CAFE: Explainable Corruption-Aware Federated Aggregation for Robust Document Understanding

Muhammad Rafsan Kabir, Md Shopon, and Marina L. Gavrilova

Department of Computer Science, University of Calgary, Canada
{rafsan.kabir, md.shopon, mgavrilo}@ucalgary.ca

Abstract. Privacy-preserving document analysis systems deployed across institutions often rely on federated learning (FL) to collaboratively train models without sharing sensitive data. However, document images collected across institutions suffer from heterogeneous visual corruptions such as shadows, blur, and noise, introducing non-IID client environments that degrade the performance of standard FL methods. Existing reliability-aware aggregation strategies partially address client heterogeneity but typically estimate client reliability solely from clean validation performance, failing to capture robustness under corruption-induced distribution shifts. Furthermore, most methods evaluate client quality primarily through predictive accuracy, without assessing whether learned representations remain stable under document degradation. In this work, we propose **CAFE**, an explainable corruption-aware federated aggregation framework for robust document classification. CAFE introduces a dual-track reliability estimation strategy that jointly models clean validation performance and corruption robustness using both clean and corrupted server-held validation sets. The resulting reliability scores guide greedy client selection and reliability-weighted aggregation, effectively mitigating negative transfer from locally corrupted clients while preserving data privacy. Experiments on the RVL-CDIP benchmark across multiple CNN and transformer models demonstrate that CAFE consistently outperforms existing FL baselines. Furthermore, Grad-CAM-based explainability analysis shows that CAFE produces more spatially consistent and semantically stable patterns under document corruptions, achieving improved cosine similarity and IoU over competing methods, establishing its effectiveness for federated document understanding.

Keywords: Federated Learning · Document Understanding · Privacy-Preserving Learning · Trustworthy Document AI · Explainable AI

1 Introduction

Document analysis and understanding have become essential components of modern intelligent information systems, with applications spanning enterprise automation, healthcare records, financial document processing, legal archives, and administrative workflows [5]. Recent advances in deep learning have significantly improved document classification, enabling models to learn rich structural, semantic, and layout-aware representations from large-scale document

datasets [17]. However, many document understanding applications involve highly sensitive and confidential information, making centralized data collection undesirable or infeasible due to privacy and institutional constraints.

Federated learning (FL) [12] has emerged as a promising solution for privacy-preserving collaborative learning by enabling distributed institutions to jointly train machine learning models without sharing raw client data. This paradigm is particularly attractive for document intelligence systems deployed across enterprises, healthcare organizations, financial institutions, and government agencies, where document privacy and data ownership are critical concerns [19]. However, applying federated learning to document understanding tasks remains highly challenging in practice. Real-world documents are often collected under highly heterogeneous conditions. Scanned pages, photographed documents, mobile captures, and archival records may suffer from shadows, blur, noise, and uneven illumination [2], resulting in substantial data distribution shifts [11, 13, 18]. These factors produce highly non-identically distributed (non-IID) client environments, leading to degraded performance and unstable generalization.

Conventional federated learning methods such as FedAvg [12] assume all client updates to be equally reliable and therefore aggregate client models uniformly. In practice, client models are often not equally reliable due to heterogeneous document distributions and corruptions across institutions. As a result, some clients may overfit to locally corrupted data, producing biased updates that negatively affect the global model. Recent FL approaches have focused on improving optimization stability, correcting client drift, or mitigating statistical heterogeneity through robust aggregation and reliability-aware weighting strategies [7,9,10,14]. While these methods improve federated optimization, they exhibit important limitations for real-world document analysis systems. First, many aggregation strategies estimate client reliability using only clean validation performance, which may not adequately capture robustness under corruption-induced distribution shifts. Second, most methods evaluate client reliability using predictive accuracy, without assessing whether learned representations remain stable and interpretable under document degradation. These limitations are particularly critical for trustworthy document AI systems, where robustness, interpretability, and stability are essential for real-world deployment [1].

To address these challenges, we propose **CAFE** (**C**orruption-**A**ware **F**ederated document understanding with **E**xplainability), an explainable corruption-aware reliability-guided federated aggregation framework for robust document classification under heterogeneous document degradation settings. The proposed framework introduces a dual-track reliability estimation strategy that jointly models clean validation performance and corruption robustness using both clean and corrupted server-held validation distributions. Specifically, we introduce a corruption consistency formulation that evaluates how well a client model preserves predictive behavior under document degradation shifts. The resulting reliability scores are employed for greedy client selection and reliability-weighted aggregation, enabling the server to prioritize more robust client updates. Additionally, we incorporate Grad-CAM-based explainability analysis to investigate

whether corruption-aware aggregation leads to more stable and semantically meaningful representations under document degradation. Attribution stability between clean and corrupted document inputs is quantified using cosine similarity and Intersection-over-Union (IoU), providing insight into the trustworthiness of learned representations. We evaluate CAFE on the RVL-CDIP document classification benchmark [5] across multiple CNN and transformer-based architectures. Experimental results demonstrate that our proposed method consistently improves classification performance while producing more spatially consistent and stable attribution patterns under document corruptions.

