MEDICAL IMAGE ANALYSIS

Privacy-Preserving Medical Image Diagnosis using Federated Learning and Vision Transformers

Ishita Awasthi, Anuj Sharma, Anurag Sharma, Ankita Sharma

Department of Computer Science and Engineering, Chandigarh University, Mohali, Punjab, India

Published November 2026

Abstract

Deep learning model training is a central approach, which poses risks to patient data privacy and hampers data sharing for medical image classification, essential for computer-assisted diagnosis. In this study, a privacy-preserving medical image classification system is presented that combines Vision Transformers (ViTs) with Federated Learning (FL) for the classification of four categories of white blood cells: eosinophils, lymphocytes, monocytes, and neutrophils. The framework uses model training on local data in three simulated clients and uses Federated Averaging (FedAvg) to merge model updates without sharing raw images. A conventional Convolutional Neural Network (CNN) and a centralized ViT are used as baseline models. The preliminary results indicate that the centralized ViT can attain an accuracy of 58.50% and a macro-F1 score of 51.82%, which is better than the CNN baseline (53.50% accuracy). The overall performance of the federated ViT can be seen as good with an accuracy score of 49.50% and a recall score of 96.00% for the class of monocytes, which suggests good class-specific recognition and, simultaneously, room for improvement at the overall level. An additional enhancement that can help prevent information leakage from model updates is proposed as differential privacy. Additional rounds of communication, larger sets of training data, hyperparameter optimization and formal guarantees of privacy will be explored in further work. This is a framework to give a basis for co-operative privacy-aware medical image classification.

Keywords: Federated learning; Vision Transformer; medical image classification; white blood cell classification; FedAvg; differential privacy; privacy-preserving machine learning

1 Introduction

1.1 Background

As a component of computer-assisted diagnosis, medical image classification is a process that aims at detecting clinically relevant patterns within images captured at the micro-level. Vision Transformers (ViTs) rely on self-attention to extract relationships between image patches [11–14], whereas conventional Convolutional Neural Networks (CNNs) are good at visual feature extraction. Centralized medical image analysis, however, requires gathering sensitive data in a single place, thus posing privacy and data-sharing concerns. Federated Learning (FL) overcomes these challenges by enabling joint training of models on clients in a distributed network without the need to share raw images directly [1–5].

1.2 Motivation

Medical images often need to be kept confidential and healthcare organizations might have privacy, governance and institutional restrictions that prevent the aggregation of medical images [4–9]. FL with ViTs offers a chance to build joint classification systems that are based on representations from

transformers but keep local data ownership. Moreover, differential privacy can also provide more protection from information leakage from model updates [6, 8, 9]. These considerations make a unified framework for privacy-preserving classification of medical images desirable.

1.3 Problem statement

Although great advances have been made in deep learning and federated optimization, medical image classification still faces issues like data scarcity, class imbalance, communication costs and suboptimal class-wise performance [1–3, 15–18]. Federated training does not guarantee privacy, and applying differential privacy can also lead to some loss in predictive performance. Thus, an important research question is how to balance federated classification using ViT between diagnostic accuracy, efficient distributed training, and privacy protection.

1.4 Objectives and scope

In this current work, a privacy-centric framework is proposed which combines the advantages of ViTs and FedAvg for classifying four types of white blood cells (WBCs): eosinophils, lymphocytes, monocytes, and neutrophils. The goals are the comparison of centralized CNN, centralized ViT, and federated ViT models, as well as the analysis of the impact of communication rounds, amount of training data, and hyperparameter tuning; and the evaluation of a federated ViT with a differential privacy extension with the help of formal privacy accounting. The scope covers classification accuracy, macro-F1, class-wise recall, communication overhead, and the privacy–utility trade-off. The framework will be used to perform collaborative medical image analysis without centralizing the raw images.

2 Literature review

For medical image analysis, Federated Learning (FL) is a promising method that allows multiple institutions to jointly train a shared model while avoiding the sharing of raw patient information. McMahan et al. [1] proposed Federated Averaging (FedAvg) that computes a global model by aggregating local model parameters. Further research looked at efficiency of communication, optimization stability and statistical heterogeneity in federated environments [2, 3]. Such problems become particularly relevant in healthcare where datasets can have varying class distributions, image characteristics and sizes.

Medical applications of FL have been highlighted by the need for protecting clinical data. Federated approaches to healthcare research collaboration were considered by Rieke et al. [4], while opportunities and challenges of using FL in healthcare were highlighted by Li et al. [5]. However, local storage of raw data does not eliminate all privacy risks since model updates can be used to infer training samples' data. One of the ways to prevent information leaks based on model updates is the use of differential privacy [6, 8, 9]. In addition to this, secure aggregation allows providing extra privacy guarantees since it prevents access to individual client updates [10]. It is possible to develop privacy-preserving predictive image classification models for medical purposes.

