ensuring that different clients received different proportions of images of cells related to two subtypes of non-small cell lung cancer (NSCLC): Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC).

(Cetinkaya et al., 2021) obtained a non-IID data distribution scenario by having chest X-ray images labelled into 4 different classes: Covid-19, Pneumonia, Lung Opacity, and Normal.

In our case, the dataset we are using corresponds to images of single cells, classified into labels corresponding to the cellular type they belong to (lymphocytes, eosinophils, etc.).

There is no clear criterion for distributing the dataset's data in a non-IID manner. It has finally been decided to merge two similar datasets and use every image's property of belonging to one another original dataset to generate IID and non-IID client scenarios.

Dataset from Hospital Clinic de Barcelona

The dataset by (Acevedo et al., 2020), from now on referred as Acevedo et al or PBC_DIB, contains 17,092 images of individual normal peripheral blood cells. Blood smears were automatedly stained with May Grünwald-Giemsa in the autostainer Sysmex SP1000i and images were obtained using the analyser CellaVision DM96 in the Core Laboratory at the Hospital Clinic of Barcelona. The images are in the format of jpg and have a resolution of 360 × 363 pixels. They were annotated by expert clinical pathologists and are organized in eight groups: neutrophils, eosinophils, basophils, lymphocytes, monocytes, immature granulocytes (promyelocytes, myelocytes, and metamyelocytes), erythroblasts, and platelets or thrombocytes (Table 9) (Figure9).

Images were collected during the period 2015-2019 from blood smears from patients without infection, hematologic or oncologic disease, and free of any pharmacologic treatment now of blood extraction.

Cell type	Abundance	Relative Abundance
Neutrophils (Segmented and Band)	3329	19.47
Eosinophils	3117	18.23
Basophils	1218	7.12
Lymphocytes	1214	7.10
Monocytes	1420	8.30
Immature Granulocytes	2895	16.93
Erythroblasts	1551	9.07
Platelets	2358	13.79
TOTAL	17102	100

Table 9. Abundance and Relative abundance of each class within (Acevedo et al., 2020)

This was the first dataset used to train the models. At an initial state of the model, it was conceived to be the only dataset to be used.

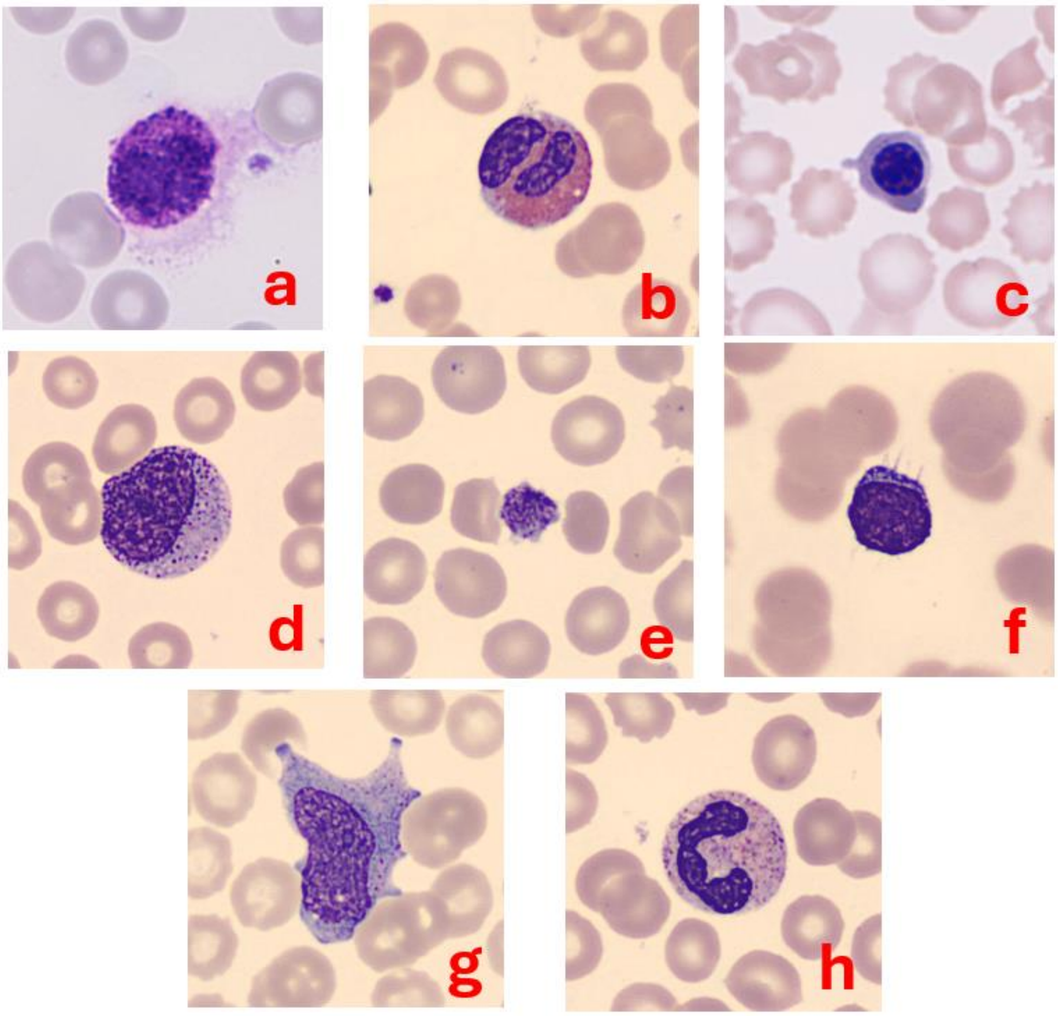


Figure 9. Examples of images belonging to different classes of PBC_DIB dataset. Basophil (a), Eosinophil (b), Erythroblast (c), Immature Granulocytes (d), Platelet (e), Lymphocyte (f), Monocyte (g), Neutrophil (h)

High Resolution Peripheral Blood Cells Dataset (HRD)

(Bodzas et al., 2023) dataset contains a total of 16,027 annotated white blood cell samples. It includes nine types of white blood cells, including neutrophil segments and bands, eosinophiles, basophiles, lymphocytes, monocytes, nucleated red blood cells (NRBC), and immature cells of both myeloid and lymphoid lineage. All blood smear samples were stained manually with May-Grünwald and Giemsa-Romanowski staining solution.

The images were captured using high-quality acquisition equipment, Basler acA5472-17uc camera mounted into an Olympus BX51 microscope, resulting in an approximate resolution of 42 pixels per 1 μm, or 5472 × 3648 pixels.

In this case, of 78 total patients from which blood smears were taken, 18 patients were diagnosed with acute myeloid leukemia, 15 patients suffered from acute lymphoid leukemia, and 45 patients did not show any pathological findings or had a non-leukemic diagnosis. This is why this dataset contains the typical 5 classes of normal blood cells found in peripheral blood samples (Basophiles, Eosinphiles, Lymphocytes, Monocytes, Neutrophile Segmented) along with 4 classes of immature blood cells (Myeloblasts, Lymphoblasts, Neutrophile Band, Normoblasts) (Table 10) (Figure 10).

Class	Abundance
Basophile	1023
Eosinophile	1017
Lymphoblasts	2557
Lymphocyte	3046
Monocyte	2040
Myeloblast	2534
Neutrophile Band	99
Neutrophile Segment	3201
Normoblast	510

Table 10. Counts of every class of HRD dataset.

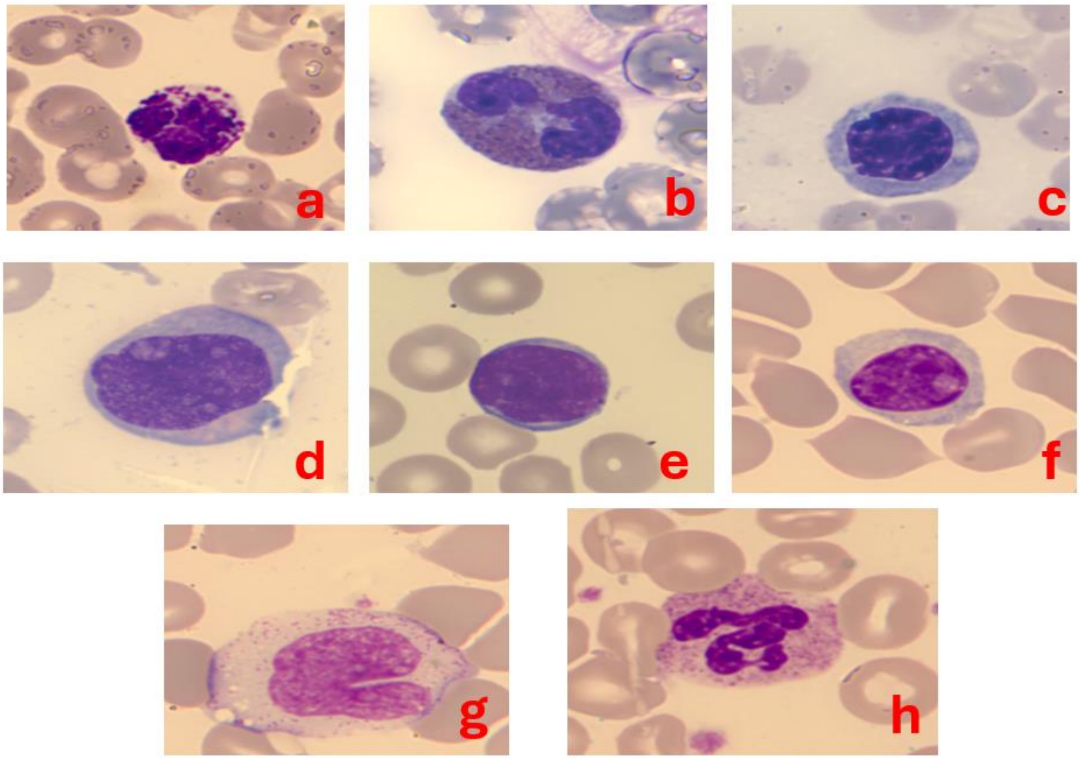


Figure 10. Examples of images belonging to different classes of HRD dataset. Basophil (a), Eosinophil (b), Erythroblast (c), Immature Granulocytes (d), Lymphoblast (e), Lymphocyte (f), Monocyte (g), Neutrophil (h)

Merged Dataset (MergeD)

MergeD is the result of merging both datasets (Figure 12). Firsts, classes of this new dataset were designed (Table 11).

MergeD	Acevedo et al class	HRD class	TOTAL
Neutrophils	Neutrophil Segmented	Neutrophile Segment	3.328
Eosinophils	Eosinophil	Eosinophile	3117
Basophils	Basophil	Basophile	2242
Lymphocytes	Lymphocyte	Lymphocyte	2428
Monocytes	Monocyte	Monocyte	2840
IG	ig	Myeloblast	2000
Erythroblasts	Erythroblast	Normoblasts	2020
Lymphoblasts	x	Lymphoblasts	2557
Platelet	Platelet	x	2345
TOTAL	x	x	**22.824**

Table 11. Original dataset class of MergeD classes dataset and counts.

Neutrophils

In HRD, Neutrophils are split on two different classes: Band and Segmented. In PBC, Band and Segmented Neutrophils are identified differently but considered the same class. To make the model more specific on blood cell type recognition, only segmented Neutrophil images are used on the new dataset.

Eosinophils, Basophils, Lymphocytes, Monocytes

These classes are merged without need of further logic, as those are the same in both datasets.

Immature Granulocytes (IG)

Immature Granulocytes (IG) is a class that compresses 3 different cellular types on the PBC dataset: Metamyelocytes, myelocytes, promyelocytes. In the HRD, it found the Myeloblast class, being those Immature Granulocytes too (Figure 11).

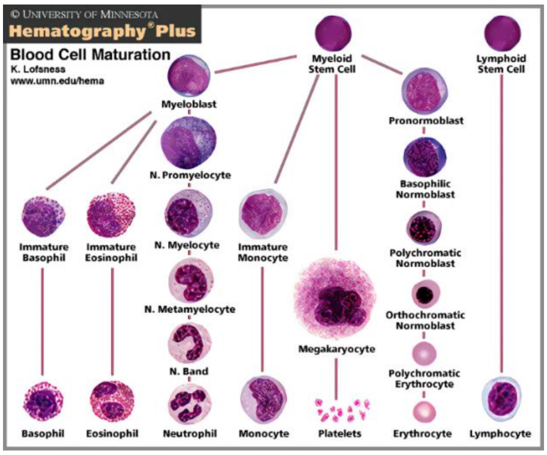


Figure 11. Description of WBC lineage, differentiation process of two main principal stem cells, Myeloid Stem Cell and Lymphoid Stem Cell.

Erythroblasts

Immature Erythrocytes are found in both datasets: Erythroblasts class for Acevedo et al and Normoblasts class for HRD.

Lymphoblasts

Lymphoblasts is a unique class for HRD, so it will not have heterogeneous composition of original datasets.

Platelet

Platelet is a unique class for Acevedo et al, so it will not have heterogeneous composition of original datasets.

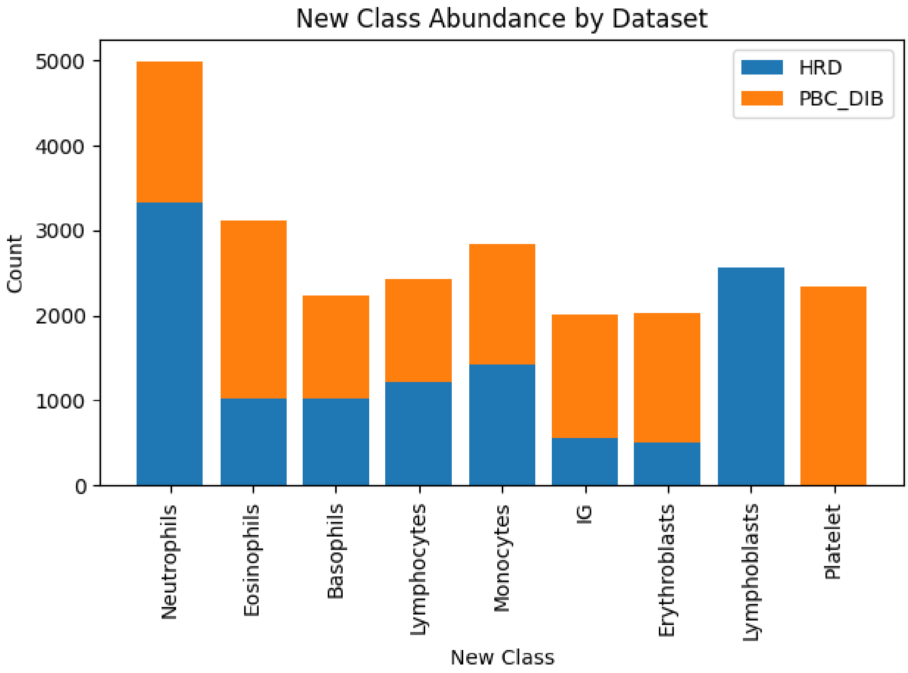


Fig 12. MergeD class composition of the original datasets of images.

3.4 Data Augmentation

To prevent model overfitting and ensure generalisation in the learning of the neural network, a series of ‘on the fly’ data transformations are applied in every model training. These transformations include:

Rotation

Images are randomly rotated in the range of -20 to +20 degrees.

Vertical and Horizontal Flip

Images are randomly flipped horizontally or vertically.

Vertical and Horizontal Shift

Images will be randomly shifted vertically and horizontally within a range of 10% of their total height or width.

Shear

Shifts each point of an image in a fixed direction, altering the shape while maintaining the image area. The bottom and top margins retain their length, while the right and left margins are elongated, distorting the image.