The main contributions of this work are as follows:

- We propose CAFE, an explainable corruption-aware reliability-guided federated aggregation framework for robust document classification under heterogeneous settings, which remains privacy-preserving and architecture-agnostic.
- We introduce a dual-track corruption-aware reliability formulation that jointly models clean validation performance and corruption robustness, enabling more reliable client selection and weighting during federated aggregation.
- We conduct extensive experiments on the RVL-CDIP benchmark demonstrating that CAFE consistently outperforms existing federated learning baselines across multiple CNN and transformer-based architectures
- We perform a robustness-aware explainability analysis that evaluates the stability of Grad-CAM explanations under document corruptions, showing that the proposed framework produces more spatially consistent and stable attribution patterns compared to existing baselines.

2 Related Work

Initial research in federated learning (FL) focused on enabling collaborative model training across distributed clients without requiring the exchange of raw data. This made FL particularly attractive for privacy-sensitive domains where data are institutionally distributed, including healthcare, finance, education, and document analysis [8,12]. The most widely adopted baseline, Federated Averaging (FedAvg), aggregates locally trained client models using data-size-weighted averaging [12]. While FedAvg provides a simple and communication-efficient solution, its performance can degrade under non-IID client distributions, where local models may overfit to client-specific data characteristics and introduce biased updates into the global model. This limitation becomes especially important in document understanding, where different institutions may collect documents using different scanners, cameras, compression pipelines, or archival procedures.

Later works explored optimization and aggregation strategies to reduce the impact of client heterogeneity in FL. FedProx introduced proximal regularization to limit local client drift, while SCAFFOLD used control variates to correct drift caused by heterogeneous local updates [9,10]. Other aggregation approaches aimed to reduce the influence of unreliable client updates during server-side aggregation [14]. Although these methods improve federated optimization under heterogeneous conditions, they mainly address general statistical heterogeneity

rather than domain-specific document degradation. This distinction is important because a client model may perform well on clean data while failing under corruptions such as blur, shadow, low resolution, or compression artifacts.

In parallel, document image classification has become a vital task in document understanding, supporting applications such as digital archiving, administrative automation, and enterprise information processing. The RVL-CDIP benchmark has been widely used for large-scale document classification across categories such as letters, forms, emails, invoices, and reports [5]. Deep convolutional and transformer-based models have improved classification performance by learning visual layout, structural, and texture cues from scanned document images. However, real-world documents are often scanned, photographed, compressed, or affected by uneven illumination and noise. Prior work on degraded document restoration has shown that such visual degradations can significantly affect recognition performance [2,6]. In federated settings, this problem becomes more complex because different clients may contain different corruption patterns, causing standard aggregation to combine updates with unequal reliability. Recent works has explored model averaging and reliability-guided aggregation as a way to improve generalization. Model souping demonstrated that averaging the weights of multiple fine-tuned models can improve accuracy without increasing inference cost [16]. Building on this idea, ReSoFed introduced reliability-guided model aggregation for federated learning by selecting client models based on validation performance and applying reliability-weighted aggregation [7]. However, ReSoFed formulation estimates reliability mainly from clean validation performance, which may not be sufficient for document classification under corruption shifts. Moreover, predictive accuracy alone does not explain whether a model relies on stable and meaningful document regions; therefore, explainable AI can be used to analyze attribution consistency under corrupted inputs [15].

As seen from the above discussion, existing FL methods support privacy-preserving distributed training and partially address non-IID heterogeneity through improved optimization, client correction, and robust aggregation. However, limited attention has been given to federated document classification where each client may contain distinct corruption patterns. The key limitation of current FL approaches is that reliability-guided aggregation mainly estimates client usefulness from clean validation performance and does not explicitly account for corruption robustness or explanation stability. We address this gap by combining corruption-aware reliability estimation, greedy reliability-guided aggregation, and Grad-CAM-based attribution stability analysis within a unified framework for robust and trustworthy federated document understanding.

3 Methodology

3.1 Problem Formulation

Consider a federated learning setting with K distributed clients, where each client $k \in \{1, 2, \ldots, K\}$ possesses a local document dataset $D_k = \{(x_i^k, y_i^k)\}_{i=1}^{N_k}$,

where x_i^k denotes a document image and $y_i^k \in \{1, \ldots, C\}$ represents the corresponding document class label. Due to variations in scanning devices, capture conditions, and document-specific degradations such as noise and blur, the local client distributions are statistically heterogeneous and non-IID.