As opposed to CNNs, where images are processed in one continuous stream, in ViTs images are represented as a collection of patches and inter-patch relationships are learned via self-attention mechanism [11]. Touvron et al. [12] found that data-efficient training strategies can improve the real-world applicability of transformer-based image classification. Since then, transformer architectures have been adopted for medical image analysis, such as classification and segmentation [13, 14]. These

developments motivate exploring the potential of a federated healthcare environment with local training and global aggregation, where computational and privacy challenges need to be addressed.

Medical image classification systems must be evaluated using datasets, and their selection is very important. MedM-NIST is a collection of standard biomedical images for simple experimentation and comparisons [15]. In the case of limited training samples, low image resolution and uneven class distribution, however, the ability of the model to generalize can be reduced. When the data distributed to the clients is not independent and identically distributed (non-IID), federated learning becomes complicated. Several techniques, like FedProx and FedBN, solve the problem of client heterogeneity and local distribution differences [3, 18]. Efficient communication and balancing computation and aggregation among local and global computations are important issues for practical deployment [2, 16, 17].

Cross-collaborative learning can be realized by optimizing with federated algorithms while maintaining the privacy of sensitive images in each participant's local memory through transformer-based classification. This will work best in specific cases for different model architectures, varying data sizes, geographical distribution of clients, local training conditions and communication frequency. Adding differential privacy also generates a trade-off between privacy protection and predictive performance. It is important to consider these factors in conjunction to establish the credibility of the framework for attaining performance in classification and providing meaningful privacy protection [6, 8–10, 19, 20].

2.1 Research gap

Researchers have studied federated learning, transformer-based image analysis, and differential privacy individually and together, but their practical integration for classifying white blood cells still needs to be evaluated. In particular, there is a need for systematic evaluation of classification performance under limited training data, for client-level data partitioning, and for different numbers of communication rounds. Local data retention is not enough to back up privacy claims; there must be formal privacy accounting.

In this study, a framework that fuses a Vision Transformer with FedAvg is used for the classification of eosinophils, lymphocytes, monocytes and neutrophils. It benchmarks the centralized CNN and ViT baselines against federated ViT training and suggests other experiments involving larger training subsets, different numbers of communication rounds, hyperparameter optimization and differential privacy. Accuracy, macro-F1, class-wise recall, communication overhead and the privacy–utility trade-off are the main evaluation criteria. The objective of the study is to set the basis for a cooperative medical image classification system, while preserving privacy without compromising performance, and to discover areas of performance and privacy issues that need to be solved before clinical use.

3 Proposed methodology

3.1 Framework overview

The proposed framework combines Federated Learning (FL) and Vision Transformers (ViTs) for the task of white blood cell image classification, while maintaining privacy. Federated learning has the idea of allowing multiple clients to train a shared model without sharing raw training data [1, 2]. This works well in healthcare applications which emphasize patient-data privacy and institutional data control [4, 5]. The framework consists of the following steps: image preprocessing, client data

partitioning, training with local models, global model aggregation, and differential privacy evaluation. Figure 1 shows the overall architecture.

The baseline models are a conventional Convolutional Neural Network (CNN) and a centralized ViT. The federated ViT uses Federated Averaging (FedAvg) for its evaluation, and differential privacy is also included as a supplementary privacy-enhancing technique [8, 9].

The expanded experimental protocol includes 4,000 training images and 400 validation images (1,000 and 100 images per class, respectively). The same subset of 600 images is used for comparison. To prevent artificial variation in the validation and testing data, augmentation is limited to training.

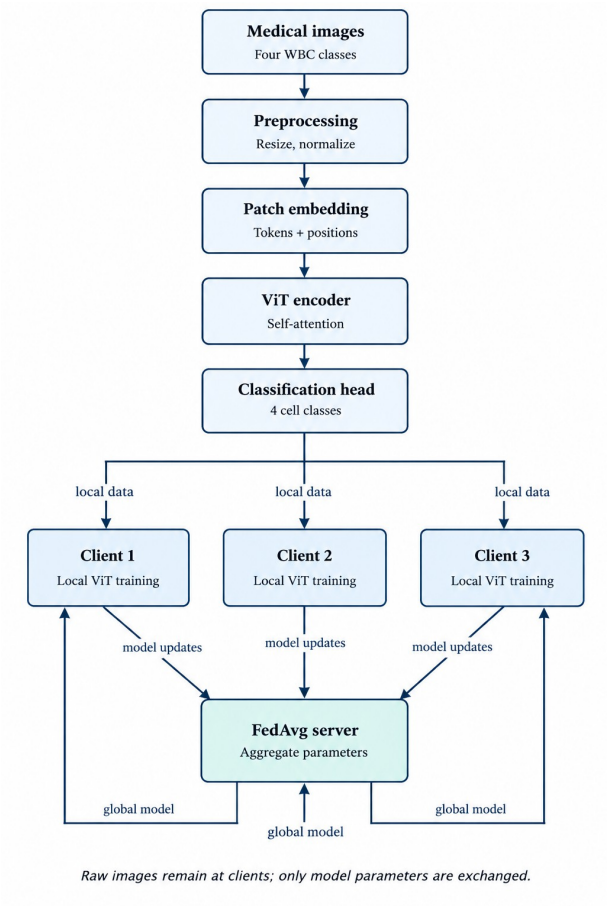


Figure 1. Overall architecture of the proposed federated Vision Transformer framework.

