International Journal on Document Analysis and Recognition (IJDAR)
https://doi.org/10.1007/s10032-024-00469-8

SPECIAL ISSUE PAPER

Towards privacy preserved document image classification: a comprehensive benchmark

Saifullah Saifullah[1,2] · Dominique Mercier[1,2] · Stefan Agne[1,3] · Andreas Dengel[1,2] · Sheraz Ahmed[1,3]

Received: 14 November 2023 / Revised: 11 March 2024 / Accepted: 8 May 2024

Abstract
As data-driven AI systems become increasingly integrated into industry, concerns have recently arisen regarding potential privacy breaches and the inadvertent leakage of sensitive user data through the exploitation of these systems. In this paper, we explore the intersection of data privacy and AI-powered document analysis systems, presenting a comprehensive benchmark of well-known privacy-preserving methods for the task of document image classification. In particular, we investigate four different privacy methods—Differential Privacy (DP), Federated Learning (FL), Differentially Private Federated Learning (DP-FL), and Secure Multi-Party Computation (SMPC)—on two well-known document benchmark datasets, namely RVL-CDIP and Tobacco3482. Furthermore, we investigate the performance of each method under a variety of configurations for thorough benchmarking. Finally, the privacy strength of each approach is assessed by subjecting the private models to well-known membership inference attacks. Our results demonstrate that, with sufficient tuning of hyperparameters, Differential Privacy (DP) can achieve reasonable performance on the task of document image classification while also ensuring rigorous privacy constraints, both in standalone and federated learning setups. On the other hand, while FL-based approaches present less implementation complexity and incur little to no loss in performance on the task, they do not offer sufficient protection against privacy attacks. By rigorously benchmarking various privacy approaches, our study paves the way for integrating deep document classification models into industrial pipelines while meeting regulatory and ethical standards, including GDPR and the AI Act 2022.

Keywords Document image classification · Privacy-preserving deep learning · Differential privacy · Federated learning · Secure multi-party computation

1 Introduction

The rapid evolution of artificial intelligence (AI), notably in computer vision [1, 2] and natural language processing [3, 4], has revolutionized the field of document analysis, with modern Deep Learning (DL)-based systems delivering superhuman performances across a row of document understanding tasks [5–9]. Simultaneously, however, concerns have been raised about potential privacy breaches and the inadvertent leakage of sensitive user data through the widespread use of such data-driven AI systems [10–15].

Numerous recent studies [12, 15–17] have shown that, when not trained with rigorous privacy constraints, deep learning (DL) models can readily become sources of information leakage. Reconstruction of the training data statistics [15, 18], inferring whether a sample comes from the training data distribution [12–14], or model stealing [19, 20] are just a few examples of potential privacy violations. If deep models

✉ Saifullah Saifullah
saifullah.saifullah@dfki.de

Dominique Mercier
dominique.mercier@dfki.de

Stefan Agne
stefan.agne@dfki.de

Andreas Dengel
andreas.dengel@dfki.de

Sheraz Ahmed
sheraz.ahmed@dfki.de

[1] German Research Center for Artificial Intelligence, 67663 Kaiserslautern, Germany

[2] RPTU Kaiserslautern-Landau, 67663 Kaiserslautern, Germany

[3] DeepReader GmbH, 67663 Kaiserslautern, Germany

Published online: 18 June 2024

with such vulnerabilities were to be directly trained on private document data, which often contains sensitive information—such as names, addresses, contact details, social security numbers, financial particulars, and, most critically, an organization's intellectual property—they could potentially be exploited, leading to significant harm to individuals or organizations. As a result, the integration of these models into industry and their adherence to regulatory and ethical standards, such as GDPR [21] and the AI Act 2022, still faces substantial obstacles.

To address the aforementioned privacy challenges in AI-powered systems, a number of privacy-preserving approaches have been recently developed [22–28]. Most notably, Differential Privacy (DP) [22], Federated Learning (FL) [23, 24], have demonstrated promising results across various application domains such as medical imaging [29], time series analysis [26], and natural language processing (NLP) [13, 14, 18, 30]. In the context of document AI, such privacy techniques may be applied under different settings. For instance, an organization providing document AI services may train the models under global privacy constraints [22, 30] to safeguard its own private data, or under local privacy constraints [31, 32], where each individual client only uploads a private augmented data to the service, leaving no fingerprint that can be traced to the client. On the other hand, federated learning [23, 24] may be deployed for private aggregation of data across multiple organizations or clients. In this scenario, each party organization only locally trains the model and uploads it to a global service provider, thus keeping its own data on-site.

In this paper, we focus on document image classification, a fundamental component of modern document processing pipelines [5, 6, 33, 34], typically employed at an initial stage to categorize or filter the documents prior to further processing. Numerous DL-based classification models [5–7, 33–36], have been proposed in recent years for this task, showcasing extraordinary performance gains as compared to their traditional counterparts [37, 38]. However, these models are also data-driven, relying on unaltered document images as input, and thus could easily become the target of membership inference [12] or model inversion attacks [15]. In addition, the unintentional memorization [17] of training samples in these models could directly expose information about the training dataset. Surprisingly, while a plethora of research has been conducted on both document classification [5, 6, 35] and privacy in textual documents [13, 14, 18, 30], we found no existing literature in the field that addresses the issue of data privacy and potential information leakage from AI-powered document image classification systems. In this work, therefore, we investigate the potential of latest privacy preservation techniques [22, 23, 26, 28] in combination with state-of-the-art DL-based document image classification models to assess whether they can achieve sufficient utility under strong privacy constraints. The main contributions of this paper are two-fold:

- We present a comprehensive performance benchmark of four different state-of-the-art privacy methods—Differential Privacy (DP), Federated Learning (FL), Differentially Private Federated Learning (DP-FL), and Secure Multi-Party Computation (SMPC)—on two prominent document benchmark datasets, RVL-CDIP and Tobacco 3482, for the task of document image classification. To the best of authors' knowledge, this is the first work in this direction.
- In an extensive analysis, we assess the aforementioned privacy approaches for the task of document image classification under a variety of settings, evaluating their performance, practical feasibility, robustness to membership inference attacks (MIA), and impact on model explainability.

2 Related work

2.1 Privacy preserving machine learning (PPML)

Privacy Preserving Machine Learning (PPML) has garnered significant attention in recent years, with numerous studies exploring both the vulnerability of deep networks to privacy attacks and developing safeguards in response.

2.1.1 Privacy attacks

The three most prominent types of privacy attacks are model inversion [15, 18], membership inference [12], and model extraction [19, 20].

Model Inversion: Model inversion attacks [15] may be employed to reconstruct the training dataset statistics by utilizing the model confidence information; for instance, by applying this attack to a face recognition model, Fredrikson et al. [15] were able to reconstruct the faces of individuals based on their associated identity labels. Similarly, Coavoux et al. [15] demonstrated how a malicious eavesdropper may recover information about sensitive private data samples from their neural representations. Taking a step further, Hitaj et al. [16] proposed a generative model capable of reconstructing the training data in a collaborative learning environment.

Membership inference: Shokri et al. [12] introduced membership inference attacks (MIA) for machine learning models, the goal of which is to determine whether a particular sample was part of the model's training set. They further demonstrated that MIAs can be successfully applied

to DL models, even with only a black-box access to the target model, achieved by training multiple shadow models that mimic the target model. A number of derivative works have further explored membership inference attacks in other tasks [10, 13, 14].

Model Extraction: Tramèr et al. [20] proposed model extraction attacks for machine learning models, aiming to steal the weights of target models by performing multiple queries on them. Other works have extended this approach to steal the training hyperparameters [39]. In a slightly different direction, Milli et al. [19] recently demonstrated how gradient-based explanations of DL models may also be utilized to extract model parameters.

2.1.2 Privacy defenses

Numerous methods have been developed recently to protect against the aforementioned privacy attacks. Traditionally, data anonymization [40] was a common approach for protecting an individuals' information. However, this method has proven to be insufficient in safeguarding against the more recent re-identification attacks [10]. Recently, Mohassel et al. [27] proposed a two-server model that utilizes secure multiparty computation (SMPC) to train neural networks over multiple partitions of the dataset. Similarly, Knott et al. [28] introduced an SMPC framework for DL models that allows the encryption of both models and data from multiple parties through secret sharing.

One of the most prominent techniques to safeguard against privacy attacks is Differential Privacy (DP) [41], which, by definition, offers strong privacy guarantees against membership inference [12] and linkage attacks used for de-anonymization [10]. To implement Differential Privacy (DP) in a deep supervised learning setting, Abadi et al. [22] proposed DP-SGD, a machine learning optimization method that ensures strict privacy guarantees during model training. Several derivatives of DP-SGD [30, 42] have recently been proposed, aiming to enhance the technique to increase model efficiency while maintaining robust privacy constraints.

Federated Learning (FL) [23] is another popular approach to privacy preservation. In FL, the optimization of a machine learning model is distributed among multiple parties, allowing them to keep their data confidential while safely contributing it to model training. FL has been combined with DP in multiple derivative works [25, 26] to provide even stronger privacy constraints. For a detailed overview of different types of privacy attacks and defenses, we refer the reader to related surveys [10, 11].

2.2 Document image classification

There is extensive literature on the subject of document image classification. Early work in this area mainly focused on exploiting structural similarity [43], feature matching [37], or classical machine learning approaches such as K-Nearest Neighbors [44] or Hidden Markov Models [38] to distinguish between different classes of documents. For a detailed overview of classical techniques, we refer the reader to a related survey [45].

With the advent of deep learning, the field of document image classification has experienced significant improvement, recently witnessing a surge in both image-based unimodal [5, 6, 33] and multimodal techniques [7, 36, 46, 47]. Afzal et al. [5] were the first to achieve breakthrough performance in document image classification by leveraging the potential of transfer learning in conjunction with deep convolutional neural networks (CNNs). Ferrando et al. [6] leveraged recent advances in convolutional neural networks along with parallel training techniques in deep learning to significantly improve performance in image-based classification. Saifullah et al. [35] recently introduced DocXClassifier, a state-of-the-art transformer-inspired CNN that not only attains the highest performance in image-based classification but also possesses the property of being inherently explainable. Recent works [48, 49] have also explored the use of Vision Transformers (ViTs) [50] for the document image classification task but have found it challenging to surpass CNNs using basic training approaches, even on sufficiently large datasets. However, a recent study [51] has shown that extensive pre-training enables ViTs to achieve performance levels comparable to those of CNNs but at the cost of additional training and increased complexity.

In the multimodal domain, document classification methods typically involve preprocessing documents to extract the layout and textual content from the images. Subsequently, visual, textual and layout features are utilized in combination to perform the classification task. Several approaches, including multi-stream models [46, 47] and transformer-based models [7, 36, 52], have recently been proposed in this area, demonstrating exceptional performance improvements.

Despite numerous advances in the field, there is a scarcity of literature actively addressing the problem of privacy preservation in both image-based and multimodal document classification. Moreover, we suspect that multimodal techniques, which extract both image and text data from the input, will open up new opportunities for various types of privacy attacks to extract information. Therefore, it is now of paramount importance to explore state-of-the-art privacy protection methods in this area, ensuring that existing and future document image classification systems can be safely deployed.

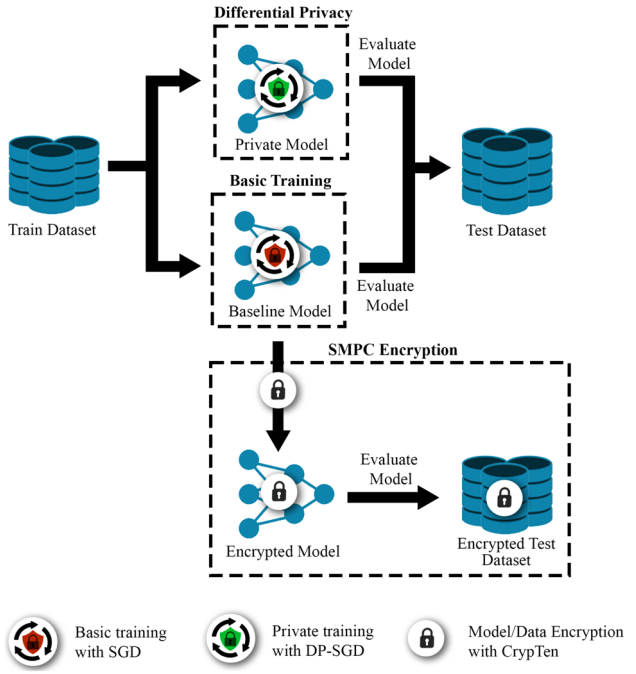


Fig. 1 An overview of standalone privacy techniques, differential privacy (DP) and secure multiparty computation (SMPC) in comparison to standard model training

3 Methods

In this section, we briefly describe the different privacy preservation techniques that we have investigated in this study.

3.1 Differential privacy (DP)

Differential privacy (DP) [41] provides a formal definition for information release from an algorithm and, by definition, offers rigorous privacy guarantees against various types of privacy attacks [12, 15]. In this work, we focus on *example-level privacy* under the global approximate-DP (also known as (ε, δ)-DP) setting, formally defined as follows:

Definition 1 A randomized algorithm $\mathcal{M} : \mathcal{D} \rightarrow \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ is (ε, δ)-differentially private if for all $S \subseteq \mathcal{R}$ and for all datasets $D, D' \in \mathcal{D}$ that differ at most in one record:

$$\mathbb{P}[\mathcal{M}(D) \in S] \leq e^{\varepsilon}\mathbb{P}[\mathcal{M}(D') \in S] + \delta$$

where the term $\mathbb{P}[\mathcal{M}(D) \in S]$ refers to the probability that the output of the algorithm $\mathcal{M}$ when applied to the dataset D lies in the subset S. Note that this a general definition of (ε, δ)-DP and may be applied to any kind of randomized algorithm $\mathcal{M}$ and dataset $\mathcal{D}$, both which may vary under different settings. Similarly, the definition of adjacency between the datasets D, D' may also vary between tasks. In this work, since we are dealing with a classification task that involves image-label pairs, we consider two datasets as adjacent if they differ only in a single image-label pair, similar to the previous work [22].