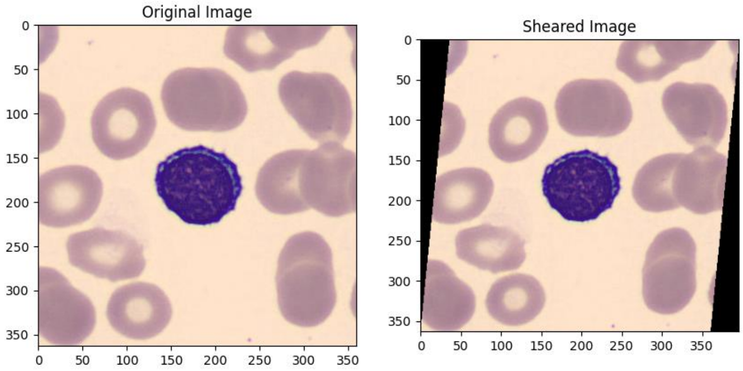


Figure 13. Example of a shear transformation using a shear factor of 0.1.

Zoom

Applies a random zoom to the images, in the range of 40% (zoom in) to 160% (zoom out).

Brightness

Randomly changes the brightness of the images, within the range of 20% (darker) to 100% (original brightness).

3.5 Pre-processing and normalization

In the context of image data, each pixel is represented by a value between 0 and 255 (for 8-bit images). By rescaling these values by 1/255, we transform them to a range between 0 and 1.

This is a common pre-processing step in image processing tasks, especially in deep learning, as it can help the model converge faster during training and can lead to improved model performance.

Also, when loaded to the training generator, all images are loaded as with a resolution of 224x224, to fit with the pre-trained VGG16 and ResNet50 models input shapes.

3.6 Data Distribution

Train/Test split

MergeD dataset is going to be randomly split into train (80% of original dataset) and test (20% of original dataset) (Figure 14). These two sets are going to be the same for global and federated training of the model, only with differences on how train data is used.

Global Training

For the comprehensive training of the classification models, the described train dataset will be employed, after being randomly divided into an 80% allocation for the training subset and a 20% allocation for the validation subset.

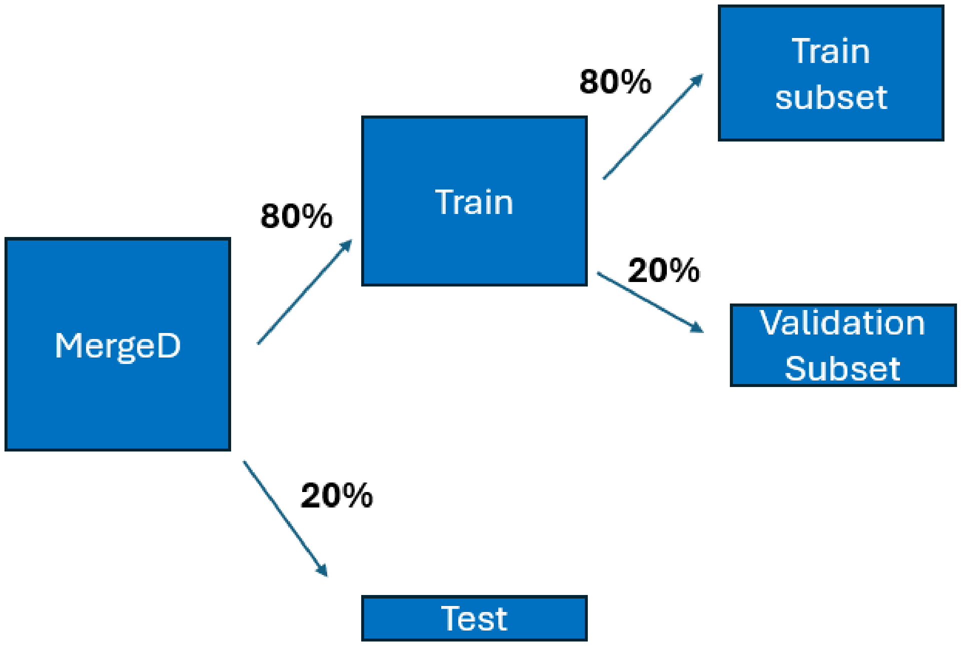


Figure 14. Scheme of train test split of the original dataset and further training and validation subsets of train images.

The validation subset is a collection of images extracted from the training partition. It is not directly utilised for training the model but serves to monitor that the model is not overfitting. At the conclusion of each training round, the model will predict the classes of the validation images and will provide a validation accuracy and loss, without these directly impacting the model's parameter training. These samples are not used for the model's backpropagation. Nevertheless, as we obtain the validation results each round, we can verify whether the model is learning general characteristics of each class or if it is learning patterns specific to the training set, leading to overfitting. For the subsequent evaluation of the models, the parameters of the model from the round in which the highest validation accuracy was obtained will be always used.

Federated Training

In the Federated Learning scheme, data is split among 5 clients, whose collaborate to train a global model capable of correctly perform as a prediction model on every client separate data.

Depending on if it's a IID (Identically Independent Distributed) scenario or a non-IID scenario, client's composition may be the same or no. It is important to note that for data to be non-IID or IID distributed, we must decide which property or characteristic is considered interesting to be differential among clients. In this project, we decided that differential property of MergeD dataset is sample's belonging to only one of the two original datasets that were merged to create MergeD, (Acevedo et al., 2020) dataset or HRD.

IID Distribution

In an IID distribution, the differential property must remain equally and independent distributed, as the acronym calls. This means that every client must be as identic to the others as possible (Figure 15, 16, 17).

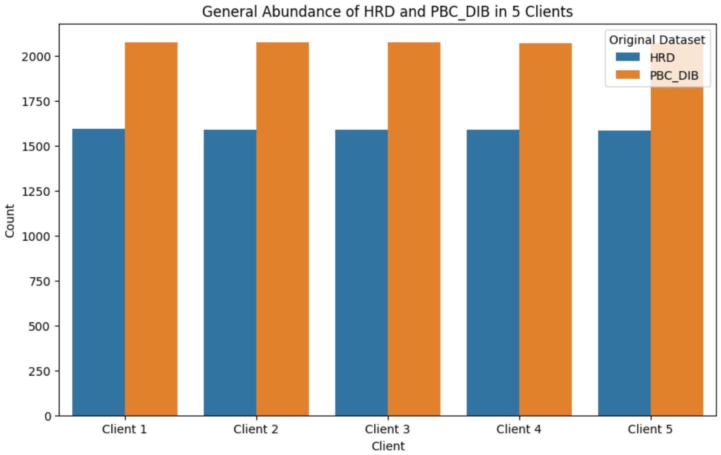


Figure 15. Distribution of HRD and PBC_DIB images within IID clients.

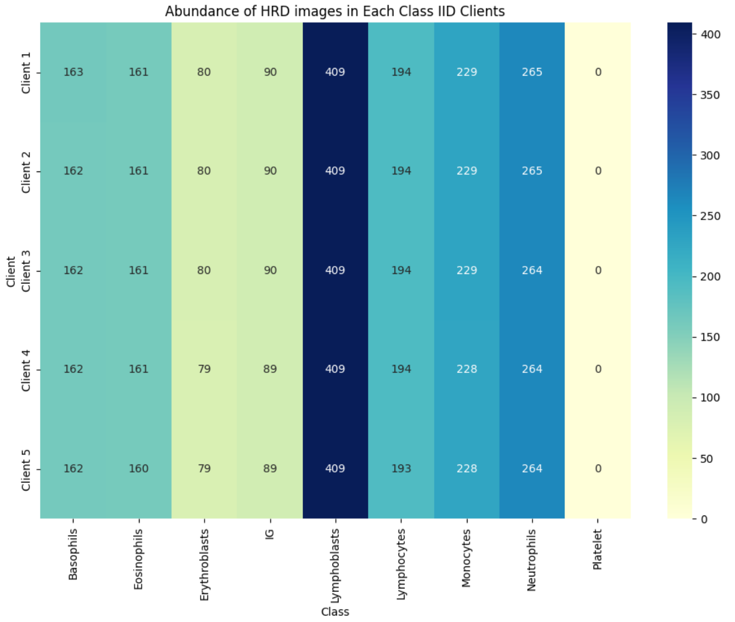


Figure 16. Heat-map of abundance of every original class of every IID client that initially belonged to HRD.

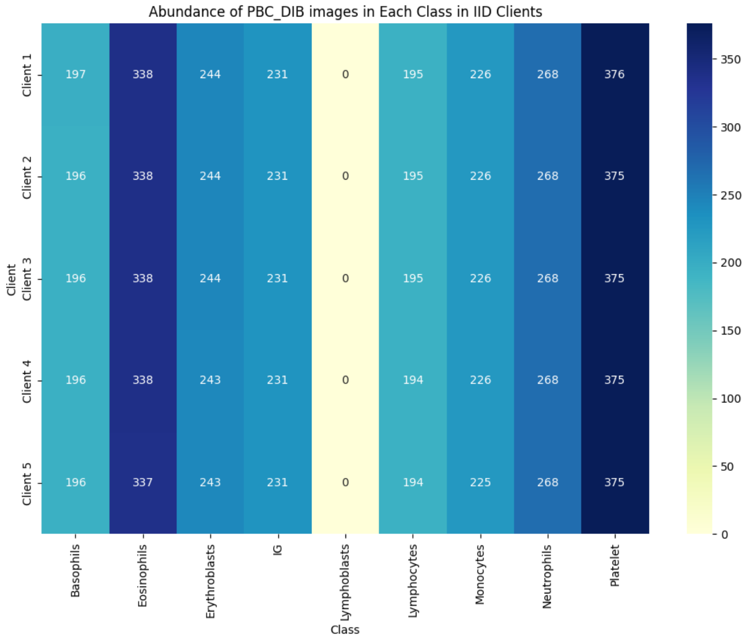


Figure 17. Heat-map of abundance of images, within IID-clients from every class, within IID-clients, of every client that initially belonged to (Acevedo et al., 2020), PBC_DIB dataset.

Despite being IID-distributed, MergeD dataset has 2 classes whose samples only exists on PBC_DIB Dataset (Platelet) and HRD (Lymphoblasts).
Also, it must be noted that data was randomly mixed before filling clients with each class’s samples, so clients’ data is as equal as possible.

No-IID

To generate clients with differently distributed data, abundance of data originally pertaining to HRD and PBC_DIB was split on an uneven way among clients(Figure 18, 19, 20).

Client 1: Fully composed of HRD images
Client 2: Client where PBC_DIB and HRD images are more levelled
Client 3: HRD images are more abundant than PBC_DIB images.
Client 4: Fully composed of PBC_DIB images.
Client 5: PBC_DIB images are much more abundant than HRD images.

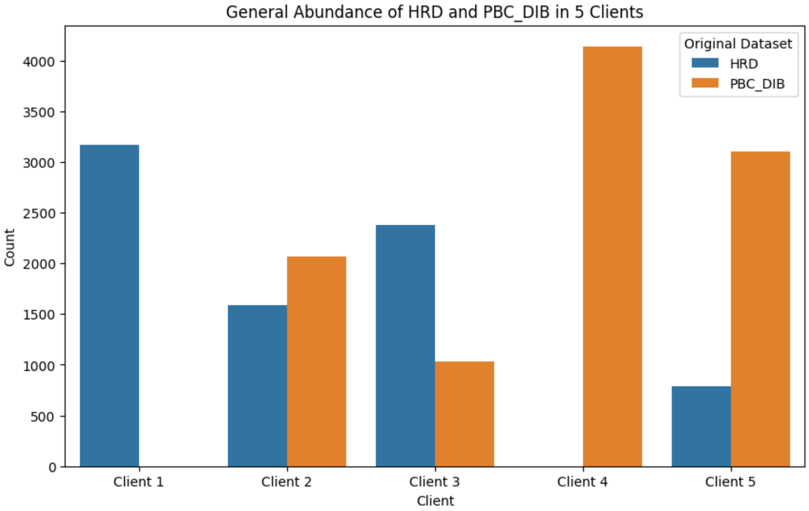


Figure 18. Distribution of original dataset property among 5 non-IID clients.

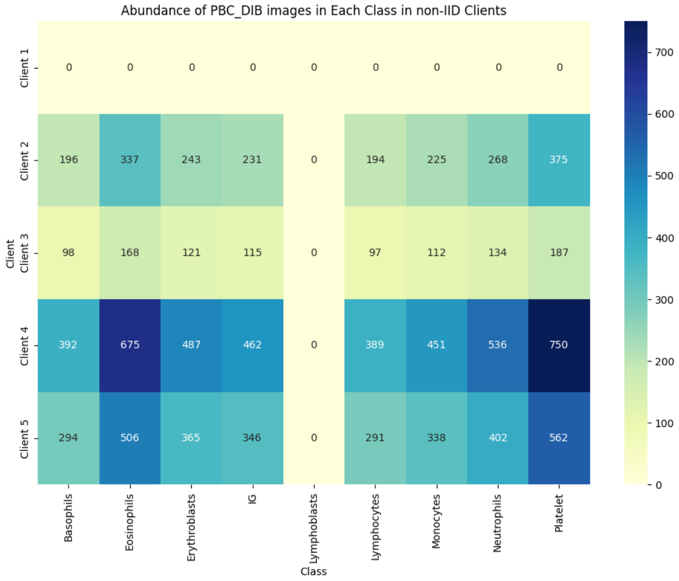


Figure 19. Heat-map of abundance of images, within non-IID clients, that initially belonged to (Acevedo et al., 2020), PBC_DIB dataset.

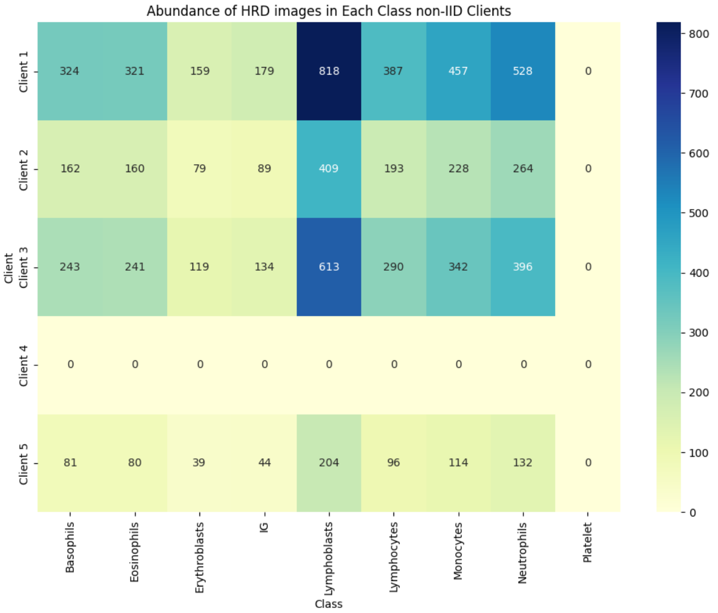


Figure 20. Heat-map of abundance of images, within non-IID clients, that initially belonged to HRD.

3.6 Training of models

Instance weighting

An imbalanced dataset is one where some classes have many more examples than others. This can cause a machine learning model to become biased towards the majority class, resulting in poor performance on the minority class.

Instance weights are a way to tell the model to "pay more attention" to certain instances during training. By assigning a higher weight to an instance, it increased the contribution to the loss function that the model is trying to minimize. This means that the model will try harder to get these instances correct.

Global Training

On the Global training, models are trained with the same data for 60 epochs. Data amplification is performed one time, on the fly, when data is loaded to the model before round 1 training.

Backpropagation and fine-tuning of model’s parameters is done by Adam optimizer, with a learning rate of 0.0001. Batch size of images on the training subset is 8. Batch size of images on the validation subset is 1. After each round of training, new model is going to be validated with the same validation subset, generating validation metrics (Figure 21). The model from the round that performs better on the validation is going to be used to evaluate the model later with the test subset.