3.2 Dataset and preprocessing

The Blood Cell Images dataset consists of images of white blood cells classified into four categories: eosinophils, lymphocytes, monocytes and neutrophils. There are standardized evaluation protocols and medical image classification datasets available to compare deep-learning models [15].

A balanced subset of training images was used for preliminary experiments, containing 200 validation images and 600 testing images, with one class containing 250 training images, 50 images for validation, and 150 images for testing. All images were resized to 32 × 32 pixels and then normalized.

3.3 Vision Transformer architecture

The ViT architecture represents images as a sequence of embedded patches, and uses transformer encoder blocks to encode the relationships between these spatial components [11]. In environments where computational resources are constrained, the use of a compact architecture is further encouraged by the possibility of training the transformer using only the data that is needed, as opposed to the readout-only option [12]. Transformer-based approaches are also explored for medical image analysis [13, 14].

The image is split into non-overlapping image patches. Positional information is added to each patch, which is then embedded in a space and passed through multi-head self-attention and feed-forward layers. The final representation is then passed to a classification head which is responsible for predicting one of the four classes of the cell. Figure 2 shows the ViT pipeline.

The attention mechanism is represented as:

$$\text{Attention}\,(Q,\ K,\ V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K and V represent the query, key and value matrices, respectively, and d_k is the key dimension.

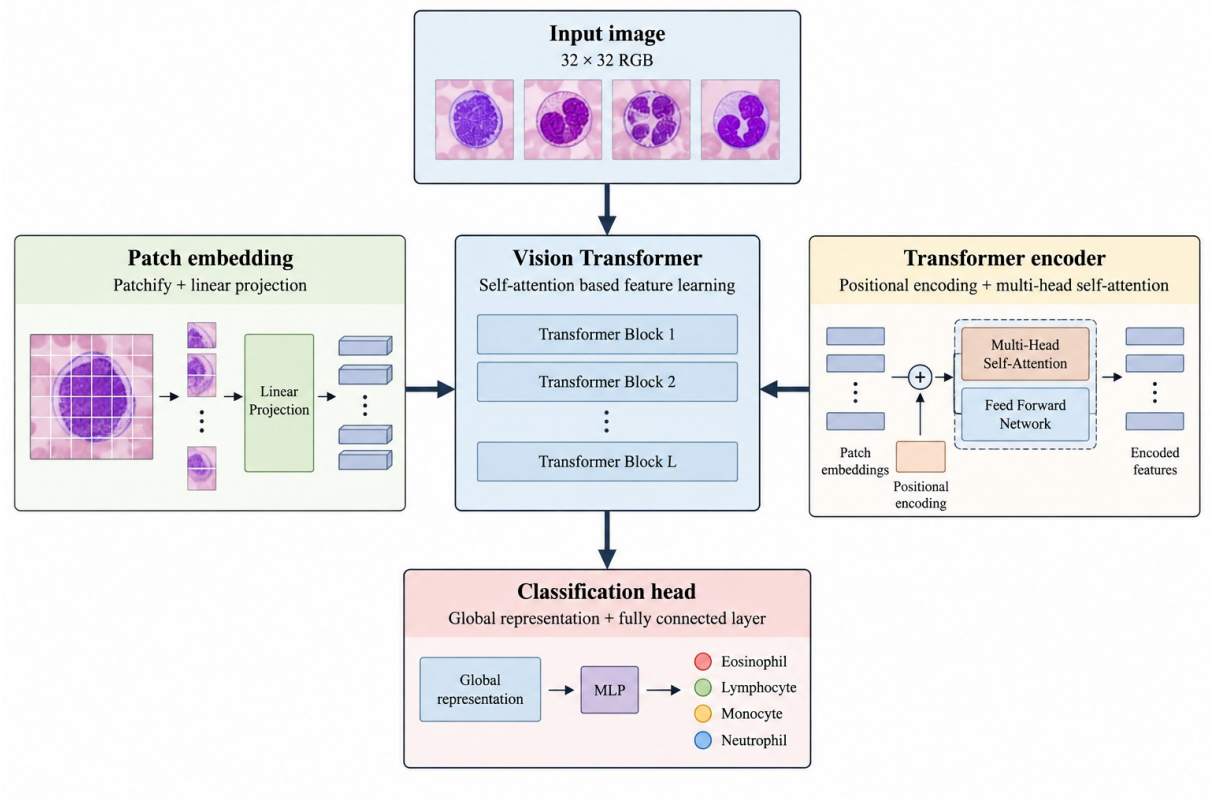


Figure 2. Vision Transformer pipeline for white blood cell image classification.