The objective of federated learning is to collaboratively optimize a global model parameterized by θ without sharing raw client data, formulated as:

$$\min_{\theta} F(\theta) = \sum_{k=1}^{K} \frac{N_k}{\sum_{j=1}^{K} N_j} F_k(\theta) \tag{1}$$

where $F_k(\theta)$ denotes the local optimization objective of client k, defined as:

$$F_k(\theta) = \frac{1}{N_k} \sum_{(x_i^k, y_i^k) \in D_k} \ell(\theta; x_i^k, y_i^k) \tag{2}$$

where $\ell(\cdot)$ represents the document classification loss function.

Solution Strategy. To improve federated robustness under heterogeneous document corruptions, we propose a dual-track corruption-aware reliability-guided aggregation strategy that estimates client reliability using both clean and corrupted validation distributions. The proposed framework prioritizes client models that exhibit stronger cross-domain generalization and corruption consistency, enabling more reliable global aggregation under distribution shifts.

3.2 Federated Learning Preliminaries

In standard federated learning, models are trained collaboratively across multiple distributed clients without exchanging raw data. At each communication round t, the server broadcasts the current global model parameters θ^t to all participating clients. Each client k performs local optimization on its private dataset D_k and updates the model parameters independently through local training. After local optimization, the updated client models are transmitted to the central server for aggregation. The most widely used aggregation strategy, Federated Averaging (FedAvg) [12], computes the global model by aggregating client updates proportionally according to the size of their local datasets.

Although FedAvg is effective under relatively homogeneous settings, it often suffers from degraded generalization under non-IID data distributions. In document understanding tasks, heterogeneous corruption patterns across clients can introduce biased local updates, causing unreliable client models to negatively influence the global aggregation process.

3.3 Reliability-Guided Federated Aggregation

Reliability-guided federated aggregation methods improve upon uniform averaging by estimating the generalization capability of each client model using a server-held validation set, enabling the server to prioritize reliable client updates and suppress weakly generalizing ones during aggregation [7].

Given locally trained client models θ_k, a reliability score r_k is computed as:

$$r_k = \mathrm{Acc}_{clean}(\theta_k) \tag{3}$$

where $\mathrm{Acc}_{clean}(\cdot)$ denotes classification accuracy on a clean server-held validation set. Clients are ranked by reliability and filtered through greedy soup-based selection [16], where models are iteratively included in the aggregation set only if their inclusion improves validation performance. The selected client subset S is then aggregated using reliability-proportional weighted averaging:

$$\theta_{global} = \sum_{k \in S} \alpha_k \theta_k, \quad \alpha_k = \frac{r_k}{\sum_{j \in S} r_j} \tag{4}$$

While this formulation effectively reduces negative transfer from weakly generalizing clients, estimating reliability solely from clean validation performance is insufficient in document classification settings where corruption-induced distribution shifts are a primary source of client heterogeneity. A client model that performs well on clean validation data may still produce unstable and unreliable representations under corrupted document inputs, motivating a more comprehensive reliability formulation that explicitly incorporates corruption robustness.

3.4 Proposed CAFE

Existing reliability-guided federated aggregation strategies primarily estimate client reliability using clean validation performance, which may not adequately capture model robustness under document corruption-induced distribution shifts. In practical document understanding systems, client data often contain degradation artifacts such as shadows, blur, noise, and compression artifacts, resulting in heterogeneous corruption patterns across distributed institutions. Consequently, a client model achieving high performance on clean validation data may still exhibit unstable performance when evaluated on corrupted document inputs. To address this limitation, we propose CAFE, an explainable corruption-aware federated aggregation framework designed for cross-silo FL environments with server-side proxy validation data [8], explicitly incorporating corruption robustness into client reliability estimation for trustworthy document classification. An overview of the proposed CAFE framework is illustrated in Fig. 1.

Let A^k_{clean} and $A^k_{corrupt}$ denote the validation accuracies of client model k on clean and corrupted validation sets, respectively. We define the corruption robustness score R_k as:

$$R_k = A^k_{corrupt} \cdot \left(\frac{A^k_{corrupt}}{A^k_{clean}} \right) \tag{5}$$

The ratio term $\frac{A^k_{corrupt}}{A^k_{clean}}$ measures corruption consistency by quantifying the performance degradation between clean and corrupted distributions. Client models exhibiting smaller degradation receive higher robustness scores. The objective