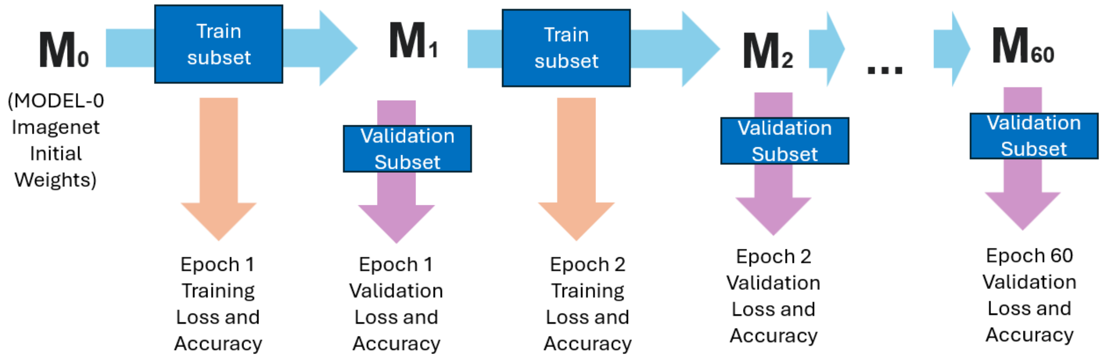

M0
(MODEL-0 Imagenet Initial Weights)
Train subset
M1
Validation Subset
Train subset
M2
Validation Subset
...
M60
Validation Subset
Epoch 1 Training Loss and Accuracy
Epoch 1 Validation Loss and Accuracy
Epoch 2 Training Loss and Accuracy
Epoch 2 Validation Loss and Accuracy
Epoch 60 Validation Loss and Accuracy

Figure 21. Simplified scheme of how global training of models is performed

Federated Training

When training a model using Federated Learning, some extra decisions must be made compared to usual global training ML.

Number of epochs and rounds

Federated Learning training involves two kinds of rounds/epochs. The times that each client trains the model with its data before sending the weights to the local server are going to be called *epochs* from now on. The times that the global server aggregates the weights received from the clients are going to be called *rounds*.

In all trainings performed in this study, the number of rounds is 40 and the number of local epochs is 2.

Federated aggregating protocol - FedAvg

This protocol is the responsible for combining all the parameters that the global server receives each round, averaging them and resolving into a new unified model that aims to correctly fit every client.
FedAvg is a simple and effective protocol for federated learning, although it assumes that all clients have equal importance and contribute equally to the global model.

Sample usage in each round

In the epochs of each round, clients could be using all their data or just a part of it. To simulate a real scenario, makes sense to fraction data in X parts and perform X rounds to train with all the data. Each client's data is split into 40 different parts, representing 40 weeks of a year of sampling at 5 different hospitals. Every client has 40 different parts of its whole data. Two cases of how these part's data is used are planned.

> Case 1
> Every round, every client is going to train the model for 2 epochs, starting with partition 1 on round 1 and finishing with partition 40 on round 40. In this case, each partition of every client only trains at the corresponding part and it's never used again (Figure 22).

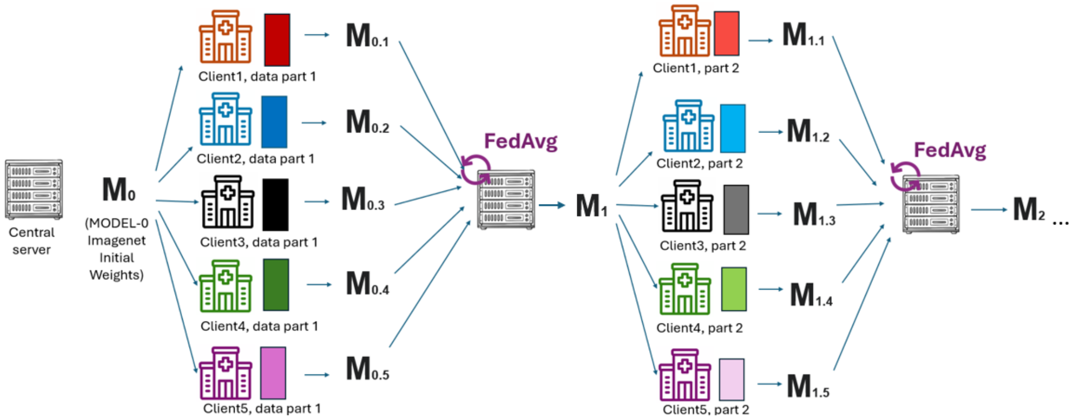


Figure 22. Scheme explanation of case 1 training. Every round, global model is trained with a partition of the data in each client. This partition is never used again to train the model.

Case 2

In this scenario, data is recycled along rounds. Every round, when a client finishes training, a percentage (in this case, 60%) of the data that has been used is added to next round's data partition of that client (Figure 23). Although this recycling, as seen in Figure 24, as the initial size of every partition is the same, this data addition reaches a top in which next partitions don't grow any larger than the last one.

This case may improve model's training, as partitions of training data are mixed along round but the size of training data for each round keeps being very short when compared to Global Training. Also, it's correct to assume that clients, in this case hospitals, may reuse its own data to try improve model's capacity of generalization.

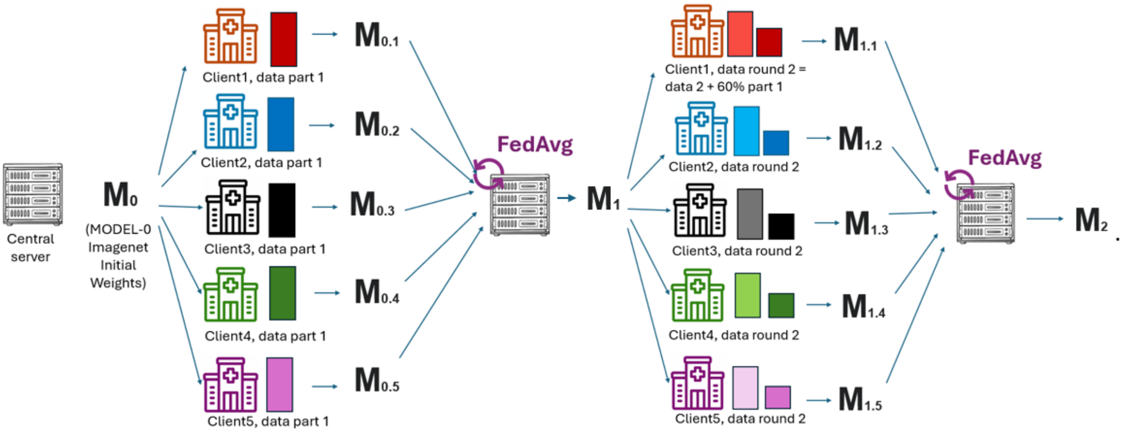


Figure 23. Scheme explanation of case 2 training. Every round, global model is trained with a partition of the data in each client. To this partition is summed a 60% of the images used in the last round.

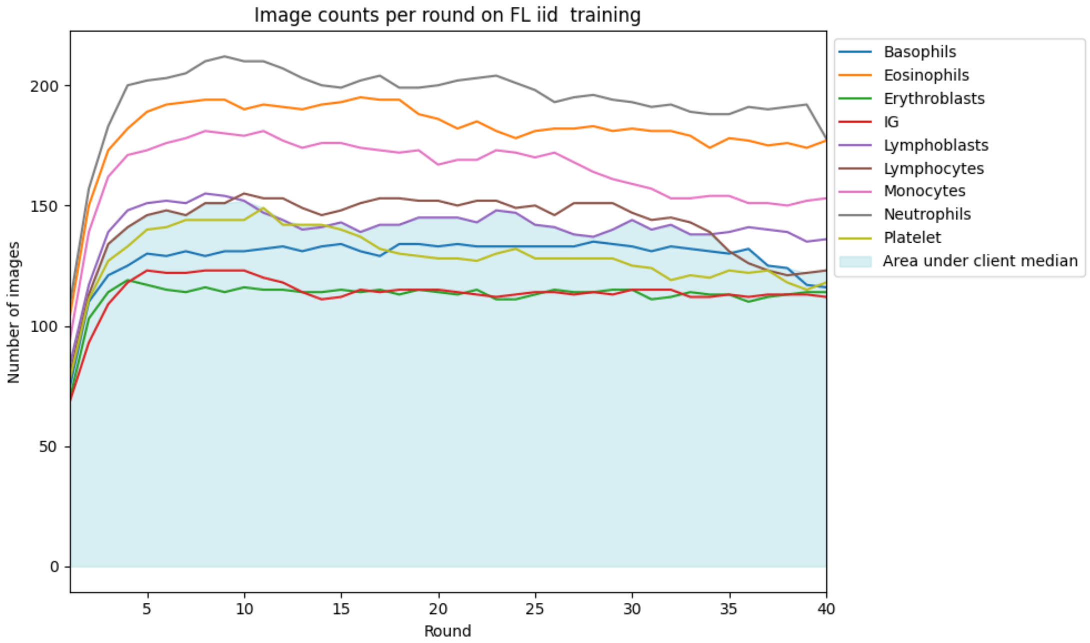


Figure 24. Example of evolution of abundance of each class at the training of local models for each round. As seen, abundance grows until reaching a top.

Validation sample usage in each round

After partitioning train subset of each client into 40 parts, each of the 40 parts are split into train (80%) and validation (20%) of each round. Training data is used as explained above, in two different cases. Validation data is used one two different ways in all the models.

First way, during local training of clients, validation data of each part, for example validation data part 3 from client 3 on round 3 training of client 3, is used to validate the model during local training (Figure 25). A median among the 5 clients of these validation metrics is calculated to study the learning of the models.

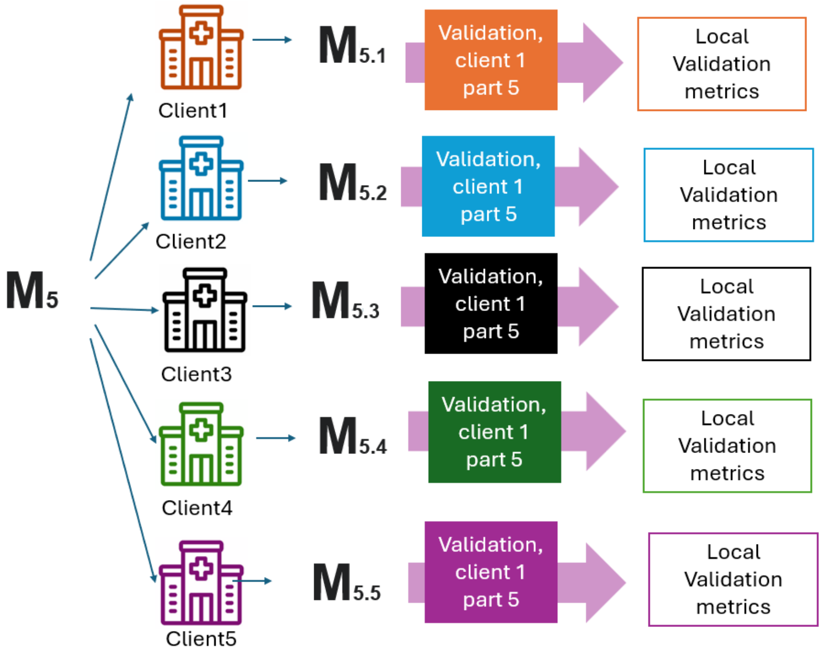


Figure 25. Local Training Validation Example Scheme

Second way, after the aggregation of the 5 models into one global model, this global model is evaluated client by client with all client's validation data of every round until the actual round. Then, the median of all 5 global validation accuracies is calculated (Figure 26). This is done to simulate a real scenario in which global server doesn't have access to client's data, but every client can send the encrypted validation accuracy to the server.

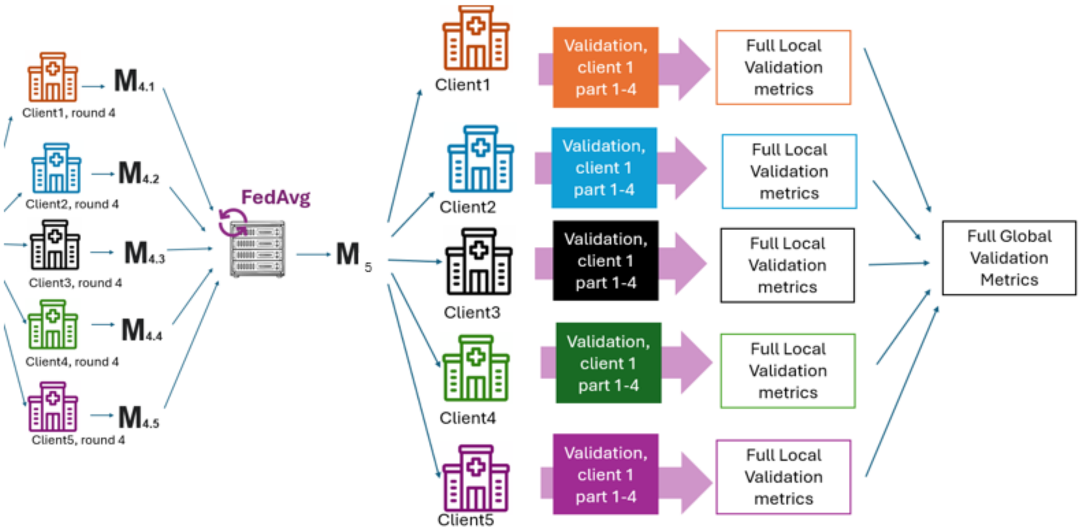

Client1, round 4
Client2, round 4
Client3, round 4
Client4, round 4
Client5, round 4
M4.1
M4.2
M4.3
M4.4
M4.5
FedAvg
M 5
Client1
Client2
Client3
Client4
Client5
Validation, client 1 part 1-4
Validation, client 1 part 1-4
Validation, client 1 part 1-4
Validation, client 1 part 1-4
Validation, client 1 part 1-4
Full Local Validation metrics
Full Local Validation metrics
Full Local Validation metrics
Full Local Validation metrics
Full Local Validation metrics
Full Global Validation Metrics

Figure 26. Global Validation Example Scheme.

Client participation in each round

As this study aims to simulate a real case, in normal conditions all clients should be training new data and sending parameters to the server every week.

3.6 Model Evaluation and Metrics

Evaluation of models is always performed with the fine-tunned model that performed better on the training validation.

Prediction and analysis

Using the best performing model, test dataset images classes are predicted. Then, a dataframe is created (Table 12), containing:

File Name: Name of the image which class is being predicted
Top 1 Predicted Class and Score.
Top 2 Predicted Class and Score.
True Class.

File_name	Predicted Class	Score1	2nd Class	Score2	True Class
106439.jpg	basophil	1.0	ig	7,90E-09	basophil
107517.jpg	basophil	0.99999964	ig	3,59E-01	basophil
113396.jpg	ig	0.67754223	basophil	0.313218419	basophil
116467.jpg	basophil	1.0	ig	4,49E-03	basophil
130237.jpg	basophil	0.9970361	ig	0.002323971	basophil

Table 12. Example of the dataset created after prediction of test images' class.

From this dataset, a lot of analysis and metrics can be obtained, as explained below..

Confusion Matrix

A confusion matrix is a table that summarizes the performance of a classification model. It provides a detailed breakdown of the model's predictions compared to the actual target values. The confusion matrix has the following key elements:

- True Positive (TP): The model correctly predicted the positive class.
- True Negative (TN): The model correctly predicted the negative class.
- False Positive (FP): The model incorrectly predicted the positive class.
- False Negative (FN): The model incorrectly predicted the negative class.

These key elements allow to calculate important evaluation metrics like accuracy, precision, recall, and F1-score Also provides a comprehensive understanding of where the model is making mistakes, which is crucial for improving its performance.

Evaluation Metrics

Accuracy

Definition: Accuracy is the proportion of true positives and true negatives among all predictions. It is calculated as the sum of true positives and true negatives divided by the total number of predictions. It is often represented as a percentage.