Intuitively speaking, the above definition of DP ensures that any output generated by the algorithm $\mathcal{M}$ on similar inputs (D, D') is difficult to distinguish. Meanwhile, the magnitude of this indistinguishability is captured by the two privacy parameters ε and δ, with ε denoting the upper bound on the overall privacy leakage and δ representing the probability of failure in preserving this bound. Consequently, smaller values of the pair (ε, δ) indicate stronger privacy guarantees.

3.1.1 DP-SGD

For the practical implementation of global approximate-DP in a machine learning setting, Abadi et al. [22] proposed the Differentially Private Stochastic Gradient Descent (DP-SGD) optimization algorithm, which ensures *example-level privacy* under (ε, δ)-DP during the training process of a machine learning model. This is achieved by first clipping the per-example gradients of the training samples to a fixed norm C and then adding Gaussian noise $n \sim \mathcal{N}(0, \sigma^2 C^2)$ to the gradients during the model optimization step, ensuring that the model's overall dependence on each input sample remains minimal. Where the noise multiplier σ determines the privacy strength, with higher values corresponding to lower ε. Its value is generally determined based on the available privacy budget (ε, δ), the privacy accountant [22, 53–55] which tracks the privacy loss ε, total number of optimization steps T, and the data sampling rate q.

For the image classification task, apart from differences in the optimization step, the training routine in DP-SGD largely remains the same as that of the standard supervised setting, including the input, target, model, and loss function (see Fig. 1). However, unlike standard supervised learning, where samples are randomly drawn from the dataset using a uniform distribution, DP-SGD [22] employs Poisson sampling, in which each sample is independently drawn from the dataset with a fixed sampling rate $q = \frac{B}{\|\mathcal{D}\|}$, where B is the training batch size, and $\mathcal{D}$ is the training dataset. Note that while initially designed for the standard SGD optimizer, extending DP-SGD [22] to other machine learning optimizers, such as Adam [56], is straightforward. For a complete pseudocode of the DP-SGD/Adam algorithm, refer to Appendix A.2.1. In this work, we investigated both DP-SGD and DP-Adam for various deep learning models for the task of document image

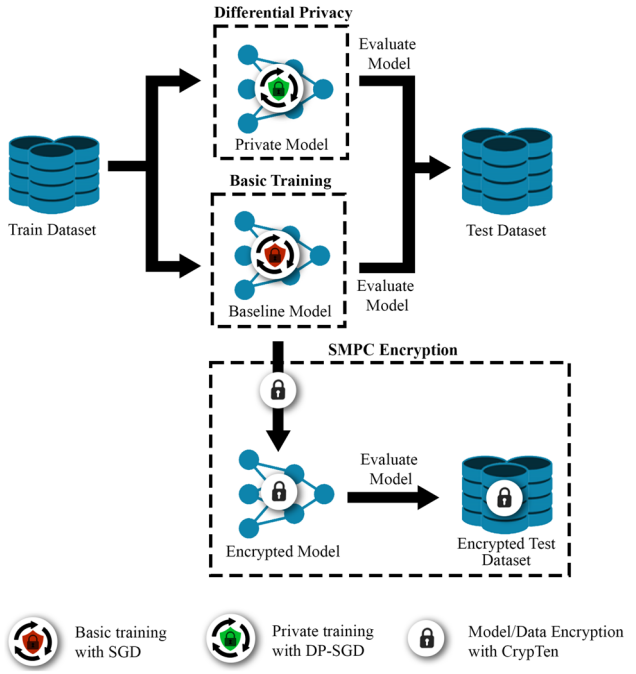


Fig. 2 An overview of the federated learning-based privacy techniques, federated averaging (FedAVG), federated ensembling (FedENS), and federated differential privacy (FedAVG-DP)

classification. For the practical implementation of these algorithms, we utilized the Pytorch Opacus[1] [57] library in this work.

3.2 Federated learning

Federated Learning (FL) represents a class of privacy-preserving techniques designed to train a centralized machine learning model in a distributed manner. This is achieved by executing model optimization updates across multiple remote clients while keeping the local data of each client private. One such algorithm is Federated Averaging (FedAVG) [23], which allows each client to safely contribute its private data to the training of a global centralized model. Given a total number of N_c clients contributing their data to the training process, in each training round of the FedAVG algorithm, a subset f_c of clients is randomly sampled to perform E_{local} local training epochs to train the local models. The weights of the local models are then sent to the central server, averaged to obtain the global model, and then sent back to the clients for the next round of local training. To assess the effectiveness of the FedAVG algorithm, we explored various configurations in this work, involving the total number of clients N_c, the fraction of sampled clients f_c, and the individual training hyperparameters.

In addition to FedAVG, we also assessed the performance of Federated Ensembling (FedENS) [26] in this work. In the FedENS algorithm, each client trains its local model on its respective local dataset using a standard training procedure. Subsequently, model ensembling is carried out on the resulting local models to evaluate their performance on the global test dataset. In this work, we performed model ensembling using weighted softmax averaging over the model outputs, with each model's weight assigned proportional to its performance on the corresponding local validation set.

Finally, we also investigated a combined setting of FedAVG with Differential Privacy (DP), namely FedAVG-DP, with several configurations. This approach provides even stronger privacy constraints by not only ensuring that the local data of each party remains on-site, but also guaranteeing that their respective gradient updates remain private. A comparison of the various federated learning approaches investigated in this work is illustrated in Fig. 2. For the practical implementation of FL-based algorithms, we utilized the Flower Federated Learning Framework[2] [58]. The complete pseudocodes of all three algorithms FedAVG, FedENS, and FedAVG-DP are provided in Appendix A.2.2.

3.3 Secure multiparty computation (SMPC)

Secure Multi-Party Computation (SMPC) enables multiple parties to perform computations on shared data through secret-sharing while maintaining the privacy of each individual party's data. SMPC holds significant potential for deep learning applications in collaborative environments, as it facilitates encrypted training and evaluation of machine learning models among multiple parties. Knott et al. [28] recently introduced CrypTen, a well-established framework for integrating SMPC-based encryption with standard Pytorch models. In this work, we specifically utilized CrypTen [28] to investigate SMPC in the context of model hiding. In model hiding, the model is first trained using a standard training procedure and then encrypted using secret-sharing, as illustrated in Fig. 1. When a client requires access to the model for making predictions on its own data, it also encrypts the data before sending it to the model, and in return receives an encrypted result from the model, which can then be safely decrypted by the client. A major advantage of SMPC-based model hiding compared to other privacy approaches is that the training procedure remains unchanged, and any pre-trained model can be easily encrypted. In this work, we investigated the overall impact of encryption on the classification performance and inference time of the models.

[1] https://github.com/pytorch/opacus.

[2] https://github.com/adap/flower.

4 Experiments and results

In this section, we present the results of our experiments conducted in this work to assess the performance of different privacy preservation methods for the task of document image classification.

4.1 Datasets

To thoroughly investigate the performance of different privacy preserving approaches, we conducted our experiments on two publicly available datasets, namely RVL-CDIP [34] and Tobacco3482[3], both of which have been extensively utilized in the field of document image classification [5, 6, 34] for benchmarking the performance of deep document classification models. RVL-CDIP [34] is a large-scale document dataset which consists of 400K labeled document images distributed across 16 document categories and has a balanced class distribution. The dataset is partitioned into training, testing, and validation sets of sizes 320K, 40K, and 40K, respectively. Tobacco3482, on the other hand, is a comparatively small-scale dataset, consisting of 3482 labeled document images grouped into 10 classes and featuring an imbalanced class distribution. In this work, we partitioned the dataset into training, testing, and validation sets of sizes 2504, 700, and 278, respectively.

4.2 Models

To perform a comprehensive comparative analysis of different privacy preserving approaches with standard non-private training, we investigated a total of 8 deep learning models, including state-of-the-art models which have been shown to perform exceptionally well on the document image classification task in the past. From the work of Afzal et al. [5], we investigated the following models: AlexNet [59], ResNet-50 [1], and VGG-16 [60]. From the work of Ferrando et al. [6], we investigate the EfficientNet-B4 [61], which showed the highest performance on the RVL-CDIP [34] dataset at the time. From the work of Saifullah et al. [35], we investigated both ConvNext-B [2] and DocXClassifier-B [35] models, which demonstrate the current state-of-the-art performance in image-based document classification on both RVL-CDIP [34] and Tobacco3482 datasets. Finally, since Vision Transformers (ViTs) have also been explored in multiple recent studies [48–51] and show promising results, we also investigated two standard ViTs—namely, ViT-B/16 [50] and ViT-L/32 [50]—to assess their performance in comparison to the CNN architectures under private training.

[3] https://www.kaggle.com/datasets/patrickaudriaz/tobacco3482jpg.

4.3 Training setup

To reproduce the performance of the models under non-private setting and to train the models under DP-excluded federated learning setups, we adopted the same training configurations as proposed in the original studies [5, 6, 35]. In particular, to train AlexNet [59], VGG-16 [60], and ResNet-50 [1], we initialized the models with ImageNet [62] pre-trained weights and then fine-tuned them on the target document datasets (RVL-CDIP [34] or Tobacco3482) with SGD optimizer and input images resized to a fixed resolution of 224×224. When training ViTs (ViT-B/16 [50] and ViT-L/32 [50]), we maintained the same approach but employed the Adam optimizer. EfficientNet-B4 [61] was also trained in a similar fashion but with a multi-GPU setting and input images of resolution 384×384, following the approach outlined by Ferrando et al. [6]. Finally, for the ConvNext-B [2] and DocXClassifier-B [35] models, we employed the training strategy proposed in [35], which involves training the models with Adam optimizer, images of resolution 384×384, and advanced regularization and data augmentation strategies applied during the process.

For experiments involving DP, we excluded all types of data augmentation and regularization techniques from the training process since the noise added by DP-SGD/Adam itself acts as a strong regularizer. Consquently, all the models were trained with the same setup, except for the image resolutions, which remained consistent with those used in non-private training. Furthermore, given previous findings that domain-specific pre-training in a DP environment can yield significant performance improvements [63], we investigated private training on the Tobacco3482 dataset under two settings: $\text{Tobacco3482}_{ImageNet}$ and $\text{Tobacco3482}_{RVL-CDIP}$. In the $\text{Tobacco3482}_{ImageNet}$ setting, models were initialized with ImageNet pre-trained weights, while in the $\text{Tobacco3482}_{RVL-CDIP}$ setting, they were initialized with RVL-CDIP [34] pre-trained weights in order to assess the effectiveness of document-specific pre-training in enhancing the utility of private training.

4.4 Evaluating differential privacy (DP-SGD/Adam)

4.4.1 Experimental setup

In this experiment, we trained all the models on target datasets (RVL-CDIP [34] and Tobacco3482) under differential privacy (DP), utilizing either the DP-SGD or DP-Adam algori thm—whichever proved best suited for the specific model architecture—as explained in Sect. 3.1.1. To track privacy loss, we employed the Rényi Differential Privacy (RDP) accountant [54] and searched for the noise multiplier σ based on the maximum target privacy budgets of $\varepsilon_{target} \in \{5, 10\}$ and $\delta = \frac{1}{\|\mathcal{D}\|}$ over a fixed number of training epochs,

Table 1 Performance comparison of baseline document image classification models with their differentially private (DP-SGD/Adam) counterparts under two configurations of target privacy budget: $\varepsilon_{target} = 5$ and $\varepsilon_{target} = 10$

Dataset	Model	Approach	$\text{Acc}_{Baseline}$		$\text{Acc}_{DP}/\varepsilon$	
			Reported	Ours	$\varepsilon_{target} = 5$	$\varepsilon_{target} = 10$
RVL-CDIP [34]	AlexNet [59]	Afzal et al. [5]	88.60	87.90	77.05 / 5.0	77.18 / 10.0
	VGG-16 [60]	Afzal et al. [5]	90.97	91.00	**81.40** / 5.0	80.63 / 10.0
	ResNet-50 [1]	Afzal et al. [5]	90.40	90.50	78.34 / 5.0	79.21 / 10.0
	EfficientNet-B4 [61]	Ferrando et al. [6]	92.31	92.60	73.54 / 5.0	73.88 / 10.0
	ConvNext-B [2]	Saifullah et al. [35]	**94.04**	93.64	79.68 / 5.0	82.32 / 9.3
	DocXClassifier-B [35]	Saifullah et al. [35]	94.00	**93.74**	80.50 / 5.0	**83.38 /6.7**
	ViT-B/16 [50]	Afzal et al. [5]	–	89.22	73.45 / 5.0	76.05 / 10.0
	ViT-L/32 [50]	Afzal et al. [5]	–	88.64	73.99 / 5.0	75.26 / 10.0
Tobacco3482 (ImageNet pre-training)	AlexNet [59]	Afzal et al. [5]	75.73	79.28	68.00 / 5.0	68.29 / 10.0
	VGG-16 [60]	Afzal et al. [5]	77.52	82.14	68.00 / 5.0	69.71 / 9.57
	ResNet-50 [1]	Afzal et al. [5]	67.93	76.57	44.44 / 5.0	46.29 / 9.13
	EfficientNet-B4 [61]	Ferrando et al. [6]	85.99	78.57	30.57 / 5.0	29.14 / 6.61
	ConvNext-B [2]	Saifullah et al. [35]	–	**89.57**	**71.58 / 5.0**	**74.29 / 9.13**
	DocXClassifier-B [35]	Saifullah et al. [35]	**87.43**	87.43	71.42 / 5.0	72.57 / 7.85
	ViT-B/16 [50]	Afzal et al. [5]	–	80.57	59.85 / 5.0	64.85 / 10.0
	ViT-L/32 [50]	Afzal et al. [5]	–	82.57	58.85 / 5.0	63.00 / 10.0
Tobacco3482 (RVL-CDIP pre-training)	AlexNet [59]	Afzal et al. [5]	90.04	89.57	87.85 / 5.0	87.71 / 9.2
	VGG-16 [60]	Afzal et al. [5]	91.01	94.14	88.57 / 5.0	89.14 / 10.0
	ResNet-50 [1]	Afzal et al. [5]	91.13	90.14	91.00 / 5.0	89.57 / 8.2
	EfficientNet-B4 [61]	Ferrando et al. [6]	94.04	94.04	90.57 / 5.0	90.14 / 9.8
	ConvNext-B [2]	Saifullah et al. [35]	**95.00**	**94.71**	**92.44 / 5.0**	**92.43 /9.8**
	DocXClassifier-B [35]	Saifullah et al. [35]	**95.00**	**94.71**	91.29 / 4.0	92.14 / 9.2
	ViT-B/16 [50]	Afzal et al. [5]	–	90.00	84.14 / 5.0	85.85 / 10.0
	ViT-L/32 [50]	Afzal et al. [5]	–	88.71	86.00 / 5.0	87.14 / 10.0

The highest accuracy scores for each setting are bolded. As can be observed, ConvNext-B [2] and DocXClassifier-B [35] significantly outperformed the other models in the majority of DP settings, whereas ResNet-50 [1] and EfficientNet-B4 [61] were severely affected by the re-initialization of normalization layers in $\text{Tobacco3482}_{ImageNet}$ setting. On the other hand, the ViTs [50] severely under-performed compared to the other models

where $\mathcal{D}$ is the target training dataset. For details on various types of privacy accountants and their significance, refer to Appendix A.1. Since it has been shown in multiple previous works that the DP-SGD/Adam algorithm is highly sensitive to the choice of hyperparameters [14, 30], such as the learning rate η, batch size B, and gradient clipping norm C, we determined the best set of hyperparameters for different models using a grid-search, the details of which are presented in Sect. 4.5. In addition, since training the models under DP necessitates the removal of batch normalization layers (BN), we substituted these layers with group normalization (GN) wherever necessary. We trained all the models for a fixed number of target epochs and report the test accuracy and privacy loss ε of the models that performed the best on the respective validation sets.