3.4 Federated training and aggregation

The training data is split over 3 simulated clients. At any point, each client has the current global model, trains it on its own images, and sends back new parameters for the model. The raw images are stored in the local client partition, following the federated optimization paradigm [1, 4, 5].

The global model is updated based on FedAvg:

$$w^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{N} w_k^{(t+1)} \tag{2}$$

Here, K is the number of clients, n_k is the number of training samples at client k, N is the total number of training samples, and $w_k^{(t+1)}$ are the updated parameters at client k.

The initial configuration requires five iterations of communication, each having two local iterations. The experiment is extended to have 5, 10, and 20 iterations of communication to observe the effects of convergence and communication cost. Another factor that needs to be considered is the heterogeneous nature of the data at the clients' ends: different distributions of data at local nodes may affect federated optimization [2, 3, 18]. The federated training cycle is shown in Figure 3.

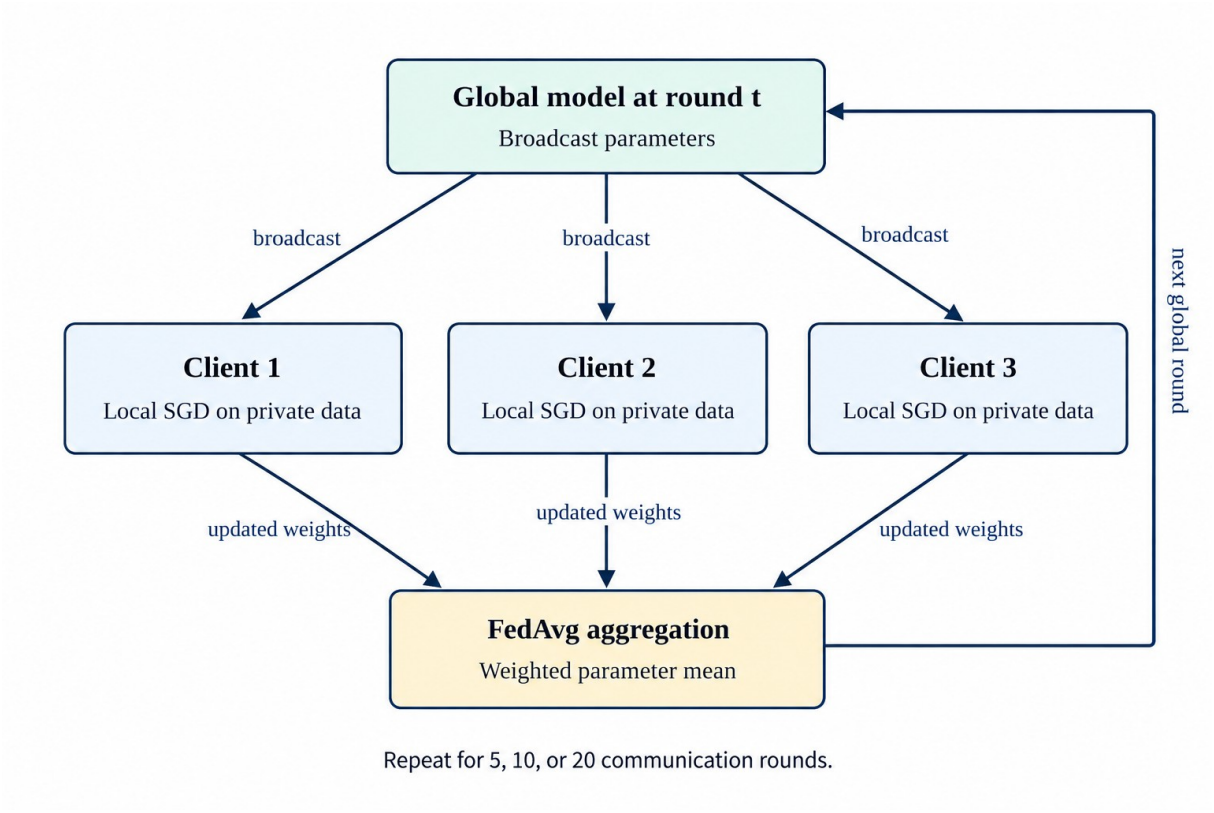


Figure 3. Federated training cycle with local optimization and global FedAvg aggregation.

3.5 Differential privacy

In federated learning, raw images are not shared with each other, but it does not prohibit updates to the models from revealing data about the training images. Therefore, differential privacy is viewed as an added layer of privacy protection [6, 8, 9]. The proposed DP-SGD adds Gaussian noise to each sample gradient and clips each gradient individually before updating the model parameters:

$$\tilde{g} = \frac{1}{B}\left[\sum_{i=1}^{B} \text{clip}\left(g_i,\ C\right) + \mathcal{N}\left(0,\ \sigma^2 C^2 I\right)\right] \tag{3}$$

where B is the batch size, g_i is the gradient for sample i, C is the clipping threshold and σ is the noise multiplier (Figure 4).