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

Precision:

Definition: Precision is the proportion of true positives among all positive predictions. It is calculated as the number of true positives divided by the sum of true positives and false positives.

$$Precision = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP) + False\ Negative\ (FP)}$$

Recall:

Definition: Recall is the proportion of true positives among all actual positive instances. It is calculated as the number of true positives divided by the sum of true positives and false negatives.

$$Recall = \frac{TP}{TP + FN}$$

F1 Score:

Definition: F1 Score is a measure that combines precision and recall. It is the harmonic mean of precision and recall. It is calculated as the harmonic mean of precision and recall. It is often represented as a percentage.

$$F1 = 2\ \times \frac{Precision\ \times\ Recall}{Precision + Recall}$$

Scalability Metrics

Training time

As all models are trained under identical GPU and CPU conditions, the duration of their training serves as a reliable indicator of the requisite computational resources.

Median CPU and GPU usage

The online platform jarvislabs.ai has been utilised for model training. Employing this platform alongside Tensorflow 2.3 has encountered certain issues and drawbacks. It appears that Tensorflow is prone to memory accumulation problems, which particularly impinged upon the simulation of federated learning. Owing to the inability to precisely gauge memory usage throughout the training process, initial CPU memory occupancy values are recorded and averaged. GPU is shown to work at its full capability almost all the time. The used Nvidia RTX5000 has 32GB of graphic memory and 16GB GDDR6 memory.

Parameters Size

One of the key factors that may constrain the feasibility of federated learning is the incessant iteration over the network between clients and the central server. To assess these costs and determine their potential impact on the standard progression of training, an average has been calculated for both the complete model's size and the parameters of the trainable layers alone.

4. Results

4.1 Training Results

General Results

After training one by one the models, the train metrics depicted in Table 13 were obtained where obtained.

N	Model	Scheme	Best Global Validation Acc.	Best Global Validation Loss	Best Median Local Validation Acc.	Training Time
1	ResNet50	Global	0.9123	0.2753	---	5h 13m
2	VGG16	Global	0.9193	0.2336	---	5h 35m
3	ResNet50	FL IID – Case 1	0.9666	0.18100	0.9459	52m
4	ResNet50	FL IID– Case 2	0.9715	0.1520	0.9523	1h 12 m
5	ResNet50	FL non-IID – Case 1	0.4951	2.2516	0.1500	48m
6	ResNet50	FL non-IID – Case 2	0.9629	0.2086	0.9500	1h 7m

Table 13. Model training results of all scenarios.

The model that best performed on training was the FL-IID Case 2 (best val. Acc. = 97, 15%, best median local val. Acc. = 95,23%). All models, except for ResNet50 FL no-IID without data recycling, obtained validation accuracies above 91%, which is a satisfactory result. Except case 3, all FL models obtained at least >5% global validation accuracy. In all cases, training time was hugely different between Global models and FL models, being the later an average of roughly 4 hours (3,997h) faster.

Model	Model size (compiled)	Params size	Trainable params size	Average Physical RAM
VGG16	272MB	58,99 MB	1,04 MB	63,82 GB
ResNet18	168 MB	94,87 MB	2,69 MB	64,11 GB

Table 14. Computational resources use and size of models and model's parameters.

Models were trained using ~64GB of RAM for all the training (Table 14). This value is approximately 50% of full capabilities of the hardware being used. This RAM usage added to the use of high-end GPU like Nvidia RTX500 has to be considered when looking at the time models took to train. Regarding the compiled model size, VGG16 was demonstrated to be a smaller model compared to ResNet50. Nevertheless, this does not significantly affect the RAM usage or the model training time. Given that VGG16 is a smaller model, the weights of its total and trainable parameters, 58.99MB and 1.04MB respectively, were less than those of ResNet50 (Total params size = 94.87MB, Trainable params size = 2.69MB).

As can be observed, there are no federated training results for VGG16. This is due to a Tensorflow error when attempting to federate VGG16 with the code that had been initially prepared for ResNet50. Having already obtained the results for ResNet50, it was considered that, taking into account temporal resources, there was no need to adapt the code to train VGG16 in a federated manner.

ResNet50 FL IID Case 1

As an example of a nice FL training evolution, ResNet50 IID without recycling of data case is shown. Train and validation accuracy evolve as expected in a properly trained DNN model (Figure 27, 28), until reaching the *plateau*

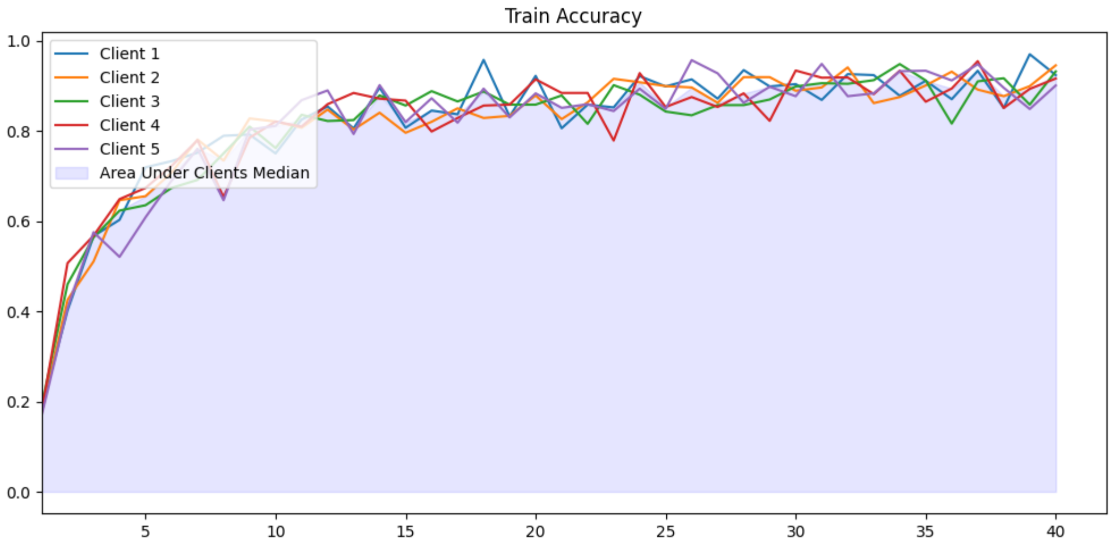


Figure 27. Train Accuracy Evolution of the 5 clients in ResNet50 FL-IID Case 1 and area under median coloured in blue.

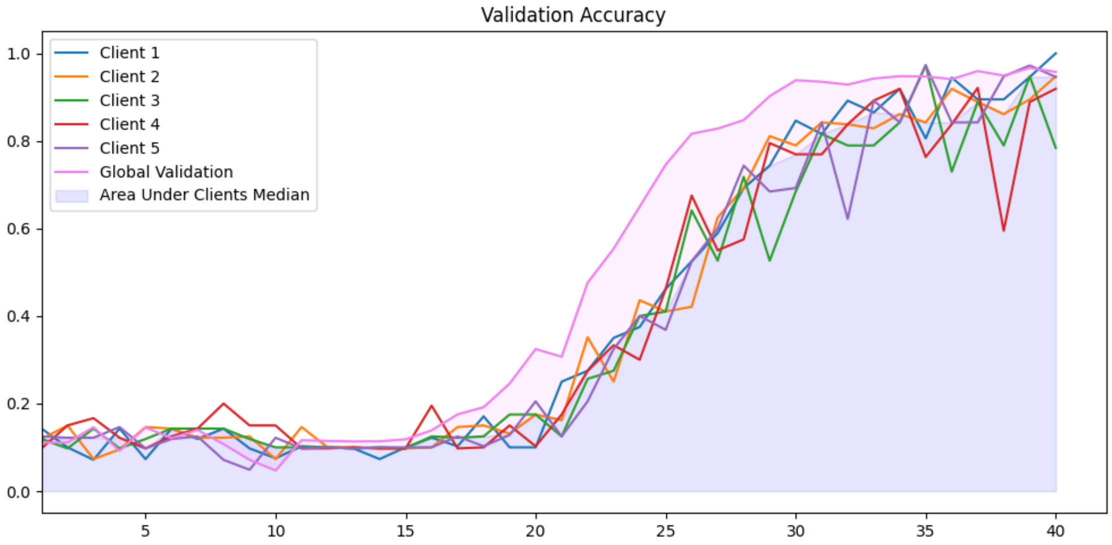


Figure 28. Validation Accuracy Evolution of the 5 clients in ResNet50 FL-IID Case 1 and area under median coloured in blue. Global Validation metric evolution line and the area between its value and median local validation value coloured in violet.

ResNet50 FL no-IID Case 1

In this instance, we have observed a proper evolution of the model's training metrics, achieving a training accuracy exceeding 80% (Figure 29). However, the evaluation metrics have remained exceedingly low (Figure 30), hovering around the values expected from random guessing (1/9 classes = 0.11) throughout the training process until the final round, where an increase in validation accuracy is noted. Yet, results beyond round 40 are not available. These outcomes clearly indicate an issue of overfitting within the model, which has learned intrinsic patterns of the training subset but has not acquired the ability to generalize its learning.

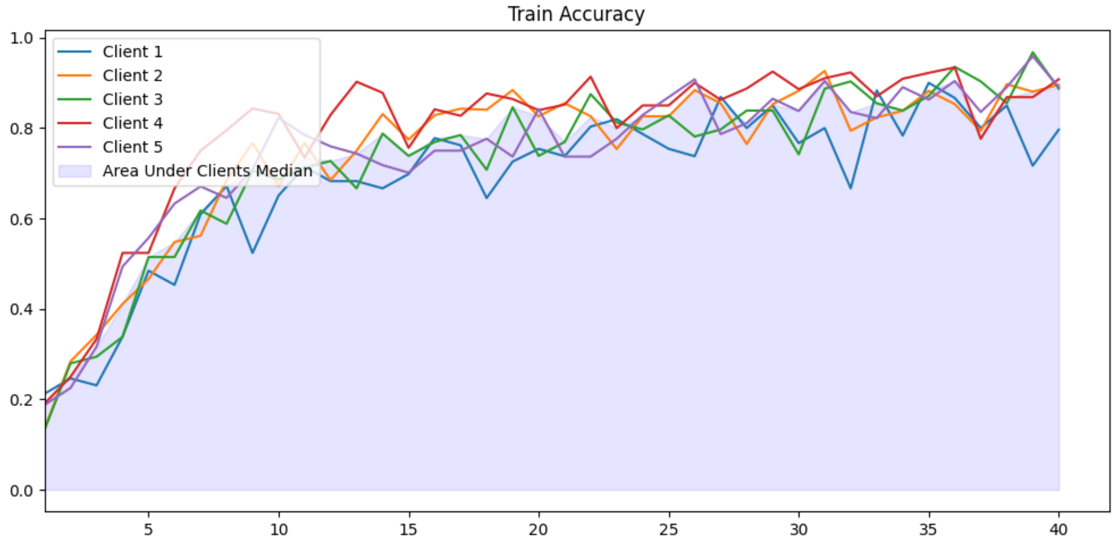


Figure 29. Train Accuracy Evolution of the 5 clients in ResNet50 FL-no IID Case 1 and area under median coloured in blue.

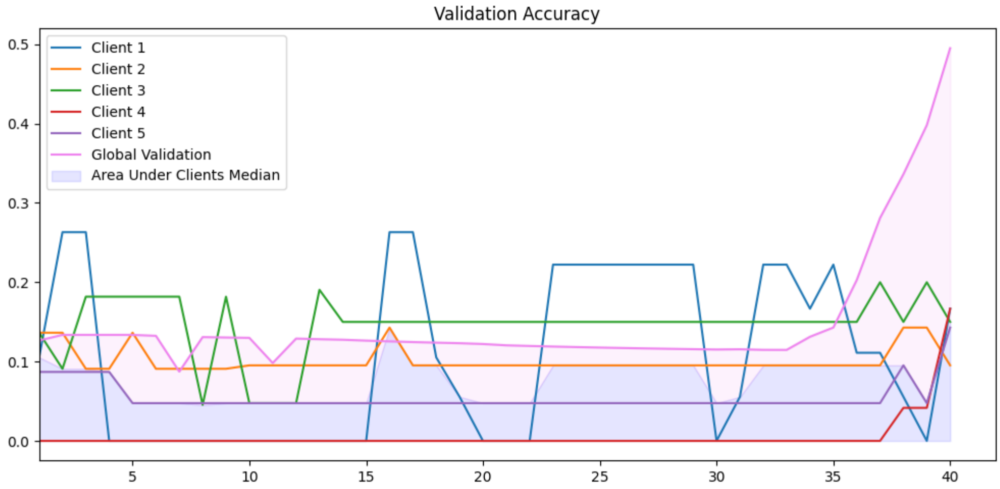


Figure 30. Validation Accuracy Evolution in ResNet50 FL no IID Case 1 of the 5 clients and area under median coloured in blue. Global Validation metric evolution line and the area between its value and median local validation value coloured in violet.

Other Training evolution graphs

The rest of the training evolution graphs can be found in the annexes.

4.2 Test Results

General Table

The evaluation results on Table 15 show that both ResNet50 and VGG16 perform well in global training, with VGG16 slightly outperforming ResNet50. In federated learning with independent and identically distributed (IID) data, ResNet50 achieves high accuracy and F1-score. However, when the data is non-IID, ResNet50's performance drops significantly, indicating challenges in training models with heterogeneous data distributions. Interestingly, when techniques like client selection or data augmentation are used, ResNet50's performance recovers, suggesting that these methods can mitigate the negative impact of non-IID data.

Model	Scheme	Accuracy	Precision	Recall	F1-Score
ResNet50	Global	0.9502	0.9484	0.9477	0.9466
VGG16	Global	0.9620	0.9641	0.9603	0.9617
ResNet50	FL IID – Case 1	0.9700	0.9710	0.9700	0.9703
ResNet50	FL IID– Case 2	0.9757	0.9758	0.9758	0.9757
ResNet50	FL non-IID – Case 1	0.5053	0.7002	0.4933	0.4379
ResNet50	FL non-IID – Case 2	0.9707	0.9705	0.9707	0.9704

Table 15. Evaluation metrics for all the models trained.

ResNet50 FL-IID Case 2

In the case of federated training on IID with data recycling for the ResNet50 model, the model evaluation results have been highly satisfactory. This confusion matrix (Figure 31) provides a more detailed and graphical representation of the excellent Recall, Precision, and F1-Score results achieved by this model.

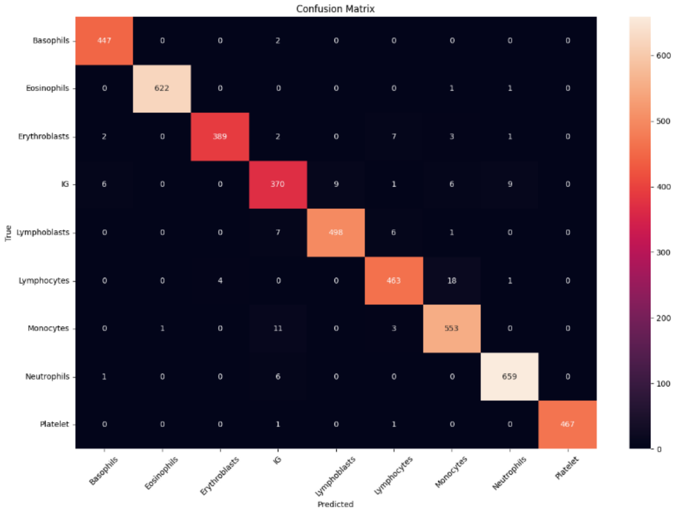


Figure 31. Confusion Matrix Heat-map comparing True Labels (y axis) with Predicted Labels (x-axis) in the evaluation of ResNet50 FL-IID Case 2. White indicates maximum matching among one class from true labels and one class from predicted labels, while black indicates 0 matching.