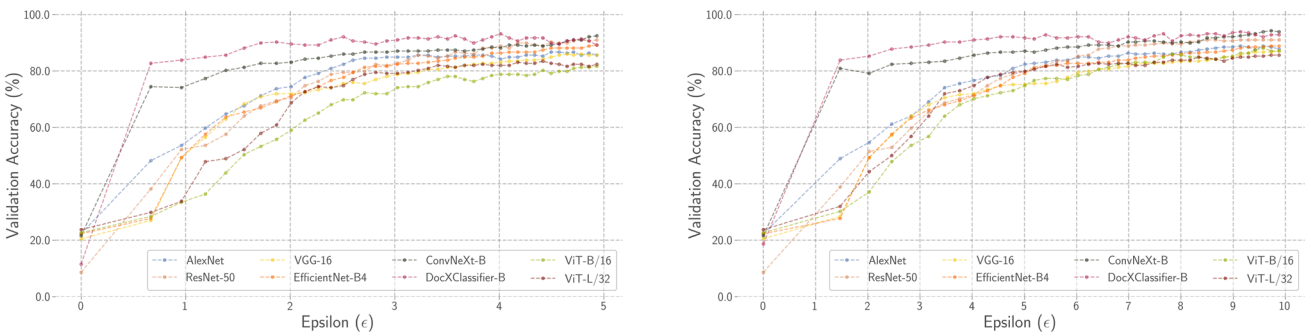
4.4.2 Experimental results

The results of these experiments are summarized in Table 1. For each model and dataset, we present the accuracy achieved by the baseline models, as well as the best accuracy and ε achieved under the DP settings $\varepsilon_{target} \in \{5, 10\}$.

Results on RVL-CDIP: It is evident that, with ImageNet pre-training on RVL-CDIP, achieving strong privacy constraints was accompanied with a significant loss of performance across all models. Interestingly, it can be observed that the larger CNNs, including VGG-16 [60], ConvNext-B [2], and DocXClassifier-B [35], slightly outperformed the smaller CNNs and showed comparable performance to each other for the $\varepsilon_{target} = 5$ setting. However, for slightly less rigorous privacy constraints with $\varepsilon_{target} = 10$, it can be noticed that the ConvNext-B [2], and DocXClassifier-B [35] performed considerably better than all other models. In addition, it can be noticed that the DocXClassifier-B [35] model converged much faster in this setting with a much lower privacy loss of $\varepsilon = 6.7$ compared to other models. It is also noticeable that EfficientNet-B4 [61] performed considerably worse than the others for both $\varepsilon_{target} = 5$ and $\varepsilon_{target} = 10$ settings, possibly due to the substitution of BN layers with newly initialized GN layers. Notably, ResNet-50 [1], despite also undergoing BN layer replacement, seemed to be less impacted in terms of its performance. Finally, it can be observed that, despite achieving comparable performance to CNNs under non-private training, ViTs severely underperformed under both DP settings.

Results on Tobacco3482$_{ImageNet}$: In the Tobacco 3482$_{ImageNet}$ setting, a similar trend was observed, with ConvNext-B [2] and DocXClassifier-B [35] models exhibiting comparably better performance under both settings: $\varepsilon_{target} = 5$ and $\varepsilon_{target} = 10$. On the other hand, it is noticeable that the ResNet-50 [1] and EfficientNet-B4 [61] models performed extremely poorly in this scenario, again due to the substitution of the BN layers. Moreover, ViTs once again exhibited subpar performance in comparison to the CNNs under both DP configurations. Overall, the performance deterioration induced by DP was notably significant across all models in this setting, likely due to the extremely small size of the dataset.

Results on Tobacco3482$_{RVL\text{-}CDIP}$: In the Tobacco3482$_{RVL-CDIP}$ setting, we observed dramatic performance improvements across all models, primarily attributed to the document-specific pre-training. Notably, all DP models under both settings achieved performance significantly closer to that of the baseline models. The ResNet-50 [1] and EfficientNet-B4 [61] models also performed significantly better in this scenario compared to the Tobacco3482$_{ImageNet}$ setting. Meanwhile, the ConvNext-B [2] model again outperformed others, achieving an accuracy of 92.44% and 92.43% on the $\varepsilon_{target} = 5$ and $\varepsilon_{target} = 10$ settings, respectively. Consistent with the previous trends, ViTs lagged behind the



(a) $\varepsilon_{target} = 5$ (b) $\varepsilon_{target} = 10$

Fig. 3 Validation accuracy versus ε over the number of training epochs for each model in the Tobacco3482$_{RVL-CDIP}$ setting. It can be observed that ConvNext-B [2] and DocXClassifier-B [35] achieved significantly faster convergence compared to other models, leading to higher performances with lower privacy loss

CNNs in this setting as well. For the Tobacco3482$_{RVL-CDIP}$ setting, we also analyze the convergence of each model in terms of the validation accuracy and privacy loss ε obtained over the number of epochs, as illustrated in Fig. 3. As evident from the figure, in this setting, both the ConvNext-B [2] and DocXClassifier-B [35] models achieved convergence in the first few epochs, while the other models, including the ViTs, exhibited a much slower convergence on both $\varepsilon_{target} = 5$ and $\varepsilon_{target} = 10$ settings. This demonstrates that with domain-specific pre-training, unlike other models, the ConvNext-B [2] and DocXClassifier-B [35] models are capable of achieving higher performances even at much lower privacy loss ($\varepsilon \approx 1$).

4.5 Hyperparameter evaluation for differential privacy

4.5.1 Experimental setup

To determine the optimal set of hyperparameters for training the models under differential privacy (DP), we conducted a grid search over three crucial parameters: learning rate η, batch size B, and gradient clipping norm C. We initially performed the search exclusively on the ResNet-50 [1] model using both optimizers, DP-SGD and DP-Adam, across all three dataset settings: RVL-CDIP [34], Tobacco3482$_{RVL-CDIP}$, and Tobacco3482$_{ImageNet}$. Since the DP-SGD/Adam-based optimization is extremely computationally intensive, requiring hundreds of GPU hours for training on large datasets, for the RVL-CDIP [34] dataset, we opted to tune the hyperparameters on a smaller subset of 50,000 training samples in combination with early stopping to prune the ineffective training runs.

4.5.2 Experimental results

Throughout our tuning experiments, we noticed a significant dependence of the model performance on both the choice of the optimizer and the hyperparameters. Moreover, even with the same optimizer, the models sometimes demonstrated different trends across the three dataset settings: RVL-CDIP [34], Tobacco3482$_{RVL-CDIP}$, and Tobacco3482$_{ImageNet}$.

DP-Adam vs DP-SGD: In Fig. 4, we compare the performance of the DP-SGD and DP-Adam algorithms on the ResNet-50 model with different settings of learning rates (η) and batch sizes (B) and a fixed gradient norm of $C = 1.0$. Overall, it can be noticed that the Adam optimizer underperformed by a wide margin compared to the SGD optimizer for the different settings of learning rate η and batch size B on the ResNet-50 model. This trend was observed across all searched values of the gradient clipping norm $C \in \{0.1, 0.5, 1.0, 2.0, 5.0, 10.0\}$.

Learning Rate versus Batch size: We also conducted an analysis of the hyperparameters for each optimizer individually

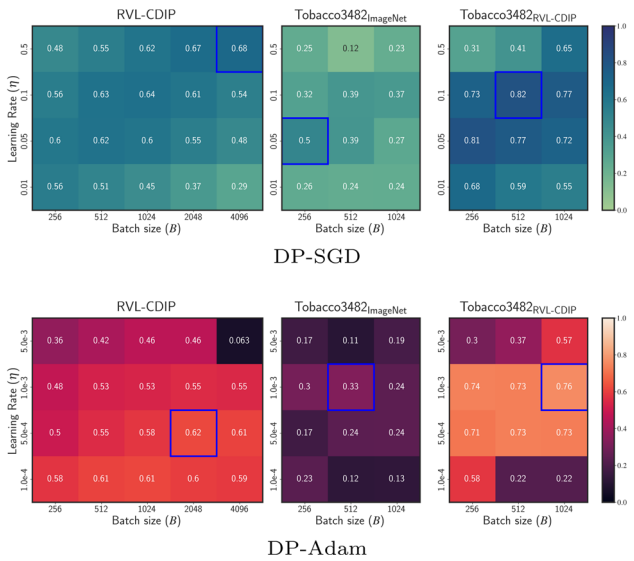


Fig. 4 Performance evaluation of the DP-SGD (top) and DP-Adam (bottom) algorithms on ResNet-50 [1] model with varying learning rates (η) and batch sizes (B) under a gradient clipping norm of $C = 1.0$. The highest performance achieved on each dataset is highlighted in blue. As shown, on ResNet-50 [1], DP-Adam severely under-performed compared to DP-SGD on all three dataset settings: RVL-CDIP, Tobacco3482$_{ImageNet}$ and Tobacco3482$_{RVL-CDIP}$ (color figure online)

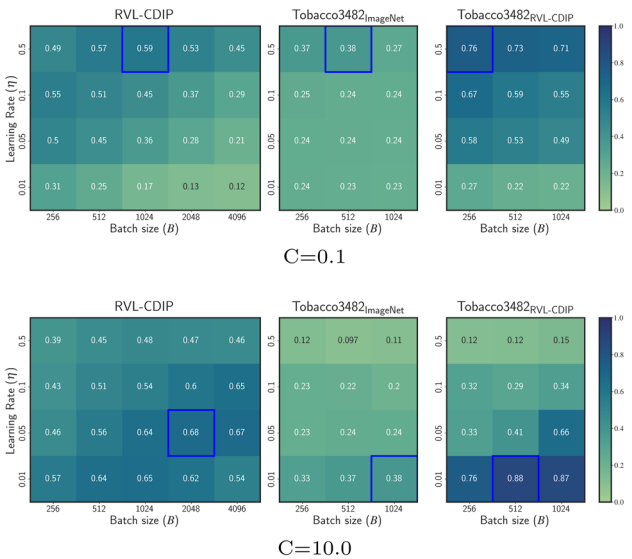


Fig. 5 Performance evaluation of the DP-SGD algorithm on ResNet-50 model with varying learning rates (η) and batch sizes (B) under two settings of gradient clipping norm, $C = 0.1$ (top) and $C = 10.0$ (bottom). The highest performance achieved on each dataset is highlighted in blue. As shown, larger batch sizes and clipping norms, coupled with smaller learning rates yielded better performance (color figure online)

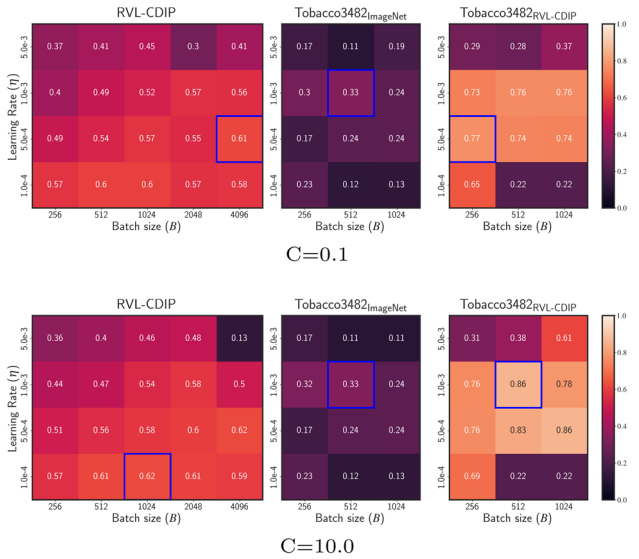


Fig. 6 Performance evaluation of the DP-Adam algorithm on ResNet-50 model with varying learning rates (η) and batch sizes (B) under two settings of gradient clipping norm, $C = 0.1$ and $C = 10.0$. The highest performance achieved on each dataset is highlighted in blue. As shown, for DP-Adam algorithm, larger batch sizes and smaller learning rates yielded better performance, regardless of the clipping norm (color figure online)

across multiple values of the gradient clipping norm C. In Fig. 5, we compare the performance of DP-SGD on ResNet-50 [1] between two settings of gradient clipping norm $C \in \{0.1, 10.0\}$. As evident from the figure, for DP-SGD, larger batch sizes (B) generally yielded better performance. Meanwhile, smaller learning rates (η) coupled with larger values of gradient clipping norm C or vice versa yielded better results. These trends remained consistent with the DP-SGD algorithm across all three dataset settings, RVL-CDIP [34], Tobacco3482$_{ImageNet}$, and Tobacco3482$_{RVL-CDIP}$. On the DP-Adam algorithm, we observed a slightly different trend, with larger batch sizes (B) coupled with smaller learning rates (η) yielding better results, irrespective of the value of the gradient clipping norm C. This behavior can be observed in Fig. 6, where we compare the performance of DP-Adam on ResNet-50 [1] under two settings of gradient clipping norm $C \in \{0.1, 10.0\}$.