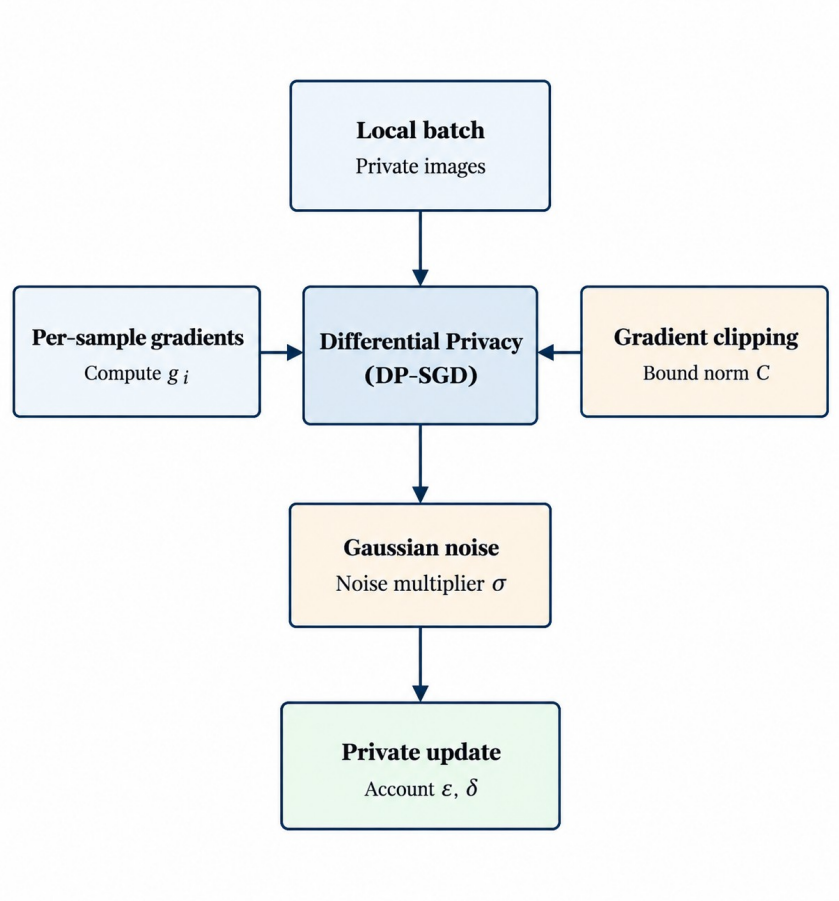


Figure 4. Conceptual differential privacy pipeline using gradient clipping, Gaussian noise, and privacy accounting.

The privacy guarantee will be reported in terms of (ε, δ) differential privacy with the privacy accountant set as the actual training procedure. This privacy budget will have to be determined after execution; it cannot be deduced from the model architecture only.

4 Experimental setup

4.1 Software and implementation environment

Experiments are conducted in a Python-based deep-learning pipeline using PyTorch. Federated training is tested by three simulated clients. The initial experiment has 32 × 32 pixel inputs and balanced class subsets, and the expanded protocol has more training data to test whether more training data leads to better generalization.

As per the literature [1, 4], the design of the experiments is divided into two parts, centralized training and federated optimization. The overall context lies in medical image analysis [13–15] and transformer-based methods.

4.2 Models and training configurations

The following configurations are used for the experimental comparison:

- **CNN ensemble:** An ensemble of conventional CNN classifiers trained at the central site.
- **Centralized ViT:** A small Vision Transformer trained with centralized data.
- **Federated ViT:** FedAvg ViT across three clients, using the same general ViT.

- **Federated DP:** Federated training with gradient clipping and Gaussian noise, to be implemented in the future.

For the preliminary centralized models, 10 epochs were used. The initial federated setup made use of 5 communication rounds and 2 local epochs per communication round. The expanded protocol runs 5, 10 and 20 rounds, and tests learning rates of 0.0001, 0.0005 and 0.001, as well as 1 and 2 local epochs.

4.3 Evaluation metrics

Overall accuracy, macro-precision, macro-recall, and macro-F1-score are used to evaluate the models. Overall accuracy is the measure of overall classification accuracy, while macro-averaged measures give each cell class the same weight. The analysis of confusion matrices and class-wise recall are used to identify the categories that are often misclassified.

5 Results and discussion

5.1 Model performance comparison

The proposed method combines the feature extraction of the Vision Transformer (ViT) with Federated Averaging (FedAvg) to enable collaborative white blood cell classification without the need to share the original training images with the clients. Unlike traditional centralized classification, it integrates representation learning via a transformer with distributed model optimization. To further protect against information leakage, differential privacy is added.