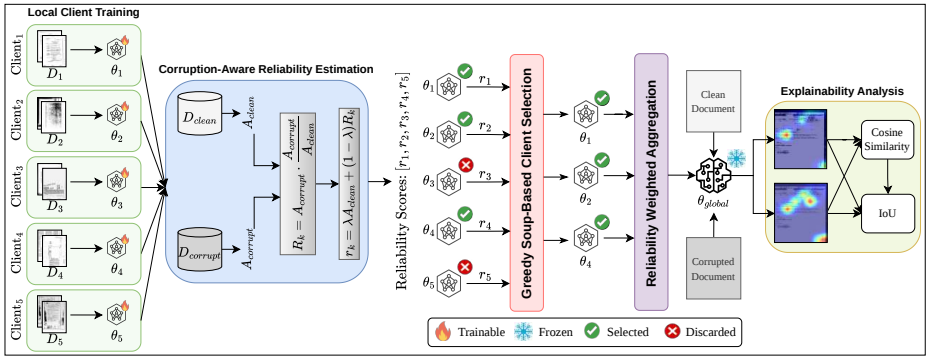


Fig. 1: Overview of the proposed CAFE framework. Each client is trained locally on private document datasets under heterogeneous document degradations. The server performs corruption-aware reliability estimation using both clean (D_{clean}) and corrupted ($D_{corrupt}$) validation sets to compute reliability scores for client selection and aggregation. Based on these scores, clients are filtered through greedy client selection, followed by reliability-weighted aggregation to construct the global model θ_{global}. Finally, explainability analysis is performed using Grad-CAM on clean and corrupted document inputs, where attribution stability is evaluated using cosine similarity and Intersection-over-Union (IoU) to assess robustness and representation consistency under document degradations.

of the robustness score is not to measure absolute predictive quality alone, but rather to prioritize client models that maintain strong performance under corruption while exhibiting limited degradation relative to their clean-data performance. Furthermore, the squared corruption accuracy term places greater emphasis on performance under corrupted conditions, while the consistency ratio penalizes models that experience substantial degradation between clean and corrupted distributions. The final reliability score r_k is computed as:

$$r_k = \lambda A^k_{clean} + (1 - \lambda) R_k \tag{6}$$

where $\lambda \in [0, 1]$ controls the trade-off between clean validation accuracy and corruption robustness.

Clients are first ranked according to their reliability scores and filtered using greedy soup-based client selection [16]. The selected client subset S is subsequently aggregated using reliability-aware weighted averaging:

$$\theta_{global} = \sum_{k \in S} \alpha_k \theta_k, \quad \alpha_k = \frac{r_k}{\sum_{j \in S} r_j} \tag{7}$$

where θ_k denotes the locally trained model parameters of client k, θ_{global} represents the aggregated global model, and α_k denotes the reliability-aware aggregation weight assigned to client k.

In addition to predictive performance evaluation, we perform explainability-driven robustness analysis using Grad-CAM to investigate attribution stability under document corruptions; details are presented in Section 4. Overall, CAFE addresses the limitations of existing reliability-guided federated aggregation methods by explicitly incorporating corruption robustness into client reliability estimation. This enables the framework to better identify robust client models, reduce negative transfer from corrupted clients, and learn more stable and trustworthy representations under heterogeneous document distributions.

4 Explainability Analysis

Beyond predictive performance, it is important to investigate whether federated aggregation strategies learn stable and semantically meaningful representations under document corruptions. In document understanding tasks, corrupted inputs may significantly alter model attention patterns, causing explanations to become spatially inconsistent even when classification predictions remain correct. Therefore, evaluating explanation stability provides valuable insight into the trustworthiness of models under heterogeneous document degradations.

To analyze explanation consistency, we employ Grad-CAM [15], which generates visual attribution maps by highlighting image regions that contribute most strongly to model predictions. Given a clean document image and its corresponding corrupted counterpart, Grad-CAM explanations are generated for both inputs across different federated learning methods, including FedAvg [12], existing reliability-aware federated aggregation (ReSoFed) [7], and the proposed CAFE framework. These attribution maps are compared to evaluate whether the learned representations remain stable under corruption shifts.

Grad-CAM Attribution Maps. Grad-CAM computes localization maps using the gradients of the target class score with respect to the feature maps of the final convolutional layer. Let A^k denote the k-th feature map and y^c represent the score corresponding to class c. The Grad-CAM map is computed as:

$$\text{Activation Map: } L^c_{GradCAM} = \text{ReLU}\left(\sum_k w^c_k A^k\right) \tag{8}$$

where w^c_k denotes the importance weight of feature map A^k.

The generated Grad-CAM heatmaps highlight the discriminative regions used by the model for document classification. By comparing explanations generated from clean and its corresponding corrupted document images, we assess whether corruption-induced distribution shifts alter the spatial attribution behavior of the learned representations.