Variable Impact of Hyperparameters Across Models: Based on the analysis of hyperparameters on the ResNet-50 [1] model, we selected the most suitable parameters for training the model under DP and maintained the same configuration for other models. However, although the hyperparameters derived from ResNet-50 [1] performed well for private training of traditional CNN architectures, including AlexNet [59], VGG-16 [60], and EfficientNet-B4 [61], we did not find them to be the most suitable for modern architectures, including, ConvNext-B [2], DocXClassifier-B [35], and ViTs (ViT-B/16 [50] and ViT-L/32 [50]). For these models, we observed that DP-Adam generally produced significantly better results compared to DP-SGD and also exhibited different performance trends. Therefore, to determine the optimal training hyperparameters for these models, we conducted a separate grid-search for the ConvNext-B [2] and ViT-B/16 [50] models, following the same approach as done for ResNet-50 [1]. The optimal hyperparameter configurations derived for ConvNext-B [2] and ViT-B/16 [50] were then translated to DocXClassifier-B [35] and ViT-L/32 [50], respectively. In contrast to the behavior observed on ResNet-50 [1], for the ConvNext-B [2] model, we noticed a general trend where larger values of batch sizes (B), gradient clipping norms (C), and learning rates (η) yielded better results. Likewise, on ViTs, we observed that larger values of gradient clipping norms (C) and learning rates (η), coupled with moderate batch sizes (B), worked the best.

4.6 Evaluating FedAVG and FedENS

4.6.1 Experimental setup

To assess the overall feasibility of Federated Learning for private training, we carried out experiments with a single group of participants for the RVL-CDIP [34] dataset and with two groups for the Tobacco3482 datasets, comparing the performance between FedAVG and FedENS. To simulate the local data for each client, we randomly shuffled the original training dataset and then created N_c equal partitions from it, which were subsequently assigned to each client. To evaluate the FedAVG algorithm on RVL-CDIP [34], we employed the following configuration: The number of clients N_c was set to 8, clients sampled per round f_c to 0.25, local epochs N_{local} to 1, and total number of rounds N_R to 40. On the other hand, to evaluate the FedENS algorithm, all local models were separately trained for a total of N_{local} epochs (set to 40) and ensembled at the end. For the Tobacco3482 dataset, we experimented with two different settings; (A) $N_c = 2$, $f_c = 1.0$, $N_{local} = 1$, $N_R = 40$ and (B) $N_c = 4$, $f_c = 0.5$, $N_{local} = 1$, $N_R = 40$.

4.6.2 Experimental results

The results of these experiments are shown in Table 2. As shown in the table, for the RVL-CDIP [34] dataset, FedAVG significantly outperformed FedENS across all models while also ensuring strong privacy with a client sampling rate of $f_c = 0.25$. This is a noteworthy result as, with a factor $f_c = 0.25$, the local dataset partitions of all clients were not trained on for a total of 40 epochs, in contrast to FedENS, and yet FedAVG demonstrated superior performance. A possible explanation of this is that the model averaging in FedAVG has a regularization effect, reducing the effects of overfitting,

Table 2 Performance comparison of the baseline document image classification models with their private counterparts trained with FedAVG and FedENS algorithms under various client configurations

Dataset	Model	Approach	$\text{Acc}_{Baseline}$		Acc_{FedAVG}	Acc_{FedENS}
			Reported	Ours	$N_c = 8,\ f_c = 0.25$	$N_c = 8$
RVL-CDIP [34]	AlexNet [59]	Afzal et al. [5]	88.60	87.90	85.54	85.21
	VGG-16 [60]	Afzal et al. [5]	90.97	91.00	89.41	84.81
	ResNet-50 [1]	Afzal et al. [5]	90.40	90.50	88.25	84.22
	EfficientNet-B4 [61]	Ferrando et al. [6]	92.31	92.60	90.59	88.73
	ConvNext-B [2]	Saifullah et al. [35]	**94.04**	93.64	**92.60**	**90.82**
	DocXClassifier-B [35]	Saifullah et al. [35]	94.00	**93.74**	91.73	90.79
	ViT-B/16 [50]	Afzal et al. [5]	–	89.22	88.65	85.71
	ViT-L/32 [50]	Afzal et al. [5]	–	88.64	88.13	85.08

Dataset	Model	Approach	$\text{Acc}_{Baseline}$		Acc_{FedAVG}		Acc_{FedENS}	
			Reported	Ours	$(A) N_c = 2,\ f_c = 1.0$	$(B) N_c = 4,\ f_c = 0.5$	$(A) N_c = 2$	$(B) N_c = 4$
Tobacco3482 (ImageNet Pre-training)	AlexNet [59]	Afzal et al. [5]	75.73	79.28	79.43	76.71	77.99	77.77
	VGG-16 [60]	Afzal et al. [5]	77.52	82.14	75.71	68.86	80.14	75.42
	ResNet-50 [1]	Afzal et al. [5]	67.93	76.57	79.00	73.00	63.14	54.00
	EfficientNet-B4 [61]	Ferrando et al. [6]	85.99	78.57	**89.71**	**88.71**	**90.42**	86.42
	ConvNext-B [2]	Saifullah et al. [35]	–	**89.57**	89.14	**88.71**	89.71	**86.57**
	DocXClassifier-B [35]	Saifullah et al. [35]	**87.43**	87.43	86.86	82.14	85.00	82.28
	ViT-B/16 [50]	Afzal et al. [5]	–	80.57	80.57	81.00	79.57	78.14
	ViT-L/32 [50]	Afzal et al. [5]	–	82.57	80.14	78.57	77.42	76.14
Tobacco3482 (RVL-CDIP Pre-training)	AlexNet [59]	Afzal et al. [5]	90.04	89.57	91.57	91.43	89.71	87.42
	VGG-16 [60]	Afzal et al. [5]	91.01	94.14	94.14	93.86	93.71	91.42
	ResNet-50 [1]	Afzal et al. [5]	91.13	90.14	93.14	92.57	91.85	85.85
	EfficientNet-B4 [61]	Ferrando et al. [6]	94.04	94.04	94.43	94.14	94.42	93.71
	ConvNext-B [2]	Saifullah et al. [35]	**95.00**	**94.71**	**95.86**	**95.71**	94.42	**94.85**
	DocXClassifier-B [35]	Saifullah et al. [35]	**95.00**	**94.71**	95.29	95.00	**94.57**	93.99
	ViT-B/16 [50]	Afzal et al. [5]	–	90.00	88.71	88.71	87.57	84.42
	ViT-L/32 [50]	Afzal et al. [5]	–	88.71	87.85	88.85	87.28	85.00

The highest accuracy scores for each setting are bolded. It can be observed that federated learning (FL) approaches only incurred minor performance losses across all training setups

and, consequently, yielding better results. We observed similar performance trends on the Tobacco3482 dataset, where in the Tobacco3482$_{RVL-CDIP}$ setting, the models sometimes outperformed even the baseline models. Notably, the best performing model, ConvNext-B [2], achieved an accuracy of 95.86% and 95.71% in settings A and B of FedAVG, respectively. Moreover, we noticed a slight deterioration in performance across all models in setting B for both FedAVG and FedENS. However, since the number of participants in this scenario is twice as high as in setting A, it also ensures better privacy in comparison, and therefore, the performance loss is expected Finally, on the Tobacco3482$_{ImageNet}$ setting, we noticed that both FedAVG and FedENS achieved similar results across all models. Interestingly, both ConvNext-B [2] and EfficientNet-B4 [61] significantly outperformed other models in this scenario. Moreover, we observed that the ViTs demonstrated a consistent trend similar to the case of DP-based private training, performing sub-optimally compared to the CNNs in all three dataset settings: RVL-CDIP [34], Tobacco3482$_{ImageNet}$, and Tobacco3482$_{RVL-CDIP}$

4.7 Evaluating federated learning with differential privacy (FedAVG-DP)

4.7.1 Experimental setup

To thoroughly evaluate the combined setting of FL with DP, we investigated the same client configurations for each dataset setting: RVL-CDIP [34], Tobacco3482$_{ImageNet}$, and Tobacco3482$_{RVL-CDIP}$, as done for the FL experiments (outlined in Sect. 4.6). Moreover, to train each local model under DP, we searched for the noise multiplier σ based on the target privacy budgets of ($\varepsilon_{target} \in \{5, 10\}$,$\delta = \frac{1}{\|\mathcal{D}_{local}\|}$) over the total number of optimization steps T per client. Where, $\mathcal{D}_{local}$ denotes the local training dataset of each client, and T is defined as the product of the total number of local optimization steps per epoch, total number of local epochs N_{local}, total number of rounds N_R and the client sampling rate f_c. Since training the models under DP is significantly difficult compared to the standard FL training setup, we raised the number of local epochs N_{local} to 4 in this setting and simultaneously reduced the total number of FL rounds to 10 to maintain consistency with the total number of epochs in the standalone DP and FL setups. Moreover, the total number of rounds were increased by a factor of $\frac{1}{f_c}$ for each run to ensure a target privacy loss of ε_{target} is achieved on each client over the complete federated learning run. Finally, deviating slightly from the standalone FL setup in this scenario, we trained the models with a sampling rate of $f_c = 0.5$ instead of $f_c = 0.25$ on the RVL-CDIP [34] dataset.

4.7.2 Experimental results

The results of these experiments are shown in Table 3. Overall, we observed similar trends in this scenario to those found in the standalone DP case, with significant performance degradation introduced by DP across all models. For instance, on the RVL-CDIP [34] dataset, we noticed that the larger models, including VGG-16 [60], ConvNext-B [2], and DocXClassifier-B [35], outperformed others under the $\varepsilon_{target} = 5$ setting. Moreover, the performance of ConvNext-B [2] and DocXClassifier-B [35] showed a slight improvement under $\varepsilon_{target} = 10$. On the other hand, ResNet-50 [1] and EfficientNet-B4 [61] once again performed poorly, possibly due to the re-initialization of the normalization layers. Similarly, both ViTs exhibited relatively poor performance, consistent with the standalone DP setting.

On the Tobacco3482 dataset, once again, in the Tobacco3482$_{ImageNet}$ setting, severe degradation of performance was observed across all models, with ResNet-50 [1] and EfficientNet-B4 [61] failing to converge even with a larger epsilon, $\varepsilon_{target} = 10$. In addition, the performance degradation was further amplified as the number of clients were increased from $N_c = 2$ to $N_c = 4$. On the other hand, in the Tobacco3482$_{RVL-CDIP}$ setting, domain-specific pre-training once again led to dramatic performance improvements across the majority of models. Surprisingly, however, the ViTs in this scenario severely unperformed in comparison to the CNNs. Meanwhile, the DocXClassifier-B [35] model exhibited the best performance across different client settings and privacy levels (ε). Overall, the results indicate that Federated Averaging with Differential Privacy has promising potential, especially when combined with domain-specific pre-training.

4.8 Evaluating SMPC for model hiding

In this section, we present the results of our evaluation of CrypTen-based Secure Multi-Party Computation (SMPC) for model hiding.

4.8.1 Experimental setup

CrypTen, being a relatively recent framework, currently lacks support for all types of PyTorch layers. As a result, the majority of models in our evaluation set were found to be unsupported by CrypTen. For this experiment, therefore, we only evaluated the first three models-AlexNet, VGG-16, and ResNet-50-since these were easily supported by CrypTen. We kept a simple experimental setup for this scenario, in which we encrypt the baseline models, encrypt the test set samples, and simply perform encrypted inference. In addition, we only performed these experiments for the Tobacco3482$_{RVL-CDIP}$ dataset setting.