Four types of white blood cells—eosinophils, lymphocytes, monocytes and neutrophils—were used in the preliminary evaluation. The test set consisted of 600 balanced images (150 per class). The CNN and centralized ViT were trained for 10 epochs. Three simulated clients, 5 communication rounds, and 2 local epochs per round were used with the federated ViT.

The classification models' performances are compared in Table 1.

Table 1. Comparative performance of classification models.

Model	Accuracy (%)	Macro-F1 (%)
CNN baseline	53.50	48.84
Centralized ViT	58.50	51.82
Proposed federated ViT	49.50	41.20

As expected, the centralized ViT architecture achieved an accuracy of 58.50%, representing an increase of 5.00 percentage points over the baseline CNN. This clearly proves the possibility of representing an image in a transformer-based architecture. The first five rounds configuration of the proposed federated ViT achieved an accuracy of 49.50% and a macro-F1 score of 41.20%. In contrast to centralized methods, the federated training architecture can be seen as the key feature of the proposed federated model, allowing for collaborative optimization without having to aggregate raw images, but failing to reach the results of the centralized baseline methods in terms of classification task. However, the discrepancy between these two sets of models does not refute the privacy objective of the framework but rather points out to the optimization problem. Figure 5 compares the accuracy and macro-F1 of the three models.

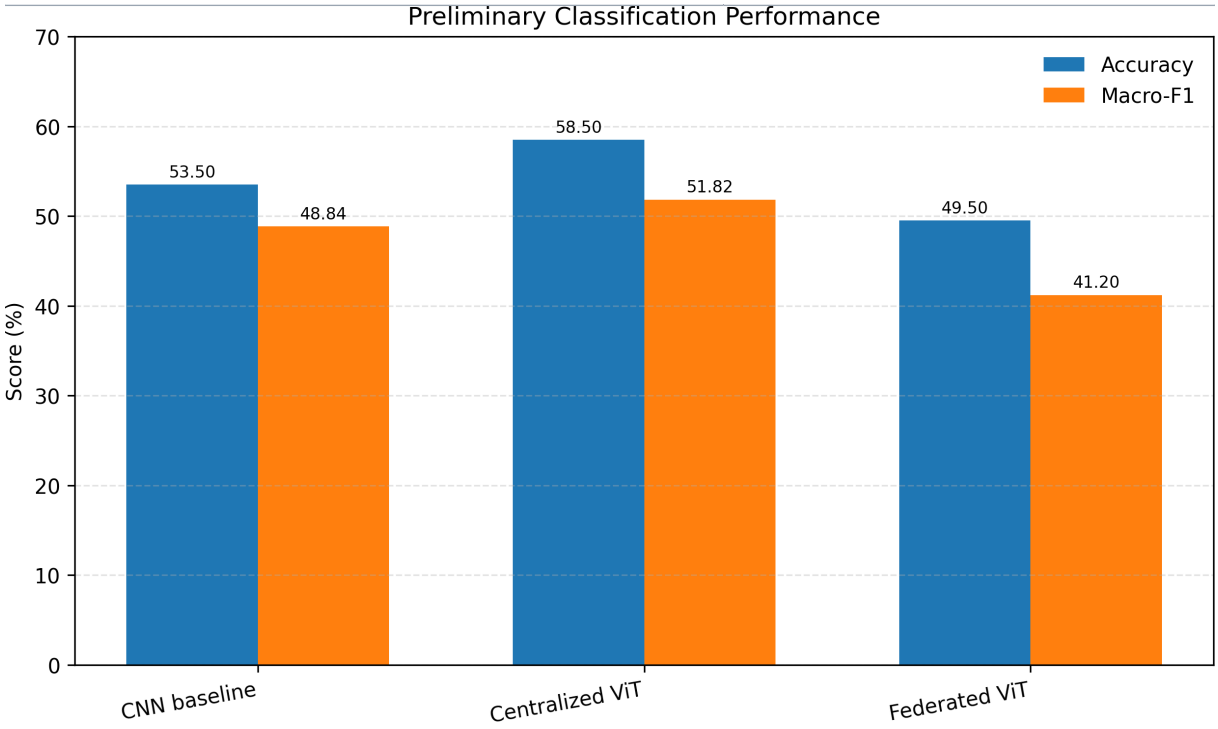


Figure 5. Preliminary accuracy and macro-F1 comparison of the CNN, centralized ViT, and federated ViT.

Table 2 and Figure 6 give the class-wise recall of the two ViT models.

Table 2. Class-wise recall of ViT models.

Cell class	Centralized ViT (%)	Proposed federated ViT (%)
Eosinophil	4.00	16.00
Lymphocyte	86.00	79.33
Monocyte	86.67	96.00
Neutrophil	57.33	6.67

Of all the models evaluated, the federated model showed the best recall performance for monocytes, at 96.00%, which is 9.33 percentage points more than the centralized ViT. Also, it improved the recall for eosinophils by 12 percentage points, from 4.00% to 16.00%. This class-dependent performance indicates that there is a differential recognition behavior in the federated optimization process.