Attribution Stability Metrics. To quantitatively evaluate attribution stability and explanation consistency under corruption, we compute similarity between Grad-CAM heatmaps generated from clean and corresponding corrupted document images using cosine similarity and Intersection-over-Union (IoU).

Cosine similarity measures the similarity between Grad-CAM attribution vectors. Let M_{clean} and $M_{corrupt}$ denote the Grad-CAM maps corresponding to

clean and corrupted document inputs, respectively. The cosine similarity score is computed as:

$$\text{CosSim}(M_{clean}, M_{corrupt}) = \frac{M_{clean} \cdot M_{corrupt}}{\|M_{clean}\| \|M_{corrupt}\|} \tag{9}$$

Higher cosine similarity values indicate that the model preserves more consistent attribution patterns under corruption.

To further evaluate spatial overlap between attribution regions, we compute the Intersection-over-Union (IoU) between thresholded Grad-CAM masks derived from clean and corrupted inputs. Grad-CAM heatmaps are first normalized to the range [0,1] and binarized using a threshold of $\tau = 0.5$. The same threshold is applied across all methods and experiments to ensure a fair comparison.

$$\text{IoU} = \frac{|M_{clean} \cap M_{corrupt}|}{|M_{clean} \cup M_{corrupt}|} \tag{10}$$

Higher IoU values indicate stronger spatial consistency between clean and corrupted explanations, suggesting that the model relies on more stable and semantically meaningful document regions despite corruption-induced perturbations.

Through this explainability-driven analysis, we investigate whether corruption-aware reliability-guided aggregation leads not only to improved predictive performance but also to more trustworthy and interpretable representation learning under heterogeneous document degradation settings.

5 Experiments

5.1 Setup

Dataset. We evaluate CAFE on the RVL-CDIP [5] benchmark, a large-scale document image classification dataset containing 16 classes, including letters, forms, emails, invoices, reports, resumes, memos, and other document types. Each image is assigned a single class label, making the task a multi-class document classification problem. The original RVL-CDIP split contains 320,000 training images and 40,000 test images. To reduce computational cost while preserving class diversity, we use 70% of the original training set, resulting in 224,000 training samples. From the original test set, 20,000 samples are used as the final centralized test set, while the remaining 20,000 samples are kept as an unused validation pool for constructing server-held clean and corrupted validation sets. All splits are stratified across the 16 document classes.

Federated Client Simulation. To simulate a federated document classification setting, the 224,000 training samples are distributed across five clients, where each client represents an independent institution or document source. The partitioning is performed in a class-balanced manner, with each client receiving approximately 44,800 samples and roughly 2,800 samples per class.

Client 1 is kept clean and serves as the reference document source. Clients 2–5 are assigned distinct document-specific corruptions to simulate heterogeneous

Table 1: Federated client simulation with document-specific corruption settings.

Clients	Corruption Type	Parameters
Client 1	Clean documents	No corruption
Client 2	Gaussian blur	Radius $r = 2.0$
Client 3	Shadow/illumination variation	Polygon opacity $\alpha = 0.45$
Client 4	Missing-region occlusion	15–25% masked area
Client 5	Compression + noise	JPEG quality $q = 30$, Gaussian noise $\sigma = 10$

acquisition conditions. For corrupted clients, 90% of the local training images are transformed using the assigned corruption, while the remaining 10% are kept clean. Specifically, Client 2 applies Gaussian blur with radius $r = 2.0$; Client 3 applies shadow/illumination variation by overlaying a semi-transparent dark polygon with opacity $\alpha = 0.45$; Client 4 applies missing-region occlusion by masking one randomly selected rectangular region covering approximately 15–25% of the image area; and Client 5 applies compression/noise degradation using JPEG compression with quality $q = 30$ followed by additive Gaussian noise with standard deviation $\sigma = 10$. These corruption choices are motivated by standard robustness benchmarks, where blur, noise, brightness variation, and compression artifacts are commonly used to evaluate distribution shifts [6]. Shadow and illumination variations are also common in captured document images [4], while missing-region occlusion is motivated by occlusion-based augmentation methods such as Cutout and Random Erasing [3, 20].

The server-held validation set is constructed from the unused validation pool and is used only for reliability estimation and client selection. CAFE utilizes both clean and corrupted validation sets to evaluate whether client models remain accurate and stable under document degradations. The corrupted validation set is generated using the same corruption families employed in the client simulation. Table 1 presents the simulated clients along with their corruption types.