Table 3 Performance comparison of baseline document image classification models with their private counterparts trained with FedAVG-DP under two configurations of target privacy budget: $\varepsilon_{target} = 5$ and $\varepsilon_{target} = 10$

Dataset	Model	Approach	$\text{Acc}_{Baseline}$		$\text{Acc}_{FedAVG-DP}(\varepsilon_{target} = 5) / \varepsilon_{max}$	$\text{Acc}_{FedAVG-DP}(\varepsilon_{target} = 10)/\varepsilon_{max}$
			Reported	Ours	$N_c = 8,\ f_c = 0.5$	$N_c = 8,\ f_c = 0.5$
RVL-CDIP [34]	AlexNet [59]	Afzal et al. [5]	88.60	87.90	70.36 / 5.5	72.18 / 11.6
	VGG-16 [60]	Afzal et al. [5]	90.97	91.00	75.00 / 6.0	77.03 / 12.5
	ResNet-50 [1]	Afzal et al. [5]	90.40	90.50	70.41 / 5.9	72.46 / 11.2
	EfficientNet-B4 [61]	Ferrando et al. [6]	92.31	92.60	56.92 / 5.3	58.14 / 11.1
	ConvNext-B [2]	Saifullah et al. [35]	94.04	93.64	75.50 / 5.8	79.08 / 11.1
	DocXClassifier-B [35]	Saifullah et al. [35]	94.00	93.74	**78.13 / 5.6**	**80.41 / 12.2**
	ViT-B/16 [50]	Afzal et al. [5]	–	89.22	68.7 / 5.8	72.4 / 11.1
	ViT-L/32 [50]	Afzal et al. [5]	–	88.64	68.6 / 6.1	72.2 / 11.1

Dataset	Model	Approach	$\text{Acc}_{Baseline}$		$\text{Acc}_{FedAVG-DP}(\varepsilon_{target} = 5)/\varepsilon_{max}$		$\text{Acc}_{FedAVG-DP}(\varepsilon_{target} = 10)/\varepsilon_{max}$	
			Reported	Ours	$N_c = 2,\ f_c = 1.0$	$N_c = 4,\ f_c = 0.5$	$N_c = 2$	$N_c = 4$
Tobacco3482 (ImageNet Pre-training)	AlexNet [59]	Afzal et al. [5]	75.73	79.28	60.28 / 5.00	**59.42 / 5.41**	64.57 / 10.00	**65.14 / 10.80**
	VGG-16 [60]	Afzal et al. [5]	77.52	82.14	58.42 / 5.00	53.85 / 5.41	61.43 / 10.00	62.29 / 10.80
	ResNet-50 [1]	Afzal et al. [5]	67.93	76.57	27.85 / 5.00	30.00 / 5.41	31.57 / 10.00	35.86 / 10.80
	EfficientNet-B4 [61]	Ferrando et al. [6]	85.99	78.57	29.43 / 5.00	30.14 / 5.41	28.57 / 10.00	26.29 / 10.80
	ConvNext-B [2]	Saifullah et al. [35]	–	89.57	**63.00 / 5.00**	55.86 / 5.41	**67.43 / 10.00**	59.43 / 10.80
	DocXClassifier-B [35]	Saifullah et al. [35]	87.43	87.43	58.42 / 5.00	46.57 / 5.41	63.43 / 10.00	59.14 / 10.80
	ViT-B/16 [50]	Afzal et al. [5]	–	80.57	50.00 / 5.00	48.71 / 5.41	56.57 / 10.00	53.85 / 10.80
	ViT-L/32 [50]	Afzal et al. [5]	–	82.57	50.28 / 5.00	47.85 / 5.41	55.14 / 10.00	53.57 / 10.80
Tobacco3482 (RVL-CDIP Pre-training)	AlexNet [59]	Afzal et al. [5]	90.04	89.57	86.57 / 5.00	82.57 / 5.41	86.14 / 10.00	86.29 / 10.90
	VGG-16 [60]	Afzal et al. [5]	91.01	94.14	84.85 / 5.00	84.57 / 5.41	85.43 / 10.00	86.71 / 10.90
	ResNet-50 [1]	Afzal et al. [5]	91.13	90.14	89.00 / 5.00	**89.00 / 5.41**	89.14 / 10.00	89.43 / 10.90
	EfficientNet-B4 [61]	Ferrando et al. [6]	94.04	94.04	87.29 / 5.00	87.29 / 5.41	87.43 / 10.00	88.29 / 10.90
	ConvNext-B [2]	Saifullah et al. [35]	95.00	94.71	87.29 / 5.00	87.00 / 5.41	88.14 / 10.00	87.76 / 10.90
	DocXClassifier-B [35]	Saifullah et al. [35]	95.00	94.71	**91.14 / 5.00**	85.57 / 5.41	**90.71 / 10.00**	**91.57 / 10.90**
	ViT-B/16 [50]	Afzal et al. [5]	–	90.00	72.14 / 5.0	68.14 / 5.41	80.71 / 10.00	74.71 / 10.90
	ViT-L/32 [50]	Afzal et al. [5]	–	88.71	81.00 / 5.0	77.43 5.41	84.29 / 10.00	82.86 / 10.90

The highest accuracy scores for each setting are bolded. It can be noted that the FedAVG-DP setting exhibited performance trends highly comparable to those of the standalone DP setting, but with slightly inferior performance. Moreover, the performance of FedAVG-DP declined with an increasing number of clients, except for the $\text{Tobacco3482}_{RVL-CDIP}$ case

Table 4 Performance and inference time comparison of baseline document image classification models with their private counterparts encrypted using SMPC

Dataset	Model	$\text{Acc}_{Baseline}$		SMPC (Crypten)	
		Reported	Ours	Acc	Inference time Inc
Tobacco3482 (RVL-CDIP pre-training)	AlexNet	90.04	89.57	89.57	4×
	VGG-16	91.01	94.14	94.14	103×
	ResNet-50	91.13	90.14	90.14	32×

As shown, SMPC resulted in a significant increase in inference costs, up to a factor of 100

4.8.2 Experimental results

The results are shown in Table 4. For each model, we present the test set accuracy achieved under encrypted inference and the increase in inference time caused by the encryption compared to the baseline model. It is evident from the results that the model and data encryption essentially causes no performance loss. However, it does introduce a significant computational overhead. For instance, while the increase in inference time is tolerable for AlexNet [59], it continues to increase as the model complexity increases, reaching 32× and 103× times that of baseline for the ResNet-50 [1] and VGG-16 [60] models, respectively. This shows that while CrypTen and SMPC appear to have promising potential in that they do not incur performance degradation, there is still much work to be done to make them more efficient for practical use.

4.9 Evaluating privacy strength using membership inference attacks

In this section, we present the results of our experimental setup, in which we apply Membership Inference Attacks (MIA) to both non-private and private models, in order to quantitatively assess and compare the privacy strength of different privacy preservation methods investigated in this work.

4.9.1 Experimental setup

The details of the experimental setup are outlined as follows: we assume that a malicious adversary has query access to the target model, along with some samples from the original training dataset, and their objective is to determine whether a specific data sample was part of the model's training dataset or not. Formally, let f_{target} represent the target model, f_{att} denote an attack model, and $\mathcal{D}_{train}$ denote the training dataset on which the target model f_{target} was trained. Then, given a sample x_i, the attack model f_{att} attempts to ascertain whether the sample x_i was part of the training dataset $\mathcal{D}_{train}$. In this work, we explored the simplest form of attack, where the adversary first queries the target model f_{target} to obtain the loss (based on a given loss function $\mathcal{L}$) and prediction scores for each sample x_i. In our experiments, we chose $\mathcal{L}$ as the cross-entropy loss, typically used for multi-class classification problems. Subsequently, using the loss, prediction scores and the true class label of each sample x_i, the adversary then generates an input for the attack model as $x_{i,att} = [\mathcal{L}(f_{target}(x_i)), \mathcal{P}(f_{target}(x_i)), x_{i,label}]$ along with its target label $y_i = 1$ or $y_i = 0$, depending on whether the sample is a member of the original training dataset or not. We refer to the samples that were part of the training dataset as member samples, which form the member dataset $D_{mem} = \{x_{i,att}, y_i = 1\}$. An equal number of samples from the original test set D_{test} are extracted to form the non-member samples, yielding the non-member dataset $D_{non-mem} = \{x_{i,att}, y_i = 0\}$. The non-member and member datasets are then combined to produce the training dataset $D_{att} = D_{non-mem} \cup D_{mem}$ on which the attack model is trained.

In this work, we investigated the performance of MIA attacks on both the RVL-CDIP [34] and Tobacco3482 datasets, considering four different model types: non-private baseline model, DP model with $\varepsilon_{target} = 5$, FedAVG model with 4 clients, and FedAVG-DP model with $\varepsilon_{target} = 5$ and clients set to 4 and 8 for the Tobacco3482 and RVL-CDIP [34] datasets, respectively. To generate the attack dataset D_{att}, on RVL-CDIP [34], we randomly selected 40K samples from the training set and combined them with the 40K test set samples. Subsequently, we divided the D_{att} dataset into two subsets, allocating 50% for training the attack model and the remaining 50% for evaluating its performance. On the Tobacco3482 dataset, we focused solely on evaluating MIA attacks in the $\text{Tobacco3482}_{RVL-CDIP}$ setting due to its superior performance. In this scenario, the attack dataset D_{att} consisted of 700 randomly selected training samples and 700 test samples, with the same train/test split ratio as in the RVL-CDIP [34] case. We experimented with multiple types of attack models; however, we found the GradientBoosting classifier [64] to work the best in our experiments.

4.9.2 Experimental results

To evaluate and compare the performance of MIA on each target model, we plot the Receiver Operating Characteristic (ROC) curves of the corresponding attack models f_{att}, for the RVL-CDIP [34] and Tobacco3482 datasets in Fig. 7 and Fig. 8, respectively. In addition, we present the Area Under the ROC Curves (AUC), attack precision, and attack recall for each setting in Table 5. From the behavior of the ROC curves on the RVL-CDIP [34] dataset, it is evident that the non-private baseline models were considerably more vulnerable to the membership inference attacks. Notably, some models such as ResNet-50 [1], VGG-16 [60], ViT-B/16 [50], and ViT-L/32 [50] appeared to be particularly susceptible in non-private setting compared to others. It can also be observed that, while the FedAVG setting does allow for client-level data privacy, it only performed slightly better against the attack compared to the non-private baseline, sometimes even exhibiting a vulnerability level similar to the baseline models (such as in the case of ViT-B/16 [50] and ViT-L/32 [50]).

In a stark contrast, the DP-SGD/Adam and FedAVG-DP settings, on the other hand, ensured a very high level of privacy. Across all settings, it is evident that when differential privacy (DP) was applied to the models, the attack model f_{att} could only perform as well as a random classifier. Similar observations can be drawn from Table 5, where, for all target models, it is evident that the attack model f_{att} achieved significantly better performance in inferring sample membership for the non-private baseline models compared to the DP models. Finally, on the Tobacco3482$_{RVL-CDIP}$ setting, similar trends can be observed from the ROC curves (see Fig. 8), with the non-private baseline and FedAVG models exhibiting significantly higher vulnerability to the attack compared to the DP models. Notably, it can also be observed that the ViTs (ViT-B/16 [50] and ViT-L/32 [50]) appeared to be considerably more vulnerable overall compared to the CNNs in this setting.

4.10 Assessing the impact of privacy on model interpretability

4.10.1 Experimental setup

In this section, we briefly assess the impact of various privacy preservation methods on the interpretability of the model. For this analysis, we specifically select the DocXClassifier-B [35] model due to its property of being inherent interpretable and exclusively analyze the results on the Tobacco3482$_{RVL-CDIP}$ setting. In particular, we generate attribution maps for various document image samples under four different settings: non-private baseline, DP setting with $\varepsilon_{\text{target}} = 5$, FedAVG setting with 4 clients, and FedAVG-DP setting with $\varepsilon_{\text{target}} = 5$ and 4 clients.

4.10.2 Experimental results

The results are depicted in Fig. 9, where we visualize the image attribution maps generated by the DocXClassifier-B [35] model for two randomly selected samples from each of the 10 document categories in the Tobacco3482 dataset. It is evident from the figure that different privacy preservation methods led to drastically different attribution maps, indicating a significant change in the underlying focus of the model under different methods. It can be noticed that for most samples, the attributions produced under the DP-SGD/Adam setting were significantly noisier compared to other methods. On the other hand, the model under the FedAVG setting produced smoother and more concentrated maps, focusing on overall regions where the text is present. Finally, FedAVG-DP showed characteristics of both DP and FedAVG, with its maps also being noisy but tending to focus on crucial class-specific regions.

5 Discussion

In this section, we summarize the key observations from our study associated with different privacy methods and discuss their practical implications in the context of document image classification.

5.1 Privacy-utility tradeoffs of different privacy methods

Several key findings were revealed from our study regarding the privacy-utility tradeoff of various privacy methods when applied to document image classification. In the following sections, we discuss these findings to present a comprehensive overview of our results.

DP Provides Rigorous Privacy Guarantees at the Cost of Utility: Through our analysis of the effectiveness of various privacy methods against membership inference attacks (see Sect. 4.9), it became evident that DP, in both standalone and federated (DP-FL) settings, offers rigorous protection against privacy attacks. However, from the general performance trends of DP on various configurations (see Sect. 4.4.2 and Sect. 4.7.2), it was also observed that if not managed appropriately, DP can result in severe performance declines for document image classification, both in standalone and federated settings. Notably, this decline was more pronounced in federated settings (DP-FL) as compared to standalone settings, with its severity increased with an increase in the number of clients. However, the utilization of document-specific pre-training proved effective in mitigating these performance degradations. Therefore, when achieving high model utility is imperative and document-specific pretrain-

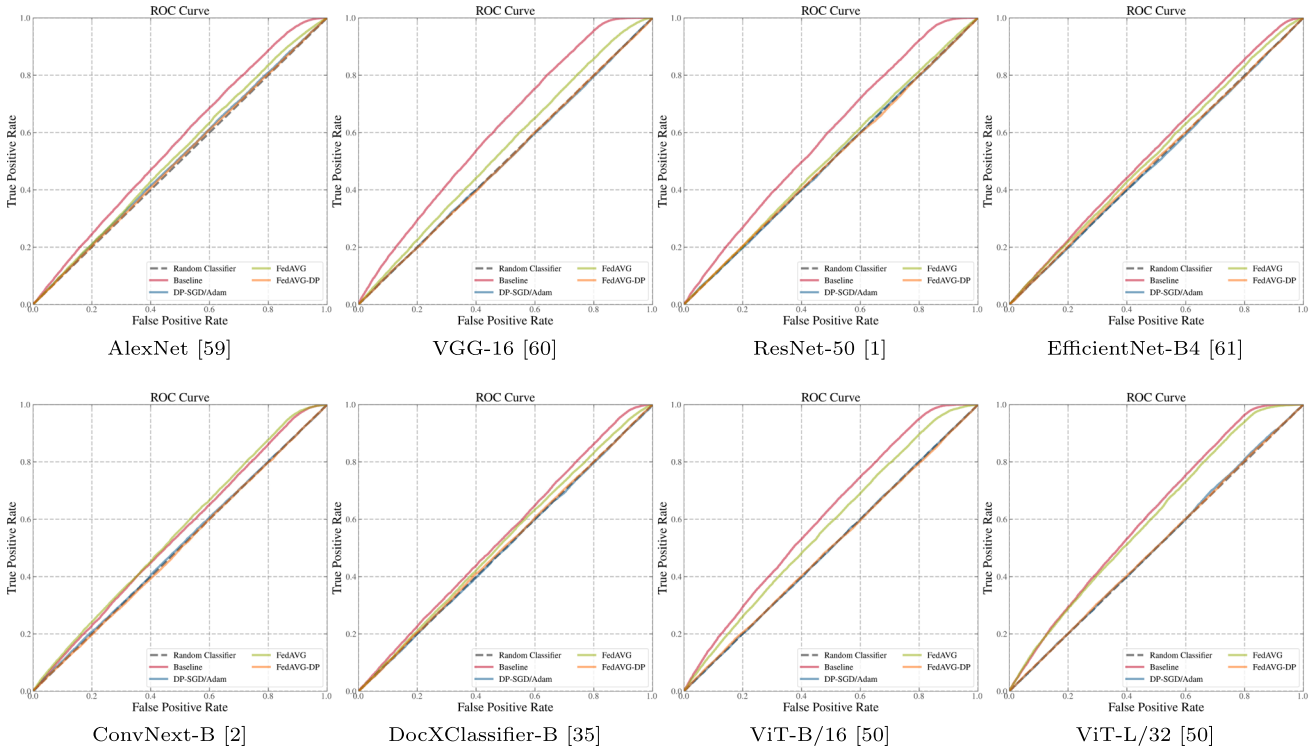


Fig. 7 ROC curves illustrating the classification performance of the MIA model f_{att} on each target model f_{target} for the RVL-CDIP [34] dataset. As shown, the attack classifier overall performed the worst on DP and FedAVG-DP approaches, whereas it showed the best performance on the non-private baseline models