However, in the federated model, there was lower recall for lymphocytes and neutrophils. Neutrophil recall is 6.67%, which is a major drawback and needs to be addressed prior to clinical use. The performance of this model appears to be dependent on the federated optimization setting and per-class feature learning. To prevent masking the regressions in some classes with the improvements in others, the macro-F1 and confusion matrices should also be reported in future comparisons, in addition to the overall accuracy.

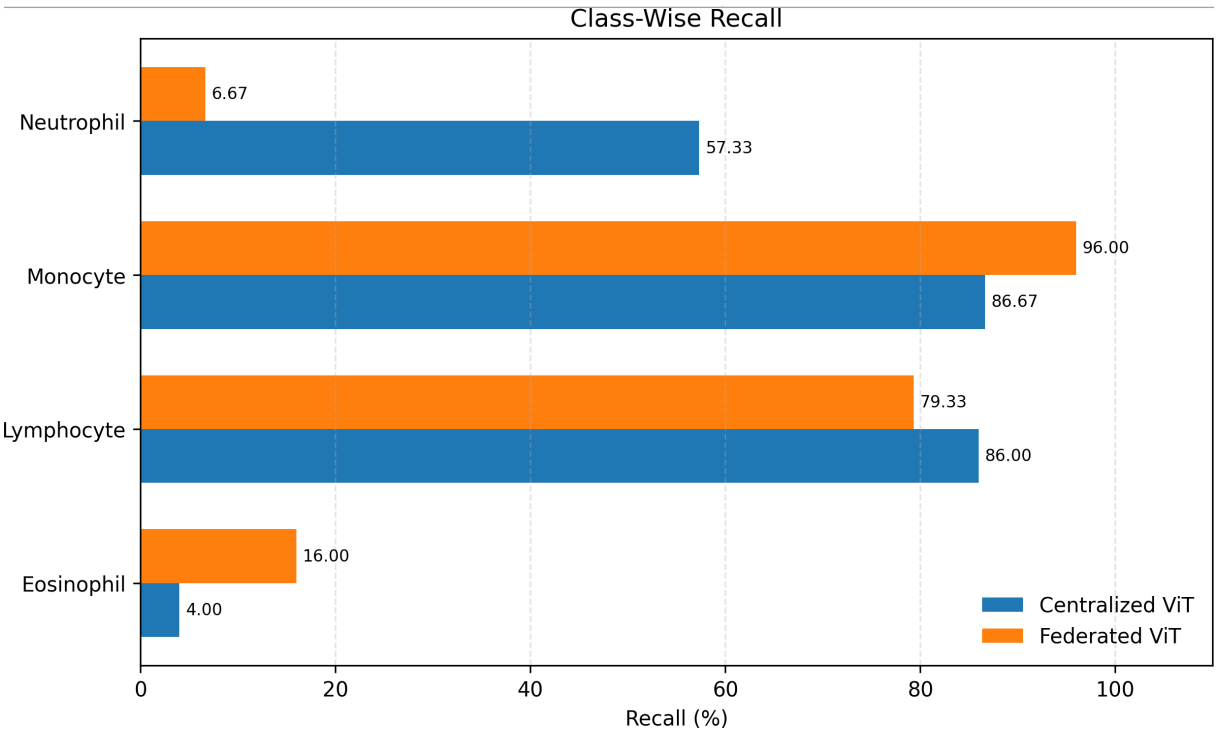


Figure 6. Class-wise recall comparison between centralized and federated Vision Transformer models.

5.2 Opportunities of the proposed framework

There are three major design benefits to the proposed approach. First, the ViT-based representation learning extracts relationships between image patches using self-attention, which is an alternative way of computing features to convolution. Second, federated aggregation enables multiple clients to feed into one model without the central collection of raw images. Third, the planned differential-privacy extension includes the addition of gradient clipping and calibrated noise, to formalize the privacy guarantee when implemented properly.

These features make the framework different from the conventional centralized classifier, as it takes into account distributed learning and privacy requirements. Raw images must, however, remain local for formal DP to be ensured, and the DP-enabled configuration needs experimental validation.

5.3 Extended evaluation and research implications

The results obtained in the current trials serve as a preliminary performance baseline. More experiments should be conducted with a larger training subset, with learning rates of 0.0001, 0.0005, and 0.001, and with one and two local epochs per round. The validation macro-F1 should serve as a guide for the selection of configurations, while the test set will be used for final evaluation.

The DP-enhanced experiment should additionally provide accuracy, macro-F1, training cost, and the estimated privacy budget (ε, δ). This will reveal if the suggested architecture is capable of preserving classification quality while providing a guaranteed level of privacy.