ResNet50 FL-nonIID Case 1

In the case of federated training on non-IID without data recycling for the ResNet50 model, it seems that being evaluated with the model from the last round, which is the only one exceeding 20% validation accuracy, has been reflected in the model's evaluation. Although the results remain quite poor (Figure 32), they may suggest that with additional training rounds, better values might have been achieved, but this is nothing more than mere speculation.

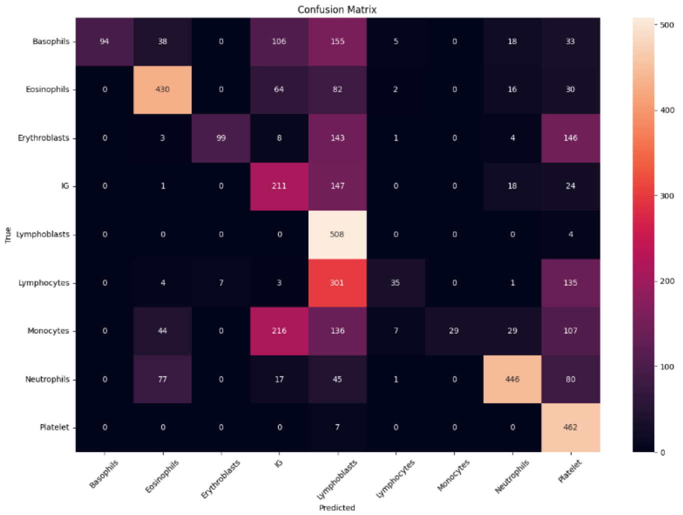


Figure 32. Confusion Matrix Heat-map comparing True Labels (y axis) with Predicted Labels (x-axis) in the evaluation of ResNet50 FL no-IID Case 1. White indicates maximum matching among one class from true labels and one class from predicted labels, while black indicates 0 matching.

4.3 Privacy results

KerasSGD_DP optimizer from the Tensorflow Privacy module has been implemented to train the ResNet50 FL without data recycling. The model required 5 hours and 54 minutes for training, and the training results were as follows:

Model name	Best Global Validation Acc.	Best Global Validation Loss	Best Median Local Validation Acc.	Training Time
DPFL Resnet50 Case 1	0.1495	2.6451	0.1621	5 h 54 m

Table 17. Training results of model ResNet50 FL-IID case 1 using DP optimizer.

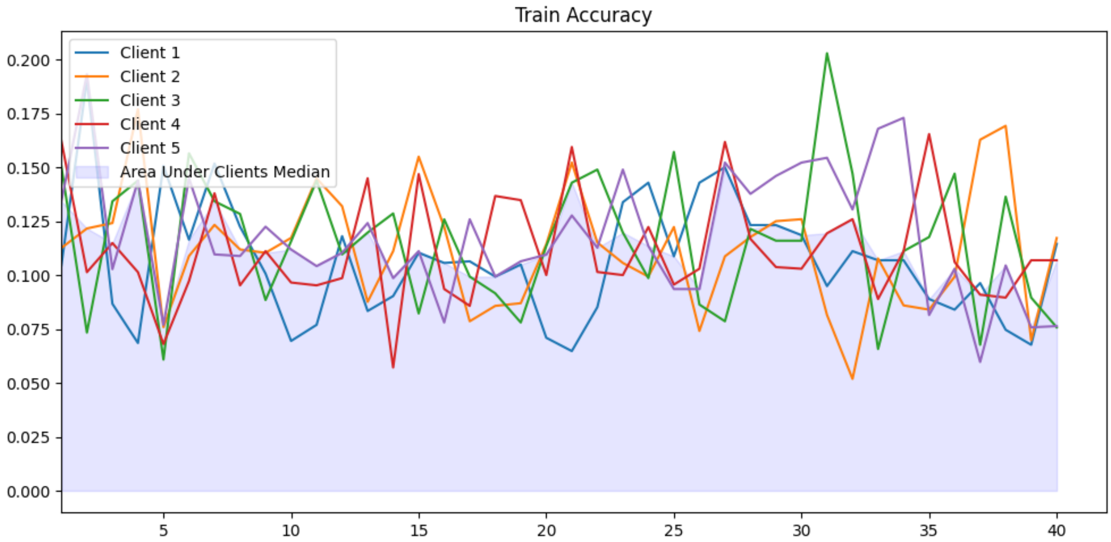


Figure 33. Train Accuracy Evolution of the 5 clients in ResNet50 FL-noIID Case 1 DP and area under median coloured in blue.

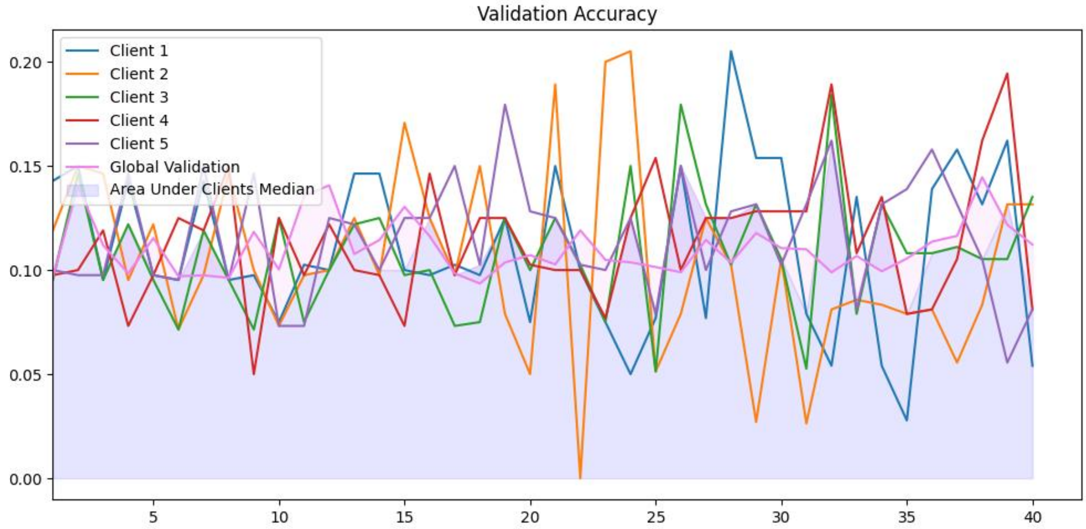


Figure 34. Validation Accuracy Evolution of the 5 clients in ResNet50 FL-IID Case 1 DP and area under median coloured in blue. Global Validation metric evolution line and the area between its value and median local validation value coloured in violet.

Furthermore, a global privacy budget value of ε=3.52 was achieved, using Gradient Clipping of 1,3 and noise multiplier of 4. The training evolution graph appears to indicate that the model is not learning to classify the images accurately, as the displayed accuracy is close to what one would expect from a random classification of the images (1 total probability / 9 classes = 0.11).

5. Conclusions

5.1 Federated Learning may increase model generalization capability

The evaluation metrics for the models, both during training and assessment, suggest that collaborative training generally exhibits superior learning and predictive capabilities in the scenarios examined. In global training, VGG16 and ResNet50 achieved maximum validation accuracies of 91.93% and 91.23% respectively, compared to 97.15% and 96.29% achieve by ResNet50 in collaborative training using data case 2, under IID and non-IID data distributions respectively.

These metrics converge more closely in the model evaluation with the test subset. The global models of ResNet50 and VGG16 attained accuracies of 95.02% and 96.20% respectively, compared to 97.57% and 97.07% achieve in the collaborative training of ResNet50, using data case 2, under IID and non-IID distributions respectively. As for the rest of the validation metrics, the progression is as anticipated; models with higher evaluation accuracy also garner the best results in F1-Score, Recall, and Precision.

In all instances, barring the federated training of ResNet50 non-IID with data use case 1, satisfactory results were obtained, demonstrating the proficiency of VGG16 and ResNet50 models in predicting cell types in peripheral blood smear images.

5.2 Federated Learning as a Private Collaboration Tool

Following the training of the federated models, it has become evident that, at least with these datasets, there is no necessity to share images with a central server to train an effective cellular type of classification model. This could already be contributing to data privacy, as the data are not directly shared at any point. Nonetheless, studies such as (Zhu & Han, 2020) have demonstrated that privacy can be compromised through the analysis of weight logs communicated between clients and the server during training iterations. A viable solution would be the application of Differential Privacy to the weights generated by the models, safeguarding them from privacy breaches. The outcomes pertaining to the federated training of ResNet50 without data recycling using Differential Privacy (DP) have been somewhat underwhelming. The training duration has escalated from t=52m without DP to t=5h 56m with DP. This denotes a substantial increment of 5 hours to train the identical model with DP, which, moreover, has not succeeded in accurately classifying the images.

5.3 Federated Learning as a Scalable Collaboration Tool

The global training of the ResNet50 and VGG16 models spanned 5 hours and 13 minutes, and 5 hours and 35 minutes respectively, averaging 5 hours and 24 minutes. Federated training of the ResNet50 model, without data recycling, recorded an average training duration of 50 minutes, while the corresponding case with data recycling averaged 1 hour and 9 minutes. These findings, along with little size of model's trainable parameters (Table 14), underscore the remarkable scalability of collaborative training in expediting the training of computationally intensive image classification models, as demonstrated in this study.

On the one hand, it should be noted that these durations do not account for the time that would be invested in real-world scenarios for data communication and sharing between clients and the central server. Nonetheless, such communication would not significantly impact the GPU or CPU resources of the nodes, as it primarily involves data upload and download processes.

On the other hand, when examining the Federated Learning (FL) training durations, it is imperative to acknowledge that it reflects the time taken by a single machine to train all clients across all rounds. While it is true that the computational resources utilised, particularly the graphics card, may not be readily available in smaller medical centres, one must consider that in a real-world scenario, the workload conducted by the hardware in this study would be distributed among the hardware of all participating nodes. Consequently, the findings suggest that in this instance, FL could be a viable paradigm for the scalability of Deep Neural Network (DNN) models.

5.4 Data Usage during Federated Learning

As previously noted, all models have achieved satisfactory predictive capabilities, except for the federated training of ResNet50 with non-IID data distribution and without data recycling, which attained a maximum training accuracy of 49.51% and an evaluation accuracy of 50.53%. The analogous training with data recycling achieved a training accuracy of 96.29% and an evaluation accuracy of 97.07%. This suggests that in cases of non-IID data distribution, which more closely resembles real-world scenarios, data recycling may be crucial for the effective training of models.

Conversely, in the federated training of ResNet50 with IID data, the outcomes without data recycling (training accuracy = 96.66%, test accuracy = 97%) and with data recycling (training accuracy = 97.15%, test accuracy = 97.57%) are remarkably similar. This indicates that in this scenario, where all clients are identical, data recycling may not be necessary, as it is inherently occurring when training five clients with the same data distribution in each round.

5.5 Objectives achieving

“To develop and evaluate a Federated Deep Learning tool for the detection and classification of cellular types on peripheral blood samples.”

The federated training of the classification models has demonstrated a satisfactory performance on training as well as on test ecaluation.

“The developed tool should be scalable and suitable (low requirements) for hospital’s computers and connection.”

While the hardware used to train the models is not the lowest requirements, it’s still not high tier technology, and it’s not difficult to assume some medium and large sized health centres may have access to similar hardware. In the case it isn’t like that, Federated Learning has proven to need much less time to learn. Although more specific work should be done, this work is positive on thinking that FL models, in this case, can be trained without GPU.

“To conceive a data privacy study/blueprint for the developed AI tool.”

This objective has not been fully achieved. While research has been conducted into potential privacy issues in Federated Learning and attempts have been made to find solutions, there has not been sufficient time to conduct a satisfactory analysis of the topic. The objective has not been met due to a lack of time and organization.

5.6 Planning analysis

The project’s planning has not been adhered to rigorously since the submission of PEC1, with a consistent delay of approximately 2-3 weeks. The primary causes of this delay have been attributed to a lack of organisation and initial challenges in securing access to powerful GPUs for model training. Furthermore, due to constraints related to space, compatibility, and features, TensorFlow Federated was not utilised for federated model training. Consequently, a Python codebase was developed to simulate this environment.

5.7 Socio-ethical Impact

Models have been developed that predict the cellular type of blood cells from individual cell images. This tool has the potential to become a valuable asset in the future, following further refinement, to assist in routine cellular counts and the identification of abnormal cells. These applications could significantly contribute to enhancing physiological health.

In terms of privacy, a framework has been established that enables collaborative training with medical data without the need to directly share these images with the cloud. Although local data training ensures that data does not leave its node, it is known that training data can be inferred from the gradient of parameters during model updates. Therefore, the achievement of privacy impact is only partial.

Regarding job displacement, it has been reaffirmed that no machine learning model is flawless, and professional expertise will always be essential for validation.

5.8 Future work

In the future, the most critical area for development, in my opinion, is privacy. At least in cellular type classification, federated learning has demonstrated learning capabilities equal to or surpassing conventional training. Therefore, the focus should now shift to ensuring that federated training upholds data privacy. For future endeavours, I would recommend exploring Pytorch as a machine learning framework, as TensorFlow has exhibited certain compatibility issues as well as memory usage concerns. Moreover, Pytorch is equipped with more established tools for privacy studies. A prospective objective would be to conduct a privacy analysis of the developed Federated Learning (FL) models, utilising Differential Privacy (DP) optimisers and examining how the epsilon value influences the training and performance of the model.

Additionally, the evaluation results obtained from federated training with non-IID data distribution and data recycling are quite remarkable. Typically, these models are expected to perform less well than their IID or globally trained counterparts. It would be intriguing to continue investigating this area to determine the full potential of non-IID training.

Discussing the applicability of science is always beneficial. The outcomes of this research are quite satisfactory, yet research does not directly impact society until it is transformed into technology. Thus, it would be prudent to explore the real-world applicability of the federated learning models developed in this study.

6 Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2016). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. http://arxiv.org/abs/1603.04467

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep Learning with Differential Privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318. https://doi.org/10.1145/2976749.2978318

Acevedo, A., Merino, A., Alférez, S., Molina, Á., Boldú, L., & Rodellar, J. (2020). A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief*, *30*. https://doi.org/10.1016/j.dib.2020.105474

Adnan, M., Kalra, S., Cresswell, J. C., Taylor, G. W., & Tizhoosh, H. R. (2022). Federated learning and differential privacy for medical image analysis. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-05539-7

Andreux, M., du Terrail, J. O., Beguier, C., & Tramel, E. W. (2020). Siloed Federated Learning for Multi-centric Histopathology Datasets. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 12444 LNCS*. https://doi.org/10.1007/978-3-030-60548-3_13

Baid, U., Pati, S., Kurc, T. M., Gupta, R., Bremer, E., Abousamra, S., Thakur, S. P., Saltz, J. H., & Bakas, S. (2022). *Federated Learning for the Classification of Tumor Infiltrating Lymphocytes*. http://arxiv.org/abs/2203.16622

Bodzas, A., Kodytek, P., & Zidek, J. (2023). A high-resolution large-scale dataset of pathological and normal white blood cells. *Scientific Data*, *10*(1). https://doi.org/10.1038/s41597-023-02378-7

Brendan McMahan Eider Moore Daniel Ramage Seth Hampson Blaise AgüeraAg, H., & Arcas, A. (2017). *Communication-Efficient Learning of Deep Networks from Decentralized Data*.