Implementation Details. Experiments are conducted in a simulated federated learning setting consisting of five distributed clients. Each client is trained under heterogeneous corruption settings to simulate realistic document acquisition variations across institutions and devices. Specifically, each client is assigned distinct document-specific corruption configurations, including shadow artifacts, blur, noise, and compression degradations, introducing non-IID visual distributions across clients. The proposed CAFE framework is evaluated across multiple backbone architectures, including two convolutional neural networks (CNNs), ResNet-50 and ConvNeXt-Base, and two transformer-based architectures, ViT B-16 and Swin-V2-Base. ImageNet-pretrained versions of all models are employed and fine-tuned for document classification on the RVL-CDIP dataset. Local client optimization is performed using the Adam optimizer with a learning rate of 1×10^{-4}. Each client independently performs local training on its private dataset without sharing raw document images. During aggregation, a server-held clean and corrupted validation set is utilized for corruption-aware

Table 2: Performance comparison of FedAvg, ReSoFed, and the proposed CAFE on the RVL-CDIP document classification benchmark. Results are reported across CNN and transformer-based backbone architectures. **Bold** values indicate the best performance per metric.

Method	Model	Architecture	Acc (%)	Pre (%)	Rec (%)	F1 (%)
FedAvg [12]	ResNet-50	CNN	79.24	77.75	79.12	78.43
	ConvNeXt-Base	CNN	85.27	86.15	85.24	85.44
	ViT B-16	Transformer	80.44	81.86	80.43	81.13
	Swin-V2-Base	Transformer	83.16	83.86	83.13	83.25
ReSoFed [7]	ResNet-50	CNN	84.78	84.97	84.74	84.76
	ConvNeXt-Base	CNN	86.57	87.03	86.54	86.65
	ViT B-16	Transformer	81.14	81.75	81.10	81.28
	Swin-V2-Base	Transformer	86.05	86.26	86.01	86.08
CAFE (Proposed)	ResNet-50	CNN	84.94	85.20	84.91	84.99
	ConvNeXt-Base	CNN	**86.72**	**87.72**	**86.68**	**86.92**
	ViT B-16	Transformer	82.55	82.82	82.51	82.68
	Swin-V2-Base	Transformer	86.43	86.71	86.39	86.53

reliability estimation and greedy soup-based client selection. All experiments are implemented using *PyTorch* and conducted on an NVIDIA RTX 3090 GPU. The hyperparameter λ is set to 0.5 to equally weight clean validation accuracy and corruption robustness during reliability estimation.

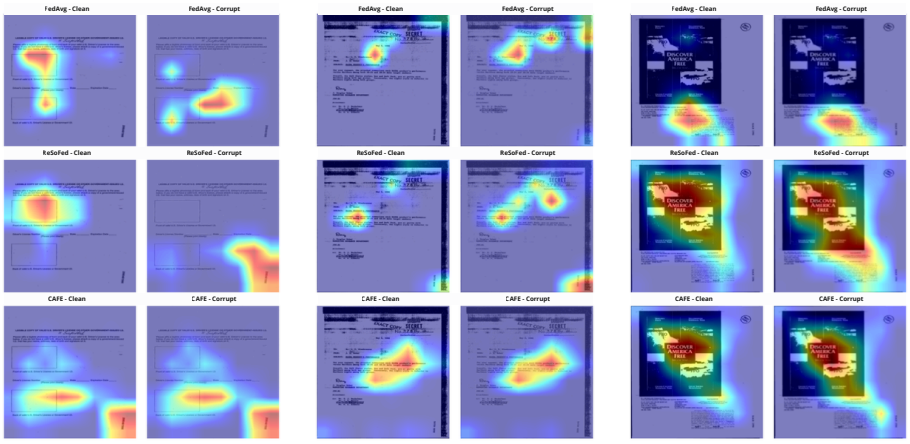
Evaluation Metrics. To evaluate classification performance, we report standard classification metrics including Accuracy, Precision, Recall, and F1-score. These metrics are computed on the clean centralized test set to assess the generalization capability of the aggregated global model.

In addition to classification performance, we evaluate attribution stability under document corruptions using explainability-based metrics. Specifically, Grad-CAM explanations generated from clean and corresponding corrupted document images are compared using cosine similarity and Intersection-over-Union (IoU). Cosine similarity measures directional consistency between attribution maps, while IoU evaluates spatial overlap between salient regions. Higher values indicate stronger explanation consistency and stability under corruption shifts.

5.2 Results and Analysis

Classification Performance. Table 2 presents the document classification performance of Federated Averaging (FedAvg), existing reliability-guided federated aggregation (ReSoFed), and the proposed CAFE framework on the RVL-CDIP benchmark. Experiments are conducted across four backbone architectures, including CNNs and transformer-based models, to evaluate the effectiveness and architecture generalizability of the proposed strategy. Performance is evaluated on the centralized clean test set using Accuracy, Precision, Recall, and F1-score.