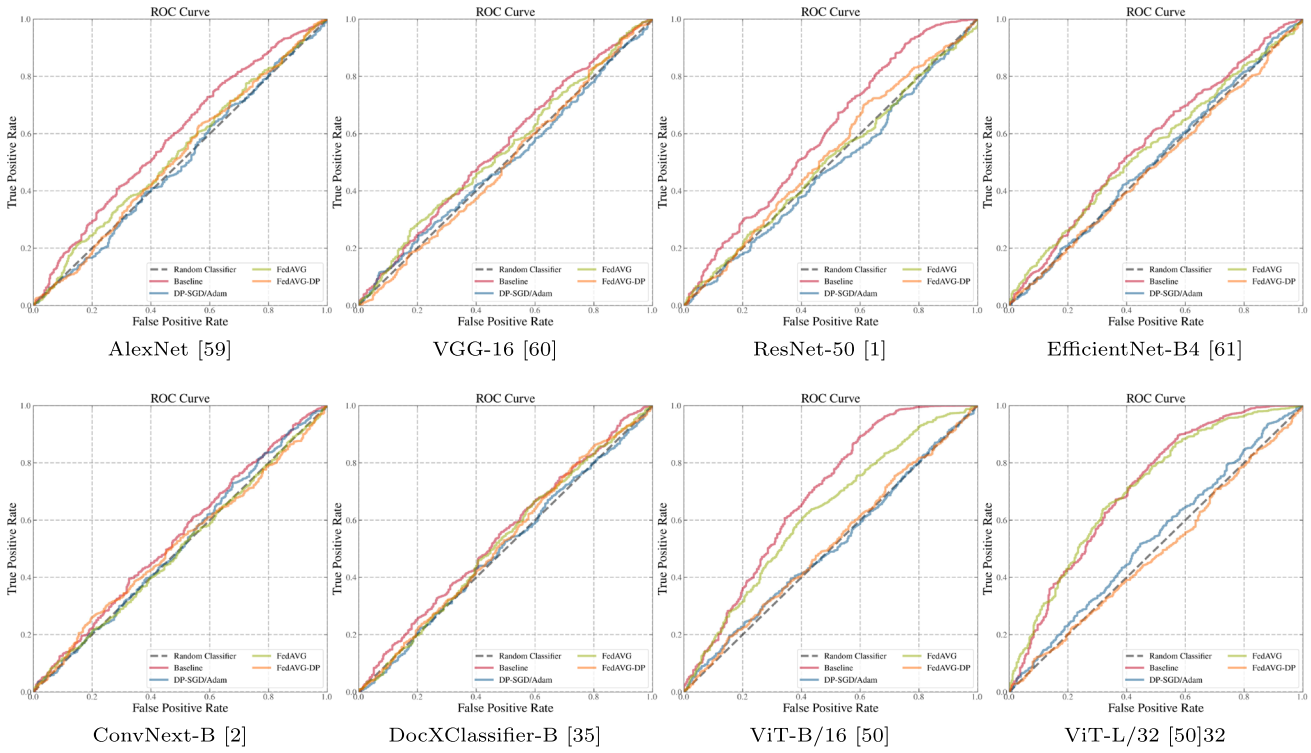


Fig. 8 ROC curves illustrating the classification performance of the MIA model f_{att} on each target model f_{target} for the Tobacco3482 dataset. As shown, the attack classifier overall performed the worst on DP and FedAVG-DP approaches, whereas it showed the best performance on the non-private baseline models

Table 5 AUC scores and accuracy of the membership inference attack model f_{att} for each target model under different privacy settings

Dataset	Model	Approach	Baseline			DP-SGD/Adam			FedAVG			FedAVG-DP		
			Prec	Recall	AUC	Prec	Recall	AUC	Prec	Recall	AUC	Prec	Recall	AUC
RVL-CDIP [34]	AlexNet [59]	Afzal et al. [5]	0.56	0.54	0.56	0.51	0.51	0.50	0.52	0.52	0.52	0.50	0.50	0.50
	VGG-16 [60]	Afzal et al. [5]	0.61	0.58	0.61	0.50	0.50	0.49	0.53	0.53	0.54	0.50	0.50	0.49
	ResNet-50 [1]	Afzal et al. [5]	0.59	0.56	0.59	0.50	0.50	0.49	0.51	0.51	0.51	0.50	0.50	0.49
	EfficientNet-B4 [61]	Ferrando et al. [6]	0.54	0.53	0.54	0.50	0.50	0.49	0.52	0.52	0.52	0.50	0.50	0.50
	ConvNext-B [2]	Saifullah et al. [35]	0.54	0.53	0.54	0.50	0.50	0.50	0.58	0.54	0.56	0.50	0.50	0.50
	DocXClassifier-B [35]	Saifullah et al. [35]	0.54	0.53	0.54	0.50	0.50	0.50	0.52	0.52	0.52	0.50	0.50	0.50
	ViT-B/16 [50]	Afzal et al. [5]	0.60	0.58	0.61	0.50	0.50	0.50	0.56	0.55	0.57	0.50	0.50	0.50
	ViT-L/32 [50]	Afzal et al. [5]	0.63	0.58	0.61	0.50	0.50	0.50	0.61	0.57	0.57	0.60	0.50	0.50
		Avg. =	0.58	0.55	0.58	0.50	0.50	0.50	0.54	0.53	0.54	0.51	0.50	0.50
Tobacco3482 (RVL-CDIP pre-training)	AlexNet [59]	Afzal et al. [5]	0.53	0.55	0.59	0.49	0.49	0.49	0.51	0.52	0.53	0.51	0.51	0.51
	VGG-16 [60]	Afzal et al. [5]	0.51	0.53	0.55	0.50	0.49	0.50	0.51	0.52	0.54	0.50	0.50	0.49
	ResNet-50 [1]	Afzal et al. [5]	0.53	0.56	0.60	0.49	0.49	0.48	0.50	0.51	0.50	0.51	0.52	0.52
	EfficientNet-B4 [61]	Ferrando et al. [6]	0.53	0.56	0.57	0.50	0.51	0.51	0.52	0.55	0.55	0.50	0.49	0.49
	ConvNext-B [2]	Saifullah et al. [35]	0.51	0.52	0.54	0.50	0.50	0.51	0.50	0.50	0.49	0.50	0.51	0.51
	DocXClassifier-B [35]	Saifullah et al. [35]	0.51	0.53	0.55	0.50	0.50	0.50	0.51	0.53	0.52	0.51	0.51	0.52
	ViT-B/16 [50]	Afzal et al. [5]	0.56	0.63	0.68	0.50	0.49	0.50	0.55	0.60	0.62	0.50	0.50	0.51
	ViT-L/32 [50]	Afzal et al. [5]	0.57	0.65	0.70	0.51	0.53	0.53	0.58	0.66	0.70	0.49	0.48	0.48
		Avg. =	0.53	0.57	0.60	0.50	0.50	0.50	0.52	0.55	0.56	0.50	0.50	0.50

As shown, the attack classifier overall performed the worst on DP and FedAVG-DP approaches, whereas it showed the best performance on the non-private baseline models

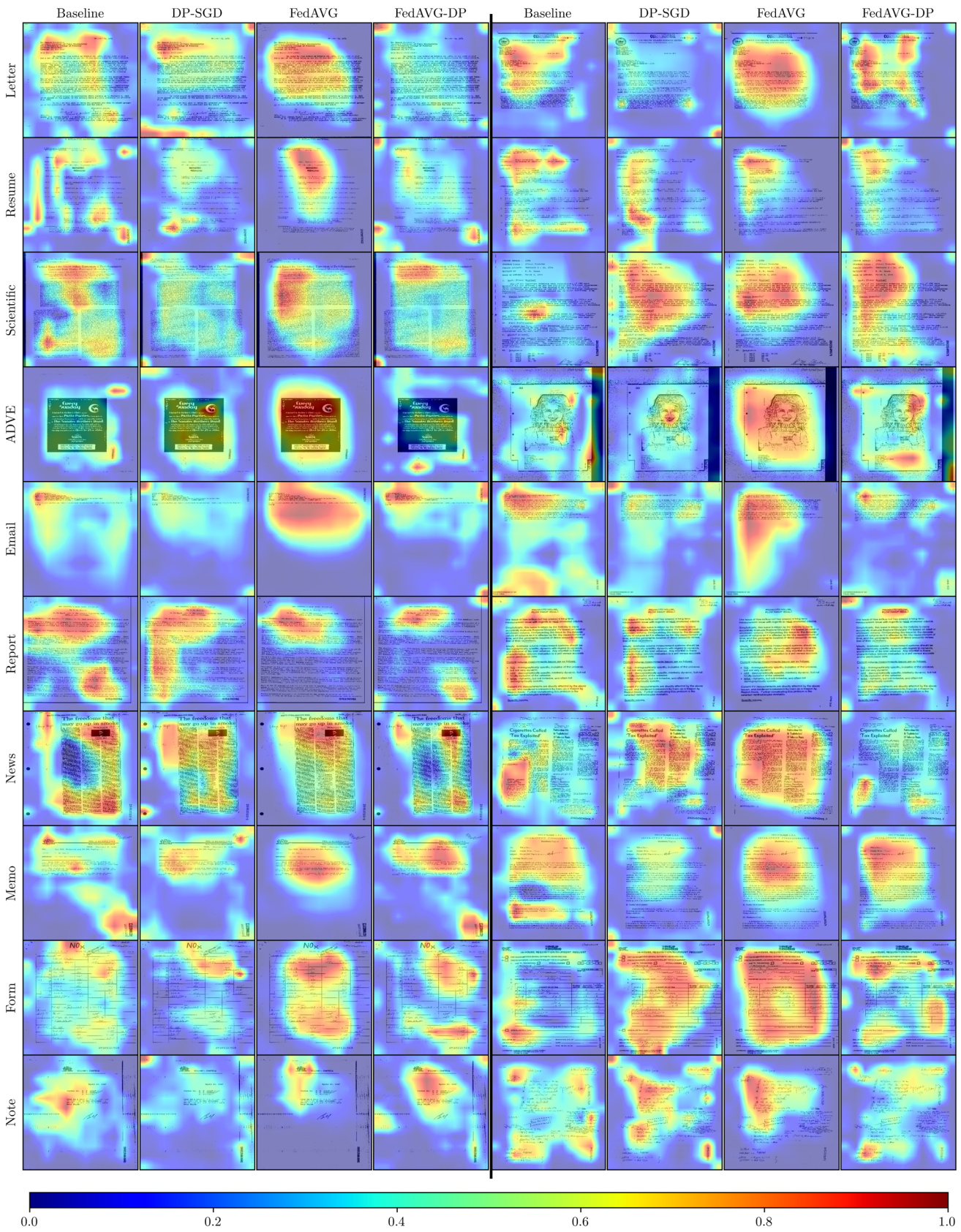


Fig. 9 Attention heatmaps generated by the DocXClassifier model on the Tobacco3482 dataset. For each of the 10 dataset classes (rows), heatmaps for two randomly selected samples (columns) are shown. It can be observed that when the model was trained under DP and FedAVG-DP settings, it produced noisier heatmaps compared to the FedAVG setting. This demonstrates a notable shift in the underlying focus of the model under different privacy approaches

ing is not feasible, DP may not be the most viable option, especially when dealing with a large number of clients in federated settings. However, in cases where privacy is of utmost importance, the utilization of DP is essential to prevent any leak of private information about the models or the training datasets.

FL Provides Better Utility but Insufficient Protection: In a striking contrast to DP, the Federated Learning (FL) approaches, FedAVG and FedENS, incurred only minor performance losses across all configurations, and this trend remained consistent across the models, irrespective of whether the document-specific pre-training was utilized. In some instances, we observed that FedAVG even led to an improvement in model performance compared to the baseline setting, potentially due to the regularization effect introduced by averaging the models after each training round. However, from our analysis in Sect. 4.9, it was quite evident that FL-based approaches fail to provide sufficient protection against membership inference attacks. These findings suggest that, as long as strong privacy guarantees are not required, FL approaches may be effectively applied for distributed training of document image classifiers while ensuring the privacy of local client data.

SMPC Enables Encrypted Information Sharing but at High Computational Costs: CrypTen-based model-hiding, which was investigated in this study, presented several advantages and disadvantages of its own. As this approach simply encrypts the data and models, unlike other methods, SMPC required no additional training, allowing users to use existing document classification models for inference in a private manner. Therefore, in practical scenarios, SMPC can be an effective approach for sharing data and models between multiple parties while keeping the local data of each party private. However, it must be noted that SMPC provides a similar level of protection to FL approaches, meaning it does not guarantee protection against privacy attacks. For instance, the membership inference attacks could still be applied to the model once its output has been revealed after decryption. Because of its simplicity, we believe SMPC can be a promising approach for introducing privacy into existing document analysis pipelines. However, unlike other privacy strategies, the current implementations of SMPC come with considerable inference costs, even when GPUs are utilized. Therefore, we believe further improvements are necessary to make this approach viable for practical applications.