Cetinkaya, A. E., Akin, M., & Sagiroglu, S. (2021). Improving Performance of Federated Learning based Medical Image Analysis in Non-IID Settings using Image Augmentation. *14th International Conference on Information Security and Cryptology, ISCTURKEY 2021 - Proceedings*, 69–74. https://doi.org/10.1109/ISCTURKEY53027.2021.9654356

Crosby, D., Bhatia, S., Brindle, K. M., Coussens, L. M., Dive, C., Emberton, M., Esener, S., Fitzgerald, R. C., Gambhir, S. S., Kuhn, P., Rebbeck, T. R., & Balasubramanian, S. (2022). Early detection of cancer. In *Science* (Vol. 375, Issue 6586). American Association for the Advancement of Science. https://doi.org/10.1126/science.aay9040

Dwork, C. (2006). *Differential Privacy* (pp. 1–12). https://doi.org/10.1007/11787006_1

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition*. http://arxiv.org/abs/1512.03385

Hosseini, S. M., Sikaroudi, M., Babaie, M., & Tizhoosh, H. R. (2023). Proportionally Fair Hospital Collaborations in Federated Learning of Histopathology Images. *IEEE Transactions on Medical Imaging*, *42*(7), 1982–1995. https://doi.org/10.1109/TMI.2023.3234450

Hu, Y., & Chaddad, A. (2023). Potential of Federated Learning in Healthcare. *2023 IEEE International Conference on E-Health Networking, Application and Services, Healthcom 2023*, 216–217. https://doi.org/10.1109/Healthcom56612.2023.10472378

Kohsasih, K. L., Zarlis, M., & Hayadi, B. H. (2022). Comparison of CNN Architecture for White Blood Cells Image Classification. *ICOSNIKOM 2022 - 2022 IEEE International Conference of Computer Science and Information Technology: Boundary Free: Preparing Indonesia for Metaverse Society*. https://doi.org/10.1109/ICOSNIKOM56551.2022.10034875

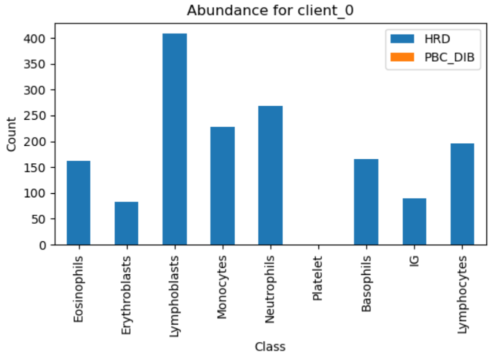
Shen, Y., Sowmya, A., Luo, Y., Liang, X., Shen, D., & Ke, J. (2023). A Federated Learning System for Histopathology Image Analysis With an Orchestral Stain-Normalization GAN. *IEEE Transactions on Medical Imaging*, *42*(7), 1969–1981. https://doi.org/10.1109/TMI.2022.3221724

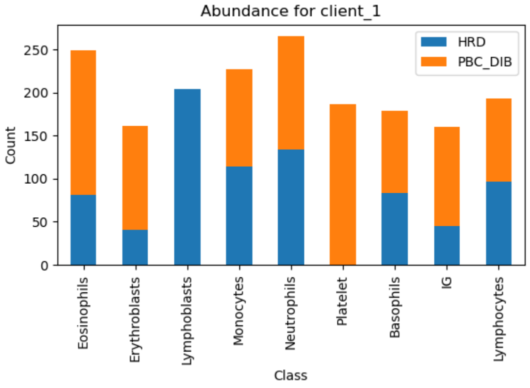
Simonyan, K., & Zisserman, A. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. http://arxiv.org/abs/1409.1556

Temprana-Salvador, J., López-García, P., Vives, J. C., de Haro, L., Ballesta, E., Abusleme, M. R., Arrufat, M., Marques, F., Casas, J. R., Gallego, C., Pons, L., Mate, J. L., Fernández, P. L., López-Bonet, E., Bosch, R., Martínez, S., Cajal, S. R. Y., & Matias-Guiu, X. (2022). DigiPatICS: Digital Pathology Transformation of the Catalan Health Institute Network of 8 Hospitals—Planification, Implementation, and Preliminary Results. *Diagnostics*, *12*(4). https://doi.org/10.3390/diagnostics12040852

Zhu, L., & Han, S. (2020). *Deep Leakage from Gradients* (pp. 17–31). https://doi.org/10.1007/978-3-030-63076-8_2

7 Annexes





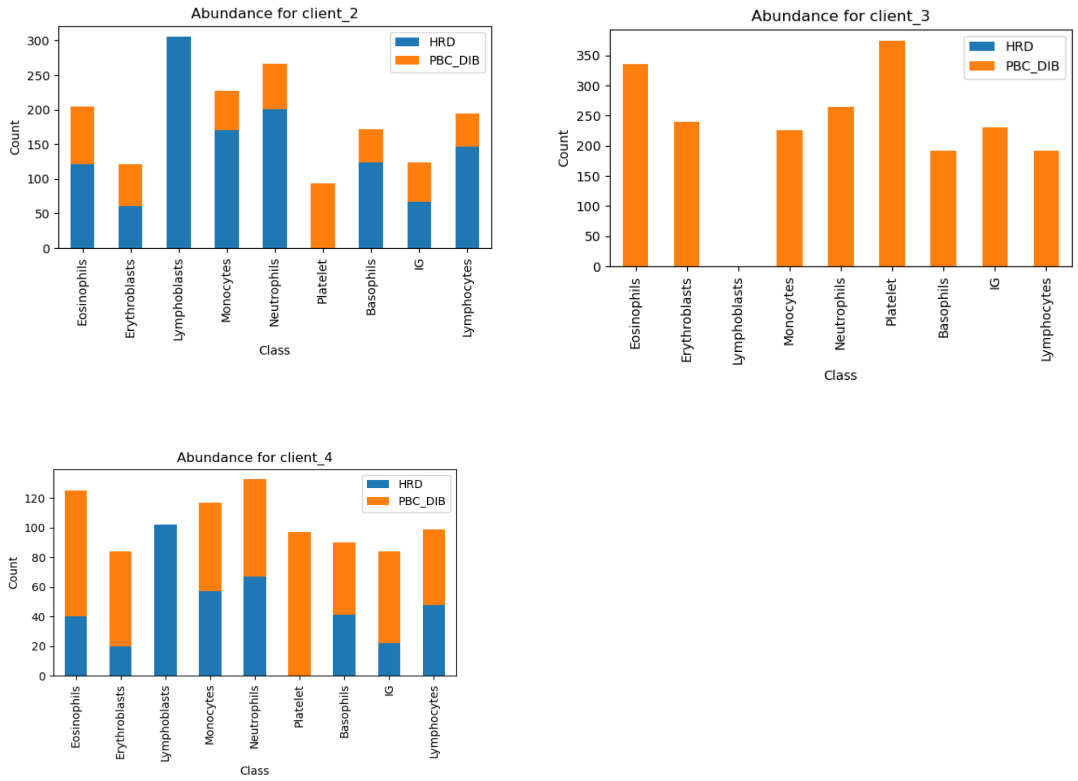


Figure A1. Original dataset property within every class of every client on no-IID distribution.

All Training Plots

FL-IID Case 1

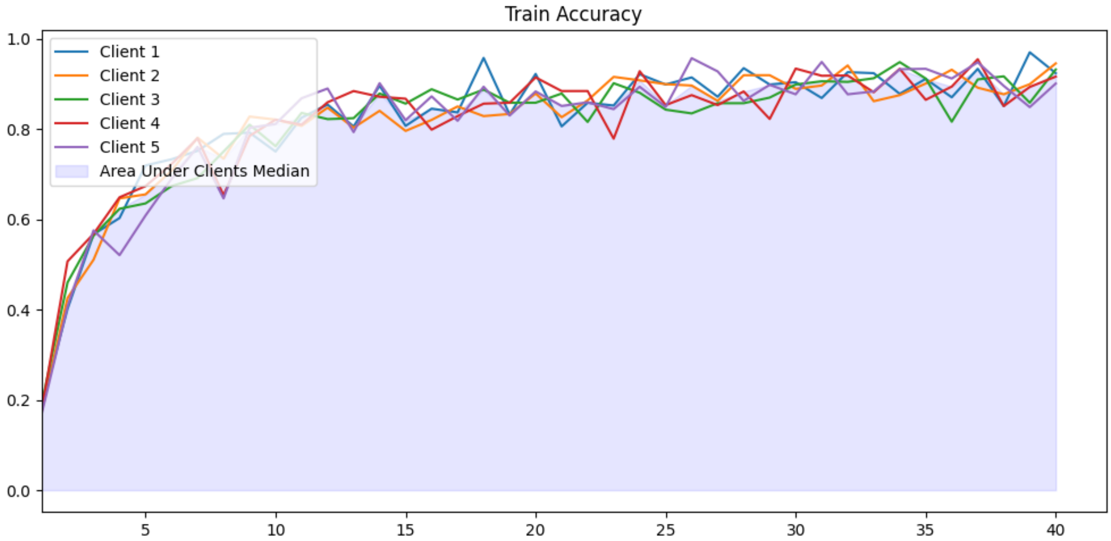


Figure A2. Train Accuracy Evolution of the 5 clients of FL-IID without data recycling and its median.

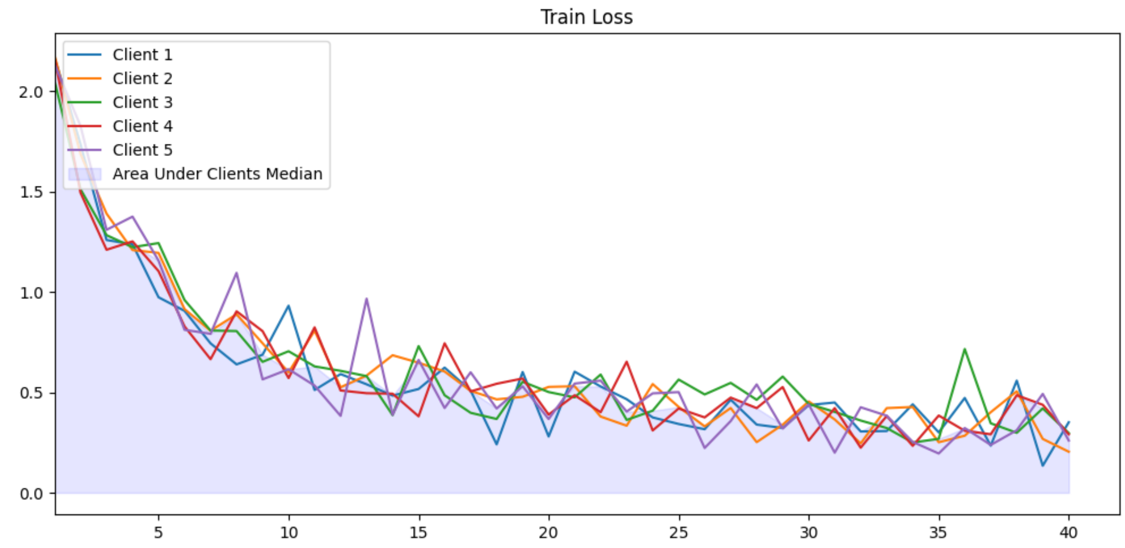


Figure A3. Train Loss Evolution of the 5 clients of FL-IID without data recycling and its median

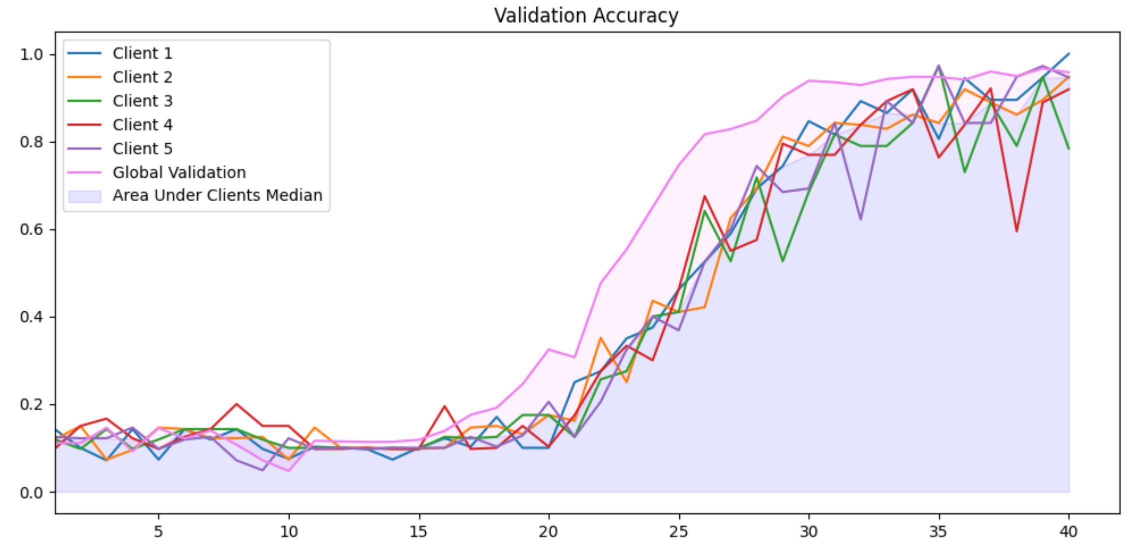


Figure A4. Validation Accuracy Evolution of the 5 clients of FL-IID without data recycling and its median. Also average global accuracy of each round is shown in violet.

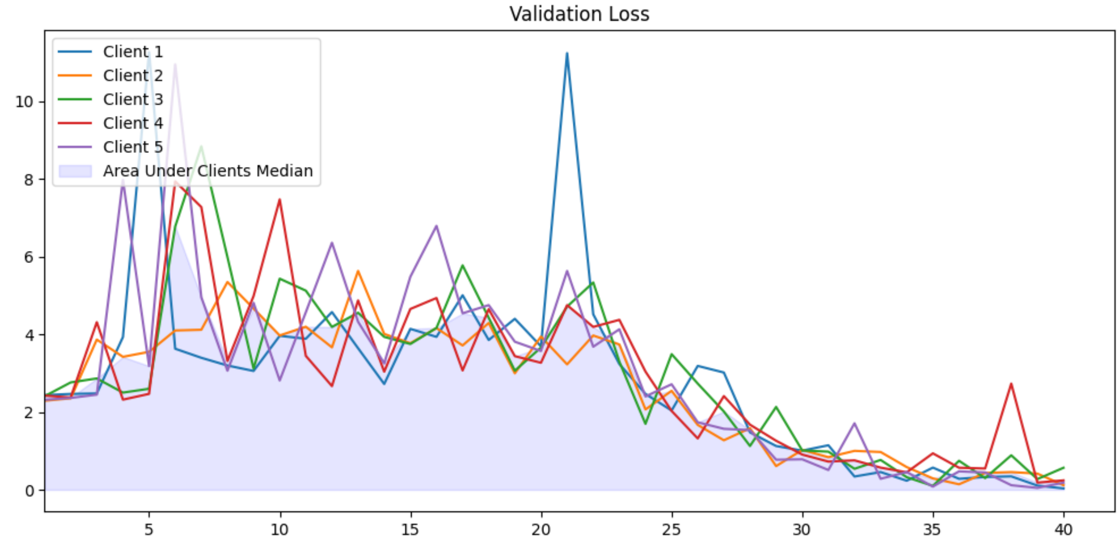


Figure A5. Validation Loss Evolution of the 5 clients of FL-IID without data recycling and its median