From Table 2, we observe the following: ***(1)*** The proposed CAFE framework consistently outperforms the standard federated learning baseline, FedAvg, across all architectures, demonstrating the effectiveness of corruption-aware reliability-guided aggregation under heterogeneous document settings. ***(2)*** Compared to



(a) Sample 1 (b) Sample 2 (c) Sample 3

Fig. 2: Grad-CAM visualizations comparing attribution stability across FedAvg, ReSoFed, and the proposed CAFE under clean and corrupted document inputs.

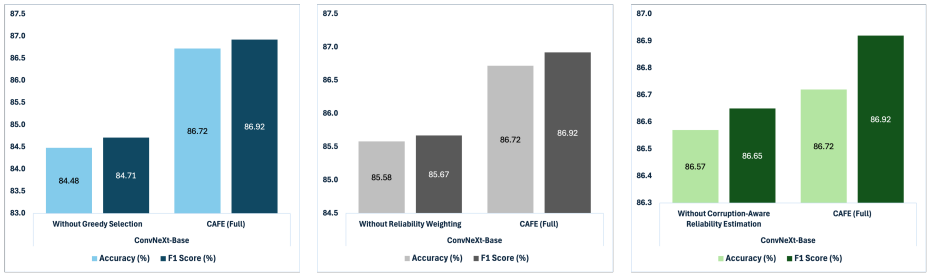
the ReSoFed framework, CAFE achieves consistent performance improvements across all models, indicating that incorporating corruption robustness into reliability estimation provides a more discriminative signal for client selection and weighted aggregation. ***(3)*** The performance gains are particularly noticeable for transformer-based models. For instance, ViT B-16 improves from 81.14% to 82.55% accuracy (+1.41%), while Swin-V2-Base improves from 86.05% to 86.43% (+0.38%), highlighting the effectiveness of the proposed method under distribution shifts. ***(4)*** Among all evaluated models, ConvNeXt-Base combined with CAFE achieves the best overall performance, reaching 86.72% accuracy and 86.92% F1-score, suggesting that modern CNN models strongly benefit from corruption-aware reliability-guided federated aggregation under corrupted document environments. ***(5)*** Overall, the consistent improvements across all architectures and evaluation metrics confirm that CAFE is architecture-agnostic, generalizing effectively across both convolutional and transformer-based models. This validates the proposed framework as a practical strategy for privacy-preserving federated document classification under real-world heterogeneous conditions.

Explainability Results. In addition to document classification performance, we analyze the robustness and interpretability using Grad-CAM-based explainability analysis. Specifically, we evaluate attribution stability between clean and corrupted document image pairs across FedAvg, ReSoFed, and CAFE using ConvNeXt-Base, the best-performing backbone, to assess the spatial consistency of learned representations under document corruptions.

The qualitative Grad-CAM results in Fig. 2 show that FedAvg produces unstable and spatially shifted attention under corrupted inputs, while ReSoFed improves consistency by filtering weaker client updates. In contrast, CAFE gen-

Table 3: Grad-CAM attribution stability comparison across federated learning methods using cosine similarity and IoU.

Method	Cosine Similarity ↑	IoU ↑
FedAvg	0.3677	0.0677
ReSoFed	0.5145	0.1510
CAFE	**0.5346**	**0.1634**



(a) Impact of greedy soup-based client selection

(b) Impact of reliability-weighted aggregation

(c) Impact of corruption-aware reliability estimation

Fig. 3: Ablation study of the CAFE framework using ConvNeXt-Base, the best-performing model. **(a)** Impact of greedy client selection. **(b)** Impact of reliability-weighted aggregation. **(c)** Impact of corruption-aware reliability estimation.

erates more coherent attribution maps across clean and corrupted document pairs, suggesting that corruption-aware reliability estimation improves attention stability under document degradation.

The quantitative results in Table 3 support the qualitative observations. CAFE achieves the highest attribution stability, with a cosine similarity of 0.5346 and an IoU of 0.1634, showing stronger consistency between clean and corrupted inputs. Although the IoU values are relatively modest, the consistent gains over FedAvg and ReSoFed indicate that corruption-aware reliability estimation improves explanation stability under document degradation. Overall, the analysis shows that CAFE improves not only classification performance but also the stability and trustworthiness of visual representations in heterogeneous settings.

5.3 Ablation Study

Impact of Greedy Client Selection. To evaluate the importance of the greedy soup-based client selection strategy, we remove the greedy selection stage and aggregate all client models directly using reliability-weighted averaging. As shown in Fig. 3a, removing greedy client selection leads to noticeable degradation in both Accuracy and F1-score. This indicates that the greedy soup mechanism effectively filters weakly generalizing client updates and reduces negative transfer during aggregation under heterogeneous document corruption settings.