Finally, the main idea behind the proposed model is to integrate transformer-based classification with federated training, while DP is an optional enhancement. Based on the preliminary analysis, the model shows a positive behavior regarding specific class classification for monocytes, but at the same time faces difficulties with optimization and generalization. In order to see if there is an improvement

in performance and privacy-utility trade-off from architectural changes, the experiments should be done.

6 Conclusion and future scope

In this study, a privacy-oriented medical image classification framework was introduced that integrates Vision Transformers (ViTs) and Federated Learning (FL) for the classification of white blood cells (WBCs). The preliminary experiments indicate that the centralized ViT outperforms the CNN baseline with an accuracy of 58.50%, while the CNN baseline achieved 53.50%. In the proposed federated ViT, the accuracy was found to be 49.50% with an excellently high monocyte recall accuracy of 96.00%, showing the high potential of collaborative learning in retaining raw images at individual clients. However, there is a need for improvement of classification performance and consistency of classification per class.

The training set will be expanded, learning rates will be adjusted, and the number of local training epochs will be increased, as well as testing the convergence performance of 5, 10, and 20 communication rounds will be added. Formal privacy-accounting technology will be used to embed and assess differential privacy, to quantify the privacy versus improvement of diagnosis trade-off. Other studies will look into heterogeneous distributions on clients, higher-resolution images, more advanced data augmentation, and larger Vision Transformer architectures. Other medical imaging datasets and healthcare centers will be important to validate the results to determine their generalizability. The idea behind the framework is to provide a way to perform collaborative medical image analysis that is accurate, scalable, and preserves privacy without centralizing sensitive clinical images.

References

[1] McMahan B, Moore E, Ramage D, Hampson S, Agüera y Arcas B. Communication-efficient learning of deep networks from decentralized data. *In: Proceedings of the* 20th International Conference on Artificial Intelligence and Statistics (AISTATS). Proc Mach Learn Res. 2017;54:1273–1282.

[2] Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*. 2020;37(3):50–60. https://doi.org/10.1109/MSP.2020.2975749

[3] Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated optimization in heterogeneous networks. In: *Proceedings of Machine Learning and Systems (MLSys), vol*. 2; 2020. p. 429–450.

[4] Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. *npj Digital Medicine*. 2020;3:119. https://doi.org/10.1038/s41746-020-00323-1

[5] Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*. 2019;10(2):12. https://doi.org/10.1145/3298981

[6] Dwork C, Roth A. *The algorithmic foundations of differential privacy*. Hanover (MA): Now Publishers; 2014. https://doi.org/10.1561/0400000042

[7] Geyer RC, Klein T, Nabi M. Differentially private federated learning: A client level perspective. In: NeurIPS Workshop on Machine Learning on the Phone and Other Consumer Devices; 2017. *arXiv*:1712.07557.

[8] Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, et al. Deep learning with differential privacy. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*; 2016. p. 308–318. https://doi.org/10.1145/2976749.2978318

[9] Papernot N, Song S, Mironov I, Raghunathan A, Talwar K, Erlingsson Ú. Scalable private learning with PATE. In: *International Conference on Learning Representations (ICLR)*; 2018.

[10] Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan HB, Patel S, et al. Practical secure aggregation for privacy-preserving machine learning. In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*; 2017. p. 1175–1191. https://doi.org/10.1145/3133956.3133982

[11] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR)*; 2021.

[12] Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. *In: Proceedings of the* 38th International Conference on Machine Learning (ICML). Proc Mach Learn Res. 2021;139:10347–10357.

[13] Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, et al. UNETR: Transformers for 3D medical image segmentation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; 2022. p. 1748–1758. https://doi.org/10.1109/WACV51458.2022.00181

[14] Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. TransUNet: Transformers make strong encoders for medical image segmentation [preprint]. *arXiv*:2102.04306. 2021.

[15] Yang J, Shi R, Wei D, Liu Z, Zhao L, Ke B, et al. MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data*. 2023;10:41. https://doi.org/10.1038/s41597-022-01721-8

[16] Wang S, Tuor T, Salonidis T, Leung KK, Makaya C, He T, et al. Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications*. 2019;37(6):1205–1221. https://doi.org/10.1109/JSAC.2019.2904348

[17] Bonawitz K, Eichner H, Grieskamp W, Huba D, Ingerman A, Ivanov V, et al. Towards federated learning at scale: System design. In: *Proceedings of Machine Learning and Systems (MLSys), vol*. 1; 2019. p. 374–388.

[18] Li X, Jiang M, Zhang X, Kamp M, Dou Q. FedBN: Federated learning on non-IID features via local batch normalization. In: *International Conference on Learning Representations (ICLR)*; 2021.

[19] Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Bhagoji AN, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*. 2021;14(1–2):1–210. https://doi.org/10.1561/2200000083

[20] Alistarh D, Hoefler T, Johansson M, Konstantinov N, Khirirat S, Renggli C. The convergence of sparsified gradient methods. In: *Advances in Neural Information Processing Systems (NeurIPS), vol*. 31; 2018.