5.2 Analyzing the factors impacting DP/DP-FL utility

Training document image classification models under DP was found to be particularly challenging, as the model utility under DP depends on several factors. In the following, we discuss these factors and suggest recommendations for achieving optimal performance.

Pre-training Significantly Improves Model Utility: Since Differential Privacy (DP) can severely hinder the convergence of deep learning models during training [13, 14, 22], we found that model pre-training (both in and out of domain) helped significantly in improving the performance of the models under DP. In Sect. 4.4.2, we observed that the models with reinitialized normalization (BN) layers, specifically EfficientNet-B4 [61] and ResNet-50 [1], performed significantly worse compared to other models that were fine-tuned directly from unaltered ImageNet weights. This suggests that the reinitialization of these layers essentially caused the models to be trained from scratch, whereas ImageNet pre-training contributed towards improving model convergence under DP. Similarly, in the Tobacco3482$_{RVL-CDIP}$ setting, we observed dramatic performance improvements across all models when document-specific pre-training was utilized. Based on these findings, our general recommendation is that whenever DP is applied to private sensitive document datasets, pre-training the models using large publicly available datasets should be considered to achieve an optimal privacy-utility trade-off.

Modern CNNs Designs Help Improve Utility: One important finding was that across different dataset settings, modern CNN architectures like ConvNext-B [2] and DocXClassifier-B [35] outperformed the previous state-of-the-art CNN models such as ResNet-50 [1] or EfficientNet-B4 [61]. Additionally, they demonstrated faster convergence compared to other models, achieving both better privacy and higher utility. Therefore, our recommendation is to prefer recent models over older architectures to achieve a better privacy-utility trade-off.

ViTs Significantly Under-perform compared to CNNs: Another noteworthy observation was that the ViTs [50] were significantly more prone to overfitting compared to the CNNs, resulting in relatively poorer performances under DP. This tendency was observed across all settings and was especially visible in the Tobacco3482$_{RVL-CDIP}$ setting, where even with document-specific pretraining, the ViTs [50] failed to achieve comparable performance to the CNNs. In the future, it will be worthwhile to explore self-supervised pretraining of ViTs [50] on large public datasets before fine-tuning them under DP to enhance their utility.

Tuning Hyperparameters is Essential for Achieving Optimal Utility: DP introduces additional hyperparameters such as noise scale σ, expected batch size B, and gradient clipping norm C, all of which were observed to have variable effects on model performance. As a result, these parameters

must be specifically tuned to find an optimal compromise between privacy and performance. We recommend tuning the three parameters-learning rate η, batch size B, and gradient clipping norm C-in combination with a fixed number of training epochs and a pre-defined target privacy budget ε_{target} that automatically determines the noise scale σ. Since DP-SGD/Adam also introduces significantly higher training costs in terms of GPU memory and compute power, we recommend tuning these hyperparameters first on a smaller subset of samples before training the models on large datasets.

5.3 Implementation challenges of different privacy methods

We encountered several practical difficulties when implementing the different privacy methods, which are briefly discussed below.

Constraints on Training Routines and Model Architectures for DP/DP-FL: First, by definition, differential privacy (DP) restricts the use of batch normalization layers in deep learning models. This limitation makes it impossible to utilize existing pre-trained models that incorporate BN layers. Additionally, differential privacy (DP) prohibits the use of stratified sampling during the training process, a method often employed to address data imbalance issues. This constraint can lead to additional performance degradation when dealing with imbalanced datasets.

High Training Costs of DP/DP-FL: Training deep learning models with DP-SGD/Adam generally requires an extensive amount of GPU memory and processing power, especially with increasing model size and image resolution. In addition, the current implementation of DP-SGD/Adam in Pytorch Opacus [57] does not support mixed-precision training, which adds an additional overhead on training time. As a result, specific GPUs with exceptionally high virtual memory were necessary to train the models under DP. In addition, for training larger models independently or in a federated environment, a distributed training setup was necessary to achieve feasible training times. In this work, we utilized the NVIDIA A100-40GB GPU, which offers 40GB of total virtual memory, for all our DP-related experiments.

Limited Support of SMPC across Model Architectures: While SMPC offers a straightforward means of providing encryption-based privacy, we faced challenges when applying it to a broad range of models. Specifically, we noted that the current implementation of this method is quite limited and does not support complex model layers. This makes it difficult to apply the approach to more recent architectures such as ConvNext-B [2], DocXClassifier-B [35], and ViTs [50].

5.4 Privacy-interpretability tradeoff: practical implications

Our results in Sect. 4.10.2, where we qualitatively assess the impact of various privacy methods on the interpretability of DocXClassifier, also led to some important findings. In particular, our observations revealed that when the model was trained under DP, it resulted in significantly noisier attribution maps, which shows that DP negatively impacted the interpretability of the model. In contrast, within the FL setting, the model produced notably smoother maps compared to the baseline, which slightly improved their interpretability. However, it occasionally over-smoothed, failing to concretely highlight the most important areas. Compared to the two approaches, the FL-DP setting offered a middle ground, where the noise from DP and the smoothing effect from FL balanced out each other, resulting in slightly improved interpretability compared to the standalone DP case.

Based on these observations, we recommend that in cases where model interpretability is crucial, FL approaches may be prioritized over DP to achieve some degree of privacy while also ensuring enhanced interpretability. However, in cases where DP is necessary for privacy concerns, we recommend combining it with FedAVG to achieve slightly improved interpretability while maintaining similar privacy levels. Overall, our findings suggest that enhancing the model's privacy may lead to a potential trade-off with its interpretability. Therefore, exploring newer approaches in the future to achieve an optimal balance for this trade-off would be worthwhile.

5.5 Advancing secure AI: the broader impact of this study

Despite facing a number of practical challenges in implementing differential privacy for document image classification, we succeeded in achieving sufficiently high performances on both large and small datasets, while maintaining robust privacy guarantees ($\varepsilon = 5$). Furthermore, through a comprehensive evaluation of the different privacy approaches under various configurations and hyperparameters, we were able to establish clear guidelines for achieving the most optimal privacy-utility tradeoffs in different scenarios. With these guidelines, we believe that our work will significantly accelerate the adoption of private document classification models in real-world applications, thus advancing the development of more secure and fortified automated document analysis pipelines that meet the standards of modern regulatory guidelines, including the GDPR [21] and the AI Act 2022.

Additionally, while our research primarily focuses on document image classification, there exist several other tasks in the document domain, such as table detection and recognition, layout analysis, and handwritten text recognition, all of which have the potential to leak critical private information. Yet, in our review of the existing literature, we noticed that while there is a significant amount of research focusing on privacy for textual document analysis tasks, not much emphasis was placed on the visual tasks. Therefore, we anticipate that our work will highlight the importance of privacy in the domain of visual document analysis and encourage further research in this area.

6 Conclusion

In this study, we conducted a comprehensive evaluation of well-known privacy-preserving methods in the context of document image classification. Our findings reveal that the application of these methods results in varying degrees of performance loss, influenced by various factors such as model architecture, size of the dataset, weight initialization, and training hyperparameters. Notably, our results demonstrate that, with sufficient hyperparameter tuning, differential privacy (DP) can achieve satisfactory utility in both standalone and collaborative learning settings, while simultaneously ensuring rigorous privacy guarantees. In addition, significant performance boosts can be achieved through domain-specific pretraining, making it preferable in most scenarios. On the other hand, our results demonstrate that while federated learning-based approaches incur only a marginal loss in performance on the task and introduce client-level data privacy, these approaches fail to provide sufficient protection against sophisticated privacy attacks. Finally, encryption-based methods also showed promise in providing privacy but the significant inference costs of their current implementations make them impractical for this task. To the best of the authors' knowledge, our work is the first that comprehensively explores modern privacy approaches in the domain of document image classification, paving the way for integrating privacy into modern automated document analysis pipelines.

Appendix A

A.1 Privacy accounting

To account for privacy loss (ε), we used the Rényi DP [54] privacy accountant in our work. However, there also exist other privacy accountants, such as, Gaussian DP [53], or Private Random Variable (PRV) Accountant [55]. Generally, all these accountants are an improvement over the moments accountant proposed by Abadi et al. [22], however, we have used RDP [54] in this work since it is not only the most widely used accountant [13, 14, 30], but also provides a strict upper bound over the privacy loss.

Given the data sampling rate q, a given noise multiplier σ and target privacy budget (ε, δ), RDP can be used to estimate the overall privacy loss over a fixed number of training steps. Numerical optimization can then be used in combination with the RDP estimation to obtain a suitable value of σ for a fixed target privacy budget (ε, δ). In this work, we perform this numerical optimization for all DP-related experiments to compute the required noise multiplier for the target epsilon ε_{target}, given a total number of training epochs.

A.2 Privacy algorithms

A.2.1 DP-SGD/Adam

In this work, we employed both DP-SGD and DP-Adam to train the target models under differential privacy (DP). For the sake of completeness, the pseudocode for both algorithms is included in Algorithm 1.

Input: $\mathcal{L}(\theta) = \frac{1}{B}\sum_i \mathcal{L}(\theta, x_i)$, Dataset $\mathcal{D} = (x_1, y_1), \ldots, (x_N, y_N)$, learning rate η, gradient clipping norm C, noise scale σ, sampling rate q, target (ε, δ), privacy accountant $\mathcal{M}$, total training steps T
Init: Initialize θ_0 randomly
for *each step* $t = 1, \ldots, T$ **do**
  $\mathcal{B} \leftarrow$ (sample a batch of size B with sampling probability q)
  for each $x_i \in \mathcal{B}$ **compute**
    // Compute gradient
    $\mathbf{g}(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$
    // Clip gradient
    $\tilde{\mathbf{g}}(x_i) \leftarrow \mathbf{g}(x_i)/max(1, \frac{||\tilde{\mathbf{g}}(x_i)||_2}{C})$
  end
  // Add noise
  $\tilde{\mathbf{g}} \leftarrow \frac{1}{B}(\sum_i \tilde{\mathbf{g}}(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})$
  if *Algorithm is DP-SGD* **then**
    // Call SGD Update
    $\theta_{t+1} \leftarrow \theta_t - \eta\tilde{\mathbf{g}}$
  else if *Algorithm is DP-Adam* **then**
    // Call Adam Update
    $m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1)\tilde{\mathbf{g}}$
    $v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2)\tilde{\mathbf{g}}^2$
    $m_t \leftarrow \frac{m_t}{1-\beta_1^t}$
    $v_t \leftarrow \frac{v_t}{1-\beta_2^t}$
    $\theta_{t+1} \leftarrow \theta_t + \eta\frac{m_t}{\sqrt{v_t}+\tau}$
  end
  print $\mathcal{M}$.get_privacy_spent(q, σ, t, δ)
end

Algorithm 1: DP-SGD/Adam

A.2.2 Federated learning algorithms

The pseudocodes for the FedAVG, FedENS, and FedAVG-DP algorithms are given in Algorithm 2, Algorithm 4, and Algorithm 3, respectively.

Input: Learning rate η, total clients N_c, clients sampling rate f_c, total federated learning rounds N_R
Server Executes:
  Init: θ_0, $m \leftarrow f_c N_c$
  for *each round* $r = 1, \ldots, N_R$ **do**
    $\mathcal{S}_r \leftarrow$ (sample a set of m clients from N_c)
    for *each client* $k \in \mathcal{S}$ **in parallel do**
      $\theta_{k,r} \leftarrow \text{ClientUpdate}(k, \theta_{r-1})$
    end
    $\theta_r \leftarrow \sum_{k \in \mathcal{S}_r} \frac{n_k}{n} \theta_{k,r}$
  end
end
ClientUpdate (k, θ)**:**
  Input: $\mathcal{L}(\theta) = \frac{1}{B} \sum_i \mathcal{L}(\theta, x_i)$, $\mathcal{D}_k$ of size $\|\mathcal{D}_k\|$
  $\mathcal{B} \leftarrow$ (sample a batch of size B)
  for *each epoch* $e = 1, \ldots, N_{local}$ **do**
    for each $b \in \mathcal{B}$ **compute**
      if *Optimizer is SGD* **then**
        // Call SGD Update
        $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta, b)$
      else if *Optimizer is Adam* **then**
        // Call Adam Update
        $\tilde{\mathbf{g}} \leftarrow \nabla_\theta \mathcal{L}(\theta, b)$
        $m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1)\tilde{\mathbf{g}}$
        $v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2)\tilde{\mathbf{g}}^2$
        $m_t \leftarrow \frac{m_t}{1-\beta_1^t}$
        $v_t \leftarrow \frac{v_t}{1-\beta_2^t}$
        $\theta \leftarrow \theta - \eta \frac{m_t}{\sqrt{v_t}+\tau}$
      end
    end
  end
return θ

Algorithm 2: FedAVG

Input: Learning rate η, total clients K, clients sampling rate C, total FL rounds T
Server Executes:
  Init: θ_0, $m \leftarrow C K$
  for *each round* $t = 1, \ldots, T$ **do**
    $\mathcal{S}_t \leftarrow$ (sample a set of m clients from K)
    for *each client* $k \in \mathcal{S}$ **in parallel do**
      $\theta_{k,t} \leftarrow \text{ClientUpdate}(k, \theta_{t-1})$
    end
    $\theta_t \leftarrow \sum_{k \in \mathcal{S}_t} \frac{n_k}{n} \theta_{k,t}$
  end
end
ClientUpdate (k, θ)**:**
  Input: $\mathcal{L}(\theta) = \frac{1}{B} \sum_i \mathcal{L}(\theta, x_i)$, $\mathcal{D}_k$ of size $\|\mathcal{D}_k\|$
  $\theta \leftarrow$ DP-SGD$(\mathcal{L}(\theta), \mathcal{D}_k)$ or DP-Adam$(\mathcal{L}(\theta), \mathcal{D}_k)$
return θ