Impact of Reliability-Weighted Aggregation. Next, we investigate the contribution of reliability-aware weighted aggregation by replacing reliability-based weighting with uniform averaging among the selected client models. Fig. 3b demonstrates that the complete CAFE framework consistently achieves higher Accuracy and F1-score compared to uniform aggregation. These results suggest that reliability-guided weighting effectively prioritizes client models with stronger cross-domain generalization during global aggregation.
Impact of Corruption-Aware Reliability Estimation. Finally, we evaluate the effect of the proposed corruption-aware reliability formulation by replacing it with a conventional reliability estimation strategy based solely on clean validation accuracy. As illustrated in Fig. 3c, incorporating corruption-aware reliability estimation further improves both Accuracy and F1-score. This demonstrates that jointly modeling clean validation performance and corruption consistency provides more reliable robustness estimation under document degradations.

6 Conclusion

In this work, we proposed CAFE, an explainable corruption-aware reliability-guided federated aggregation framework for robust document classification. Unlike conventional federated aggregation strategies that estimate client reliability solely using clean validation performance, CAFE explicitly incorporates corruption robustness by evaluating client models on both clean and corrupted server-held validation sets. By integrating corruption-aware reliability estimation, greedy soup-based client selection, and reliability-weighted aggregation, the proposed method effectively mitigates negative transfer arising from heterogeneous document corruptions. Extensive experiments on the RVL-CDIP benchmark demonstrate that the proposed framework consistently outperforms existing methods across all evaluated architectures. Furthermore, Grad-CAM-based explainability analysis demonstrates that CAFE produces more spatially consistent attribution patterns under document corruptions, confirming that corruption-aware reliability estimation improves not only predictive performance but also the trustworthiness of the global model.

For future work, we plan to extend CAFE to federated settings with naturally occurring client heterogeneity. We also aim to investigate adaptive reliability estimation strategies where the balancing parameter λ evolves dynamically across communication rounds in response to changing client distributions. Additionally, future research may explore the integration of multimodal document understanding and document foundation models within corruption-aware FL frameworks.

References

1. Anzum, F., Asha, A.Z., Dey, L., Gavrilov, A., Iffath, F., Ohi, A.Q., Pond, L., Shopon, M., Gavrilova, M.L.: A comprehensive review of trustworthy, ethical, and explainable computer vision advancements in online social media. Global Perspectives on the Applications of Computer Vision in Cybersecurity pp. 1–46 (2024)

2. Dai, Z., Luecke, J.: Autonomous document cleaning—a generative approach to reconstruct strongly corrupted scanned texts. IEEE TPAMI (2014)
3. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
4. Gatos, B., Pratikakis, I., Perantonis, S.J.: Adaptive degraded document image binarization. Pattern Recognition **39**(3), 317–327 (2006)
5. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. In: International Conference on Document Analysis and Recognition (ICDAR) (2015)
6. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (ICLR) (2019)
7. Kabir, M.R., Shopon, M., Gavrilova, M.L.: ReSoFed: reliability-guided model souping for robust federated learning in heterogeneous classroom environments. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2026)
8. Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., et al.: Advances and open problems in federated learning. Foundations and Trends in Machine Learning (2021)
9. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., Suresh, A.T.: Scaffold: Stochastic controlled averaging for federated learning. In: International Conference on Machine Learning (ICML) (2020)
10. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: Proceedings of MLSys (2020)
11. Marino, S., Vandoni, J., Aldea, E., Lemghari, I., Le Hégarat-Mascle, S., Jurie, F.: ICPR 2024 competition on domain adaptation and generalization for character classification (DAGECC). In: International Conference on Pattern Recognition (ICPR). pp. 161–172. Springer (2024)
12. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: International Conference on Artificial Intelligence and Statistics (AISTATS) (2017)
13. Pei, J., Liu, W., Li, J., Wang, L., Liu, C.: A review of federated learning methods in heterogeneous scenarios. IEEE Transactions on Consumer Electronics (2024)
14. Pillutla, K., Kakade, S.M., Harchaoui, Z.: Robust aggregation for federated learning. IEEE Transactions on Signal Processing **70**, 1142–1154 (2022)
15. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: IEEE International Conference on Computer Vision (ICCV). pp. 618–626 (2017)
16. Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., et al.: Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: International Conference on Machine Learning (ICML) (2022)
17. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: LayoutLM: pre-training of text and layout for document image understanding. In: ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1192–1200 (2020)
18. Ye, M., Fang, X., Du, B., Yuen, P.C., Tao, D.: Heterogeneous federated learning: State-of-the-art and research challenges. ACM Computing Surveys (2023)
19. Yin, X., Zhu, Y., Hu, J.: A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. ACM Computing Surveys (2021)
20. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI Conference on Artificial Intelligence (2020)