FL-IID Case 2

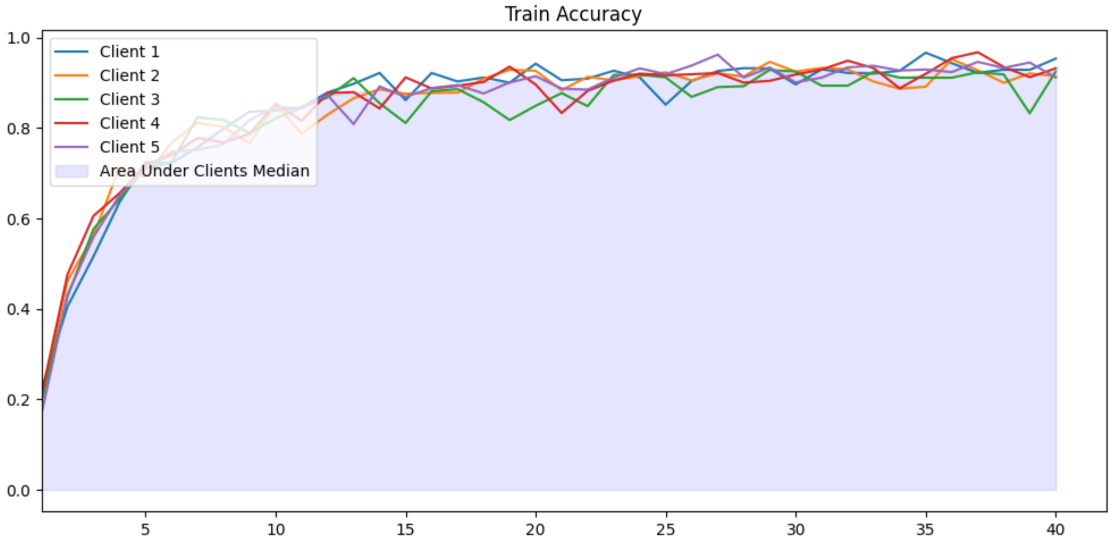


Figure A6. Train Accuracy Evolution of the 5 clients of FL-IID with data recycling and its median

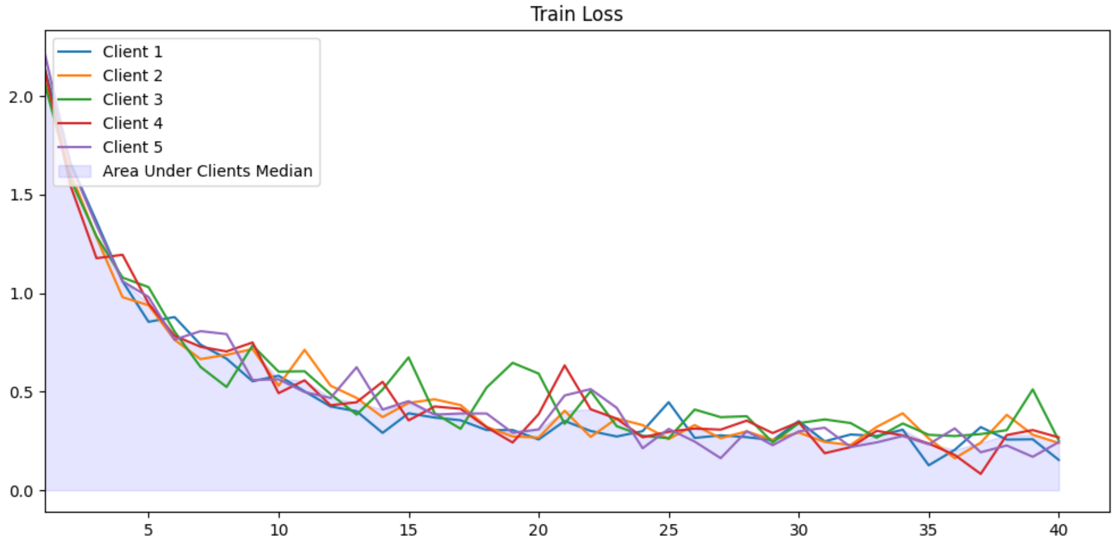


Figure A7. Train Loss Evolution of the 5 clients from FL non-IID and its median

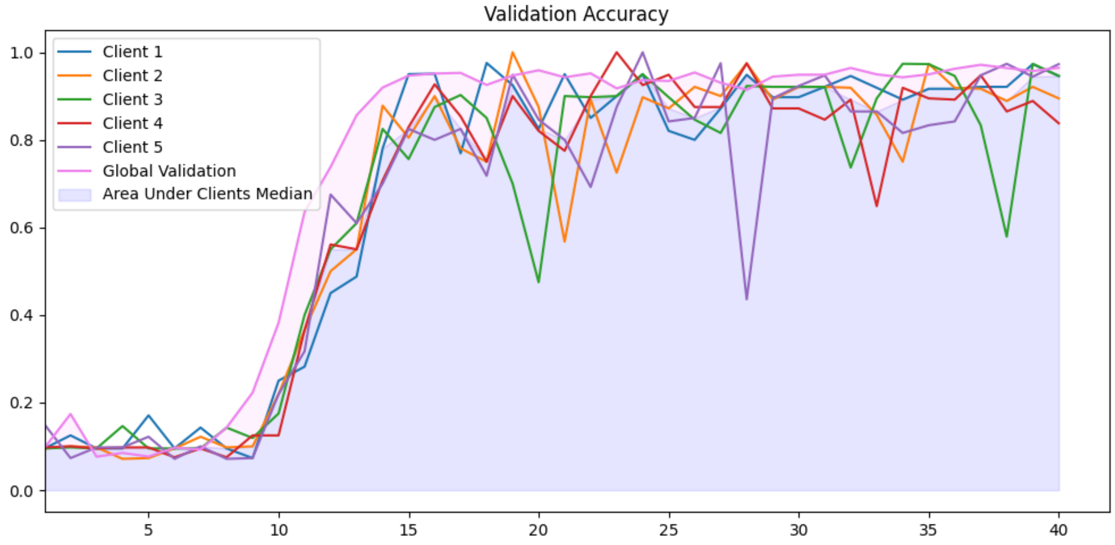


Figure A8. Validation Accuracy Evolution of the 5 clients from FL-IID with data recycling and its median. Also average global accuracy of each round is shown in violet.

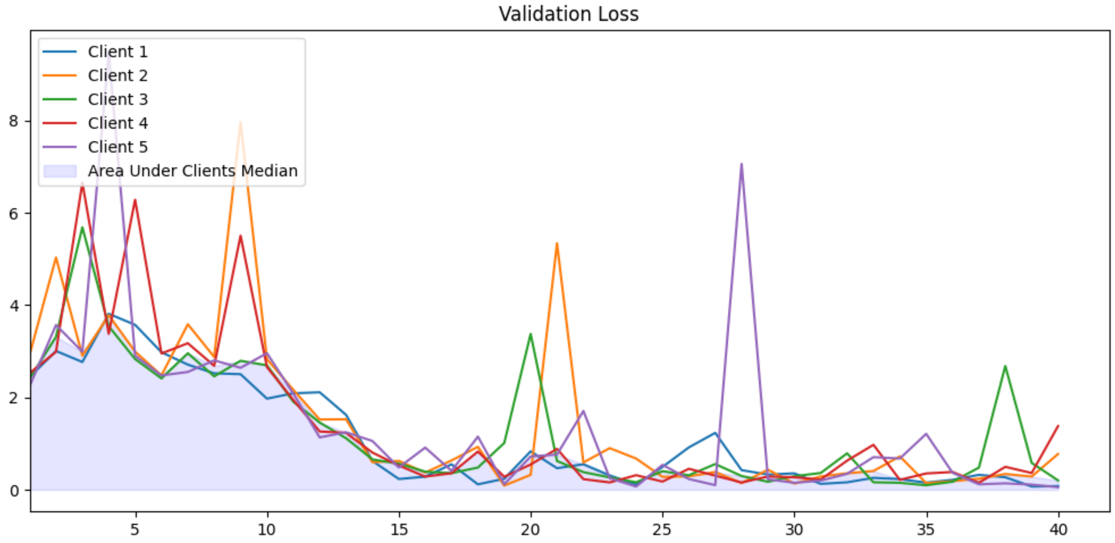


Figure A9. Validation Loss Evolution of the 5 clients from FL-IID with data recycling and its median.

FL non-IID Case 1

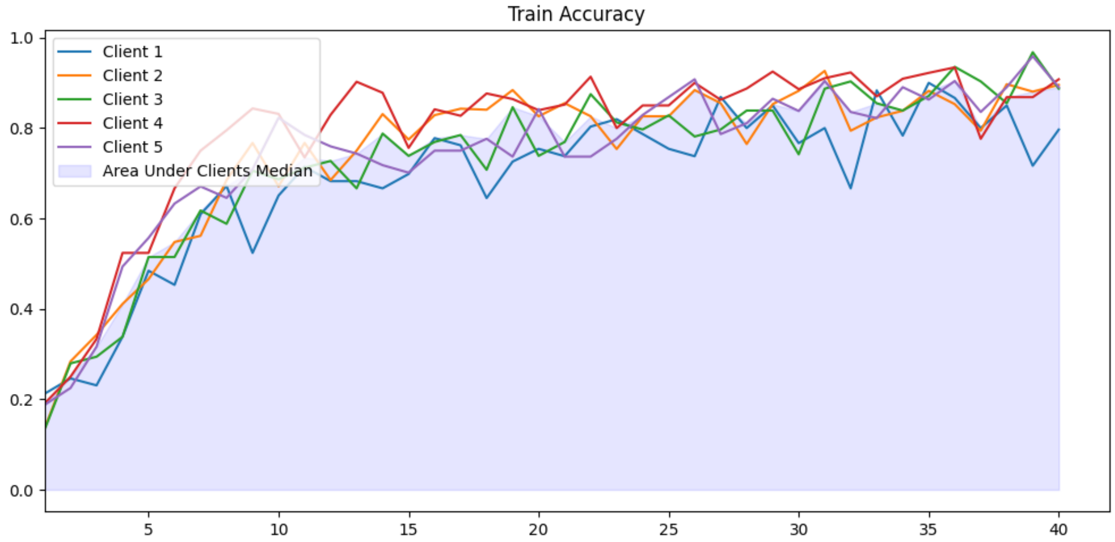


Figure A10 Train Accuracy Evolution of the 5 clients from FL-nonIID without data recycling and its median

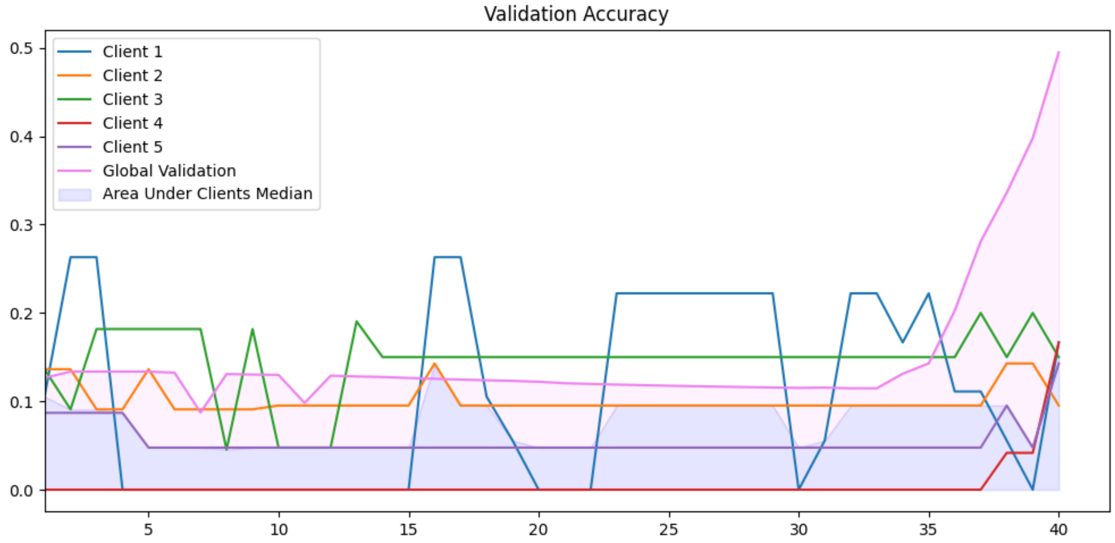


Figure A11. Validation Accuracy Evolution of the 5 clients from FL-noIID without data recycling and its median. Also average global accuracy of each round is shown in violet.

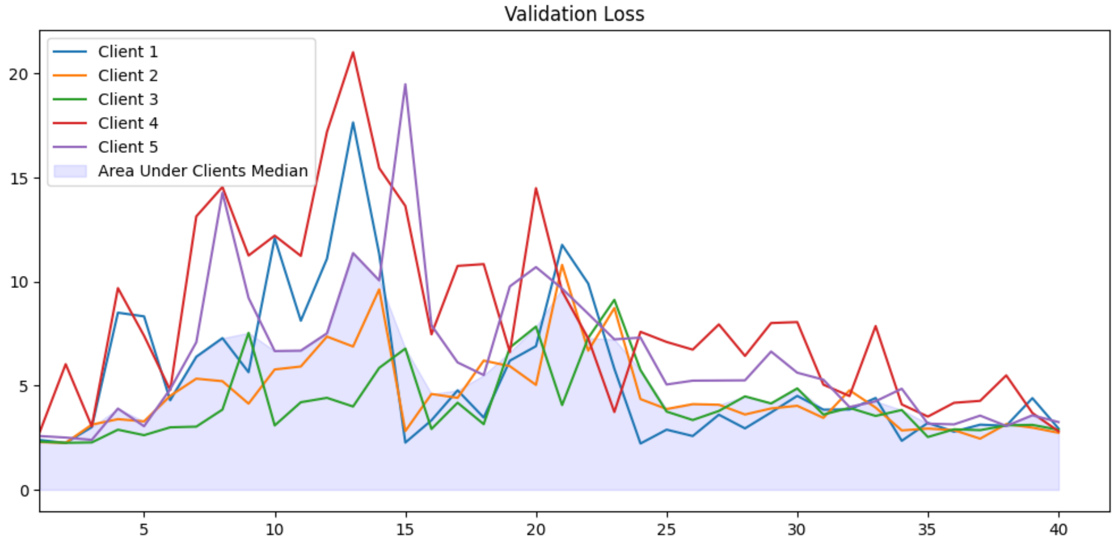


Figure A12. Validation Loss Evolution of the 5 clients frm FL-noIID without data recycling and its median

FL non-IID Case 2

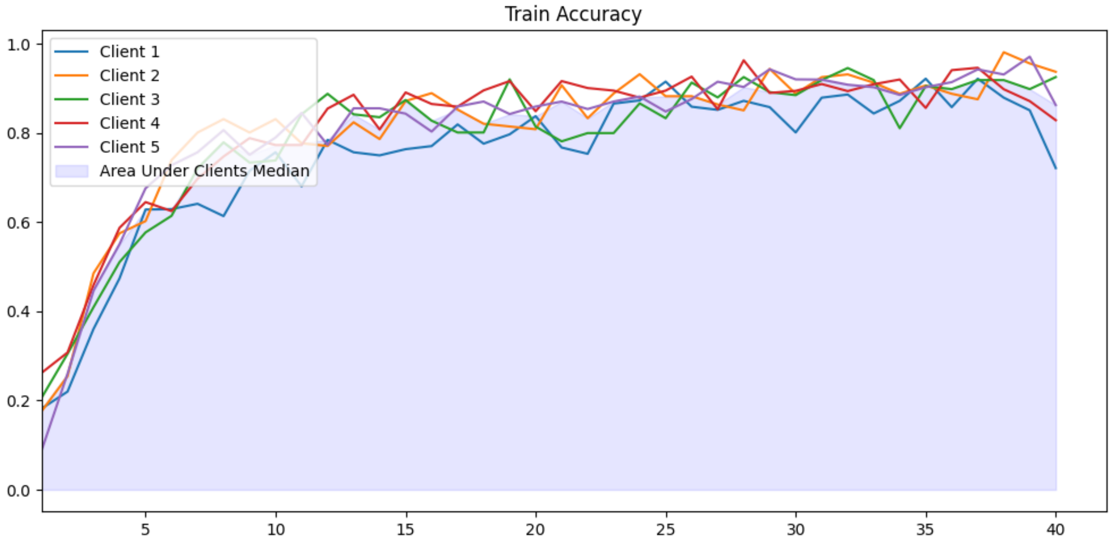


Figure A13. Train Accuracy Evolution of the 5 clients from FL-noIID with data recycling and its median