Algorithm 3: FedAVG-DP

Input: Learning rate η, total clients N_c
Server Executes:
  Init: θ_0
  $\mathcal{S} \leftarrow$ (get the set of all N_c clients)
  for *each client* $k \in \mathcal{S}$ **in parallel do**
    $\theta_k \leftarrow \text{ClientUpdate}(k, \theta_0)$
  end
  Evaluate model ensemble $\{\theta_1, \ldots, \theta_{N_c}\}$ on test set
end
ClientUpdate (k, θ)**:**
  Input: $\mathcal{L}(\theta) = \frac{1}{B} \sum_i \mathcal{L}(\theta, x_i)$, $\mathcal{D}_k$ of size $\|\mathcal{D}_k\|$
  $\mathcal{B} \leftarrow$ (sample a batch of size B)
  for *each epoch* $e = 1, \ldots, N_{local}$ **do**
    for each $b \in \mathcal{B}$ **compute**
      if *Optimizer is SGD* **then**
        // Call SGD Update
        $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta, b)$
      else if *Optimizer is Adam* **then**
        // Call Adam Update
        $\tilde{\mathbf{g}} \leftarrow \nabla_\theta \mathcal{L}(\theta, b)$
        $m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1)\tilde{\mathbf{g}}$
        $v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2)\tilde{\mathbf{g}}^2$
        $m_t \leftarrow \frac{m_t}{1-\beta_1^t}$
        $v_t \leftarrow \frac{v_t}{1-\beta_2^t}$
        $\theta \leftarrow \theta - \eta \frac{m_t}{\sqrt{v_t}+\tau}$
      end
    end
  end
return θ

Algorithm 4: FedENS

A.3 Hyperparameter evaluation

A.3.1 Default hyperparameters

The default set of hyperparameters used in this work are presented in Table 6

Table 6 Full list of default hyperparameters that were used in our study

Parameter	Value
Privacy budget (ε, δ)	$(\{5, 10\}, 1/\|\mathcal{D}_{train}\|)$
Gradient clipping norm (C)	10.0
Optimizer	SGD
Learning rate (η)	0.05
Learning rate decay	False
Epochs (E)	40
Weight decay (λ)	0
Noise multiplier (σ)	Computed such that privacy budget (ε, δ) is spent after E epochs
Total federated rounds (N_R)	40 for FL; $\frac{1}{q}$40 for FedAVG-DP

Table 7 Training hyperparameters that were used to fine-tune the models under DP and FedAVG-DP settings

Dataset	Parameters	ResNet-50	ConvNeXt	ViT
RVL-CDIP	Optimizer	SGD	Adam	Adam
	Batch size (B)	2048	4095	4095
	Clipping norm (S)	10.0	10.0	2.0
	Learning rate (η)	0.05	$5.0e-3$	$1.0e-3$
Tobacco3482$_{ImageNet}$	Optimizer	SGD	Adam	Adam
	Batch size (B)	512	1024	256
	Clipping norm (S)	10.0	10.0	10.0
	Learning rate (η)	0.01	$5.0e-3$	$1.0e-3$
Tobacco3482$_{RVL-CDIP}$	Optimizer	SGD	Adam	Adam
	Batch size (B)	512	256	256
	Clipping norm (S)	10.0	10.0	10.0
	Learning rate (η)	0.01	$5.0e-3$	$1.0e-3$

A.3.2 Tuned hyperparameters

The hyperparameters that were selected for each dataset for different private settings are listed in Table 7.

Author Contributions All authors contributed to the study conception and design. Data collection and analysis were performed by Saifullah Saifullah. The first draft of the manuscript was written by Saifullah Saifullah and Dominique Mercier, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data availability The datasets analysed during the current study are available in the following repositories: RVL-CDIP can be accessed via the official project page (https://adamharley.com/rvl-cdip/). Tobacco3482 can be accessed via Kaggle (https://www.kaggle.com/datasets/patrickaudriaz/tobacco3482jpg).

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2015). https://api.semanticscholar.org/CorpusID:206594692
2. Liu, Z., et al.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986 (2022)
3. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates Inc., New York (2017)
4. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Association for Computational Linguistics, Minneapolis, pp. 4171–4186 (2019). https://aclanthology.org/N19-1423
5. Afzal, M.Z., Kolsch, A., Ahmed, S., Liwicki, M.: Cutting the error by half: investigation of very deep CNN and advanced training strategies for document image classification. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 883–888 (2017). arXiv:1704.03557
6. Ferrando, J., et al.: Improving accuracy and speeding up document image classification through parallel systems. In: Computational Science-ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part II 20, 12138 LNCS, pp. 387–400 (2020). arXiv:2006.09141
7. Powalski Rafał Borchmann, Ł., Jurkiewicz, D., Dwojak, T., Pietruszka Michał Pałka, G., Lladós, J., Lopresti, D., Uchida, S.: Going full-TILT boogie on document understanding with text-image-layout transformer. In: Lladós, J., Lopresti, D. Uchida, S. (eds.) Document Analysis and Recognition–ICDAR 2021, Springer International Publishing, Cham (2021)
8. Lee, C.Y., et al.: FormNet: structural encoding beyond sequential modeling in form document information extraction. vol. 1, pp. 3735–3754 (Long Papers, 2022). arXiv:2203.08411
9. Shen, Z., et al.: Layoutparser: a unified toolkit for deep learning based document image analysis. In: Lladós, J., Lopresti, D., Uchida, S. (eds) Document Analysis and Recognition–ICDAR 2021, pp. 131–146. Springer International Publishing, Cham (2021)
10. Al-Rubaie, M., Chang, J.M.: Privacy-preserving machine learning: threats and solutions. IEEE Secur. Priv. **17**(2), 49–58 (2019). arXiv:1804.11238
11. Zhang, D., Chen, X., Wang, D., Shi, J.: A survey on collaborative deep learning and privacy-preserving. In: 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), pp. 652–658 (2018)
12. Shokri, R., Stronati, M., Song, C., Shmatikov, V. Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 3–18 (2017). arXiv:1610.05820
13. Dupuy, C., Arava, R., Gupta, R., Rumshisky, A.: An efficient DP-SGD mechanism for large scale NLU models, vol. 2022-May, pp. 4118–4122 (2022). https://aws.amazon.com/ec2/instance-types/. arXiv:2107.14586
14. Wunderlich, D., Bernau, D., Aldà, F., Parra-Arnau, J., Strufe, T.: On the privacy-utility trade-off in differentially private hierarchical text classification. Appl. Sci. **12**(21), 11177 (2022). https://doi.org/10.3390/app122111177
15. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Pro-

ceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, vol. 2015–Oct, pp. 1322–1333 (ACM, New York, NY, USA, 2015)
16. Hitaj, B., Ateniese, G., Perez-Cruz, F.: Deep models under the GAN: information leakage from collaborative deep learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 603–618 (2017)
17. Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., Song, D.: The secret sharer: evaluating and testing unintended memorization in neural networks. In: 28th USENIX Security Symposium (USENIX Security 19), pp. 267–284 (2019). arXiv:1802.08232
18. Coavoux, M., Narayan, S., Cohen, S.B.: Privacy-preserving neural representations of text. pp. 1–10 (2020). arXiv:1808.09408
19. Milli, S., Dragan, A.D., Schmidt, L., Hardt, M.: Model reconstruction from model explanations. In: FAT* 2019-Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 1–9 (2019). arXiv:1807.05185
20. Tramèr, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T.: Stealing machine learning models via prediction apis. In: SEC'16, pp. 601–618. USENIX Association, USA (2016)
21. European Parliament & Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. https://data.europa.eu/eli/reg/2016/679/oj
22. Abadi, M., et al.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (ACM, 2016). https://doi.org/10.1145/2976749.2978318
23. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. (2016)
24. Reddi, S., et al.: Adaptive federated optimization (2021). arXiv:2003.00295
25. McMahan, H.B., Ramage, D., Talwar, K., Zhang, L.: Learning differentially private recurrent language models. (2017)
26. Mercier, D., Lucieri, A., Munir, M., Dengel, A., Ahmed, S.: Evaluating privacy-preserving machine learning in critical infrastructures: a case study on time-series classification. IEEE Trans. Ind. Inf. **18**, 7834–7842 (2021)
27. Mohassel, P., Zhang, Y.: Secureml: a system for scalable privacy-preserving machine learning. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 19–38 (2017)
28. Knott, B., et al.: Crypten: secure multi-party computation meets machine learning (2022). arXiv:2109.00984
29. Kaissis, G., et al.: End-to-end privacy preserving deep learning on multi-institutional medical imaging. Nat. Mach. Intell. **3**(6), 473–484 (2021)
30. Li, X., Tramer, F., Liang, P., Hashimoto, T.: Large language models can be strong differentially private learners (2022). https://openreview.net/forum?id=bVuP3ltATMz
31. Plant, R., Gkatzia, D., Giuffrida, V.: CAPE: context-aware private embeddings for private language learning. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 7970–7978 (2021). https://aclanthology.org/2021.emnlp-main.628
32. Meehan, C., Mrini, K., Chaudhuri, K.: Sentence-level privacy for document embeddings. vol. 1, 3367–3380 (Association for Computational Linguistics, Stroudsburg, PA, USA, 2022). https://aclanthology.org/2022.acl-long.238. arXiv:2205.04605
33. Das, A., Roy, S., Bhattacharya, U., Parui, S.K.: Document image classification with intra-domain transfer learning and stacked generalization of deep convolutional neural networks. In: 2018 24th international conference on pattern recognition (ICPR), 2018-Augus, pp. 3180–3185 (2018). arXiv:1801.09321
34. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 2015-Novem, pp. 991–995 (2015). arXiv:1502.07058
35. Saifullah, S., Agne, S., Dengel, A., Ahmed, S.: Docxclassifier: high performance explainable deep network for document image classification (2022)
36. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: LayoutLMv3: pre-training for document AI with unified text and image masking. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 4083–4091 (ACM, New York, NY, USA, 2022). https://dl.acm.org/doi/10.1145/3503161.3548112. arXiv:2204.08387
37. Kumar, J., Ye, P., Doermann, D.: Learning document structure for retrieval and classification. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), pp. 1558–1561 (2012)
38. Diligenti, M., Frasconi, P., Gori, M.: Hidden tree Markov models for document image classification. IEEE Trans. Pattern Anal. Mach. Intell. **25**(4), 519–523 (2003)
39. Wang, B., Gong, N.Z.: Stealing hyperparameters in machine learning (2019). arXiv:1802.05351
40. Pawar, A., Ahirrao, S., Churi, P.P.: Anonymization Techniques for Protecting Privacy: A Survey. Institute of Electrical and Electronics Engineers Inc., Piscataway (2018)
41. Dwork, C.: Differential Privacy. vol. 4052 LNCS, pp. 1–12, Springer, Berlin (2006)
42. Chen, X., Wu, S.Z., Hong, M.: Understanding gradient clipping in private SGD: a geometric perspective. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 13773–13782. Curran Associates Inc., New York (2020)
43. Dengel, A., Dubiel, F.: Clustering and classification of document structure-a machine learning approach. In: Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 2, pp. 587–591 (1995)
44. Baldi, S., Marinai, S., Soda, G.: Using tree-grammars for training set expansion in page classification. In: Seventh International Conference on Document Analysis and Recognition, 2003-Janua (Icdar), pp. 829–833 (2003)
45. Chen, N., Blostein, D.: A survey of document image classification: problem statement, classifier architecture and performance evaluation. Int. J. Doc. Anal. Recognit. **10**(1), 1–16 (2007)
46. Asim, M.N., et al.: Two stream deep network for document image classification. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1410–1416. (2019)
47. Audebert, N., Herold, C., Slimani, K., Vidal, C.: Multimodal deep networks for text and image-based document classification. In: Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, vol. 1167 CCIS, pp. 427–443, Springer, Cham (2020). arXiv:1907.06370
48. Saifullah, Siddiqui, S. A., Agne, S., Dengel, A., Ahmed, S.: Are deep models robust against real distortions? A case study on document image classification. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 1628–1635 (2022)
49. Siddiqui, S.A., Dengel, A., Ahmed, S.: Analyzing the potential of zero-shot recognition for document image classification. In: International Conference on Document Analysis and Recognition, pp. 293–304, Springer-Verlag, Berlin, Heidelberg (2021). https://doi.org/10.1007/978-3-030-86337-1_20
50. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale (2021). https://openreview.net/forum?id=YicbFdNTTy
51. Li, J., et al.: Dit: self-supervised pre-training for document image transformer. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 3530–3539 (Association for Computing Machinery, New York, NY. USA (2022). https://doi.org/10.1145/3503161.3547911

52. Xu, Y., et al.: LayoutLMv2: multi-modal pre-training for visually-rich document understanding. pp. 2579–2591 (Association for Computational Linguistics, Online, 2021). https://aclanthology.org/2021.acl-long.201
53. Koskela, A., Tobaben, M., Honkela, A.: Individual privacy accounting with gaussian differential privacy (2022). arXiv:2209.15596
54. Mironov, I.: Rényi differential privacy (IEEE, 2017). https://doi.org/10.1109%2Fcsf.2017.11
55. Gopi, S., Lee, Y.T., Wutschitz, L.: Numerical composition of differential privacy. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems, vol. 34, pp. 11631–11642. Curran Associates, Inc., New York (2021)
56. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR. abs/1412.6980 (2014). https://api.semanticscholar.org/CorpusID:6628106
57. Yousefpour, A., et al.: Opacus: user-friendly differential privacy library in PyTorch. arXiv preprint arXiv:2109.12298 (2021)
58. Beutel, D.J., et al.: Flower: a friendly federated learning research framework. arXiv preprint arXiv:2007.14390 (2020)
59. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems, vol. 25. Curran Associates, Inc., New York (2012)
60. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR. abs/1409.1556 (2014). https://api.semanticscholar.org/CorpusID:14124313
61. Tan, M., Le, Q.: Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019)
62. Deng, J., et al.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
63. Kerrigan, G., Slack, D., Tuyls, J.: Differentially private language models benefit from public pre-training. In: Feyisetan, O., Ghanavati, S., Malmasi, S., Thaine, P. (eds.) Proceedings of the Second Workshop on Privacy in NLP, pp. 39–45 (Association for Computational Linguistics, Online, 2020). https://aclanthology.org/2020.privatenlp-1.5
64. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**, 1189–1232 (2001)