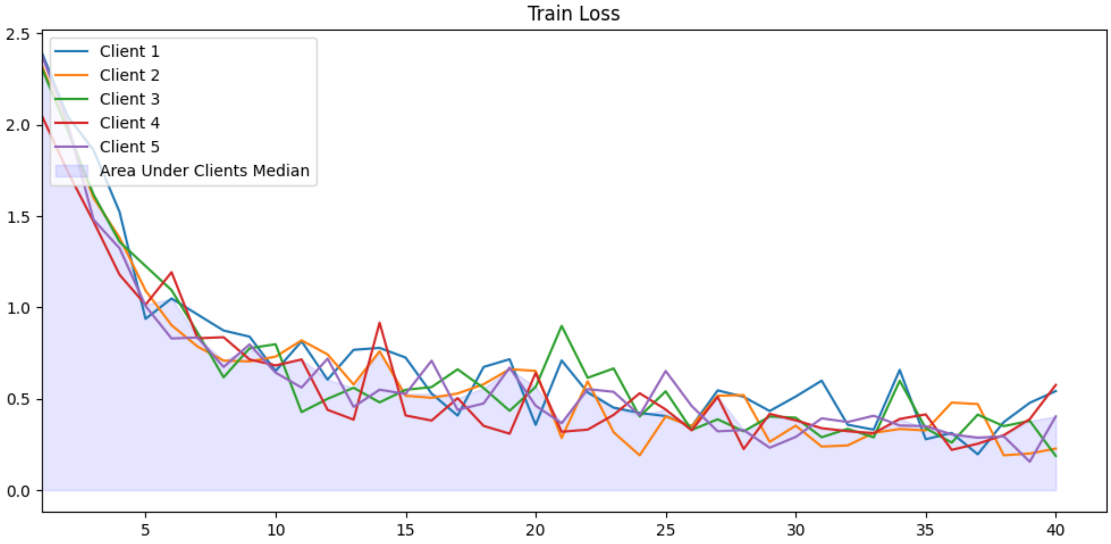


Figure A14. Train Loss Evolution of the 5 clients from FL-noIID with data recycling and its median

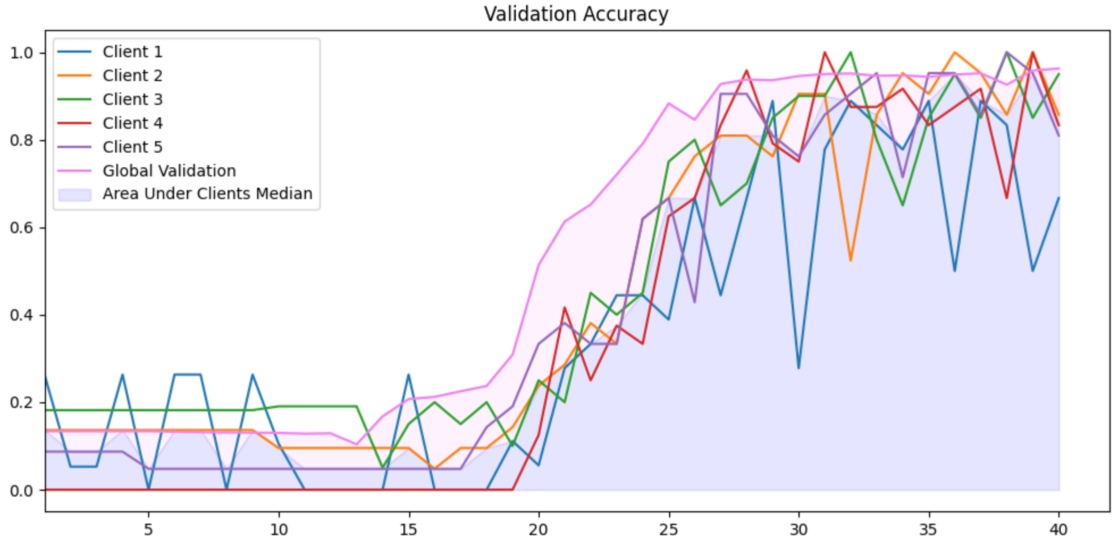


Figure A15. Validation Accuracy Evolution of the 5 clients from FL-noiid with data recycling and its median. Also average global accuracy of each round is shown in violet.

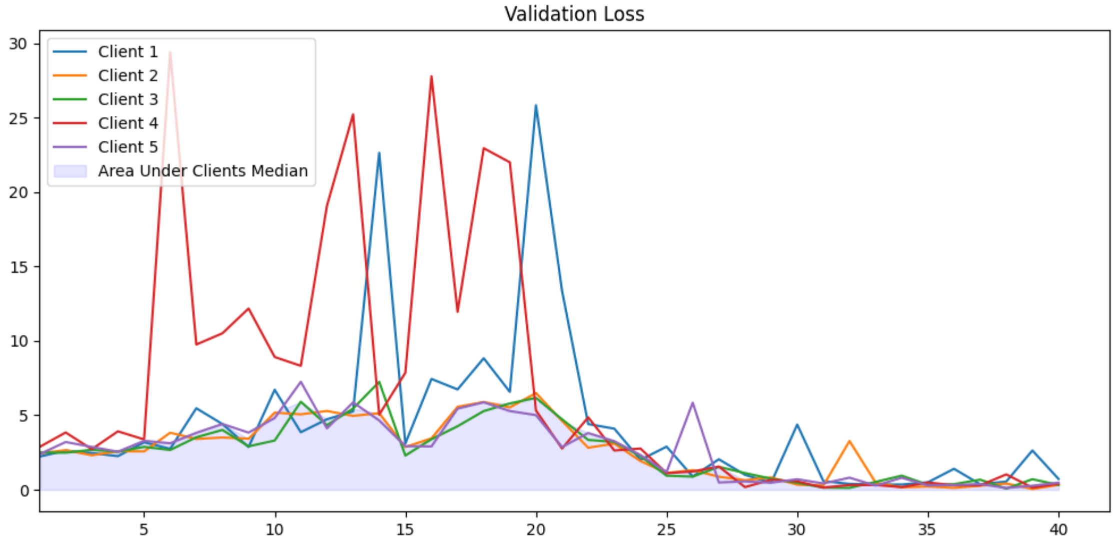


Figure A16. Validation Loss Evolution of the 5 clients from FL-noIID and its median

Test Confusion matrix

ResNet50 Global Training Confusion Matrix

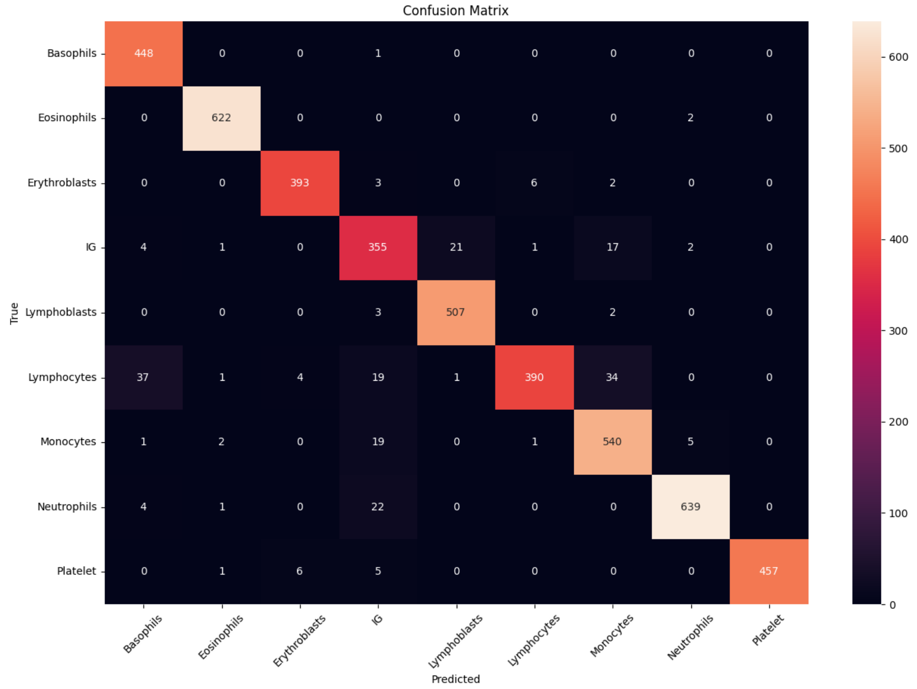


Figure A17. Confusion Matrix Heat-map comparing True Labels(y axis) with Predicted Labels (x-axis) in the evaluation of ResNet50 Global Training. White indicates maximum matching in one class true and predicted labels, while black indicates 0 matching.

VGG16 Confusion Matrix

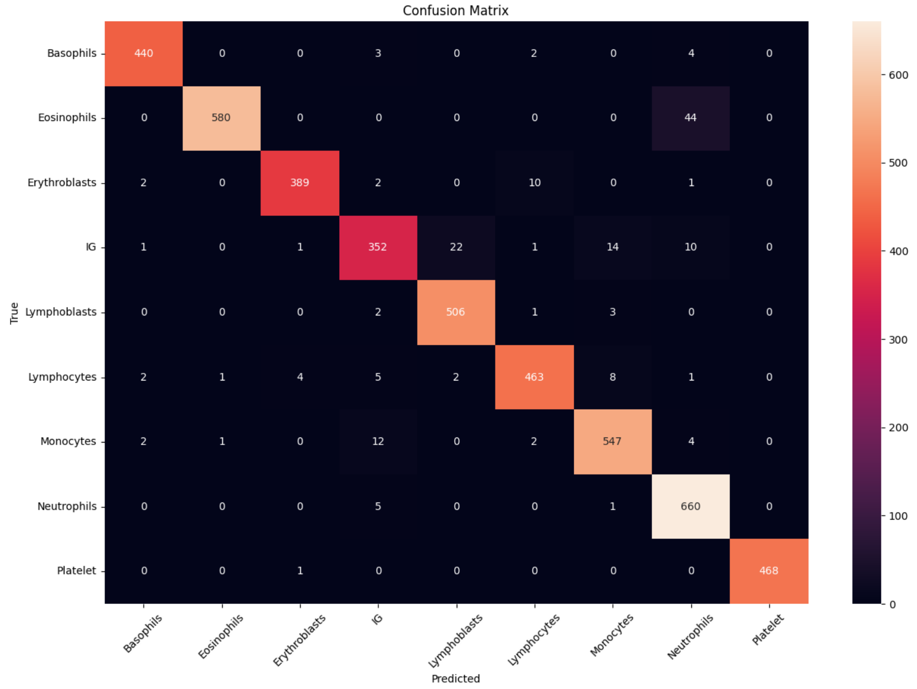
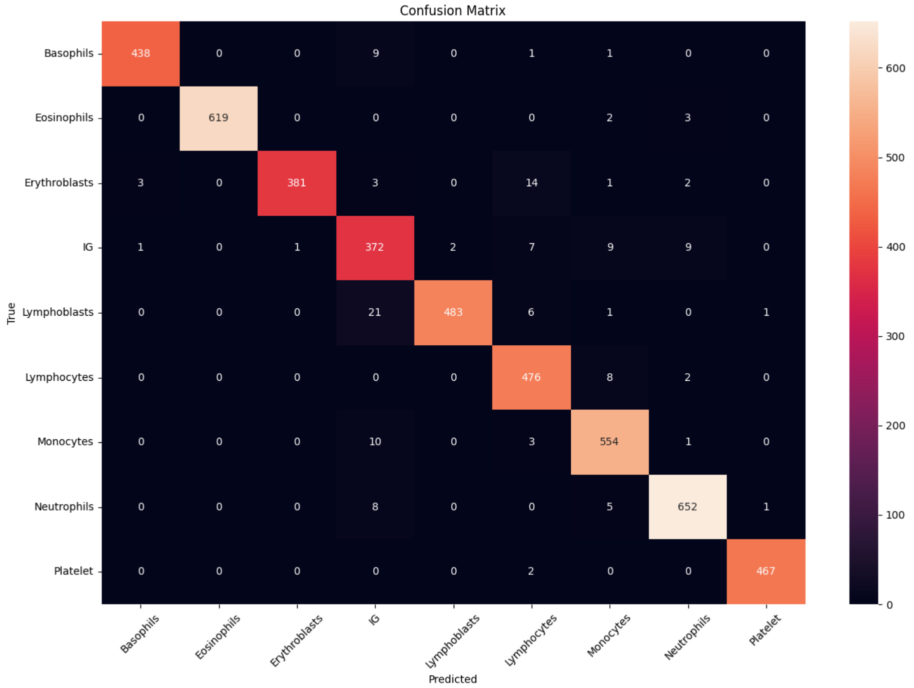


Figure A18. Confusion Matrix Heat-map comparing True Labels(y axis) with Predicted Labels (x-axis) in the evaluation of VGG16 global training. White indicates maximum matching in one class true and predicted labels, while black indicates 0 matching.

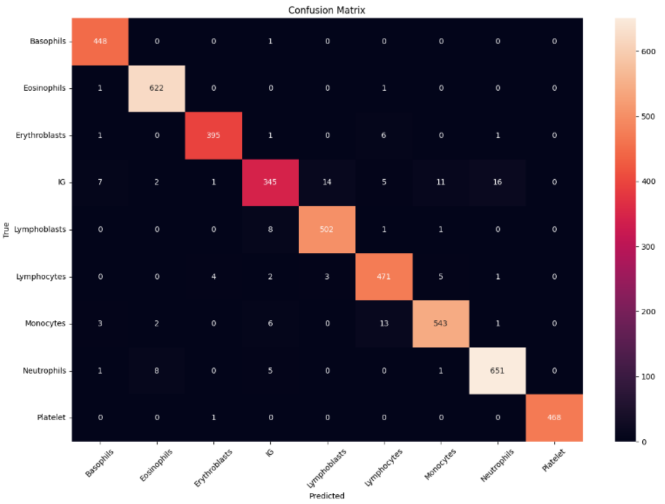
ResNet50 FL-IID Case 1



	Basophils	Eosinophils	Erythroblasts	IG	Lymphoblasts	Lymphocytes	Monocytes	Neutrophils	Platelet
Basophils	438	0	0	9	0	1	1	0	0
Eosinophils	0	619	0	0	0	0	2	3	0
Erythroblasts	3	0	381	3	0	14	1	2	0
IG	1	0	1	372	2	7	9	9	0
Lymphoblasts	0	0	0	21	483	6	1	0	1
Lymphocytes	0	0	0	0	0	476	8	2	0
Monocytes	0	0	0	10	0	3	554	1	0
Neutrophils	0	0	0	8	0	0	5	652	1
Platelet	0	0	0	0	0	2	0	0	467

Figure A19. Confusion Matrix Heat-map comparing True Labels(y axis) with Predicted Labels (x-axis) in the evaluation of ResNet50 FL IID Case 1. White indicates maximum matching in one class true and predicted labels, while black indicates 0 matching.

ResNet50 FL-nonIID Case 2



	Basophils	Eosinophils	Erythroblasts	IG	Lymphoblasts	Lymphocytes	Monocytes	Neutrophils	Platelet
Basophils	448	0	0	1	0	0	0	0	0
Eosinophils	1	622	0	0	0	1	0	0	0
Erythroblasts	1	0	395	1	0	6	0	1	0
IG	7	2	1	345	14	5	11	16	0
Lymphoblasts	0	0	0	8	502	1	1	0	0
Lymphocytes	0	0	4	2	3	471	5	1	0
Monocytes	3	2	0	6	0	13	543	1	0
Neutrophils	1	8	0	5	0	0	1	651	0
Platelet	0	0	1	0	0	0	0	0	468

Figure A20. Confusion Matrix Heat-map comparing True Labels(y axis) with Predicted Labels (x-axis) in the evaluation of ResNet50 FL nonIID Case 2. White indicates maximum matching in one class true and predicted labels, while black indicates 0 matching.