Integrating rPPG-based Physiological Cues into Federated Face Presentation Attack Detection via Stacking

Mehmet Fatih Gündoğar
Marmara University
İstanbul, Türkiye
mehmetfatihgundogar@gmail.com

Çiğdem Eroğlu Erdem
Özyeğin University
İstanbul, Türkiye
cigdem.erogluerdem@gmail.com

Ömer Korçak
Marmara University
İstanbul, Türkiye
omer.korcak@marmara.edu.tr

***Abstract*—Face presentation attack detection (FPAD) remains a critical challenge due to the significant domain shift between different recording environments and the privacy concerns associated with centralized data collection. Remote photoplethysmography (rPPG) is a non-contact physiological sensing technique that estimates heart-rate-related signals from facial videos by analyzing subtle skin color variations caused by blood volume changes, providing intrinsic liveness cues that are difficult to replicate by face spoofing attacks. Building on our previous Fed-StackFPAD framework, which relied on appearance-based features extracted using vision transformers (ViT) and stacking-based federated ensemble learning, this paper extends the framework by incorporating heart-rate-driven spatio-temporal features derived from rPPG signals. rPPG signals are extracted using both unsupervised and supervised methods, and discriminative time- and frequency-domain features are computed to capture physiologically implausible patterns introduced by face presentation attacks. These physiological cues are integrated into an extended stacking phase, where data-center-specific rPPG model outputs are fused with federated global and local appearance-based predictions to enhance cross-domain robustness. Experimental results across multiple benchmarks demonstrate that integrating heterogeneous physiological and appearance-based cues significantly improves detection accuracy and outperforms existing state-of-the-art (SOTA) federated FPAD methods. To ensure reproducibility and support further research, the source code and trained models are publicly available at https://github.com/mehmetfatihgundogar/rPPG-Enhanced-Fed-StackFPAD.**

I. Introduction

Face recognition and verification systems have become integral to various modern applications, ranging from secure building access and mobile device authentication to automated passport control at borders. However, these systems remain inherently vulnerable to face presentation attacks, where unauthorized individuals attempt to gain access using spoofing instruments such as printed photographs, video replays, or wearable masks (see Fig. 1). Consequently, face presentation attack detection (FPAD) has emerged as a critical safeguard to ensure the reliability and security of biometric authentication.

Despite significant progress in deep learning-based FPAD, most models struggle with domain shift—the performance degradation that occurs when a model trained on one dataset is deployed in an environment with different illumination, subject

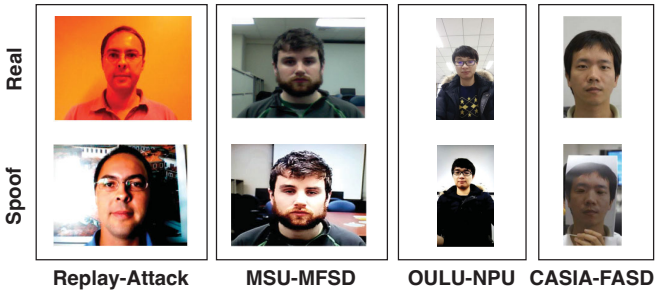

Fig. 1. Example face presentation attack (spoof) samples from Replay-Attack [1], MSU-MFSD [2], OULU-NPU [3], and CASIA-FASD [4] datasets

ethnicities, or camera resolutions. While centralizing data from multiple sources can improve diversity and robustness, this approach is often hindered by stringent privacy regulations and high computational costs. Federated learning (FL) offers a privacy-preserving alternative by allowing multiple data centers to train a global model collaboratively without sharing raw data. In our previous work, we introduced Fed-StackFPAD [5], a framework that combined a self-supervised vision transformer (ViT), specifically a masked autoencoder (MAE)-based ViT model (ViT-MAE) [6], with stacking-based ensemble learning to mitigate data heterogeneity. By fusing the predictions of a federated global model with data-center-specific local models, we achieved a significant reduction in the average half total error rate (HTER) compared to state-of-the-art (SOTA) federated methods.

While Fed-StackFPAD effectively captured spatial appearance-based features, relying solely on texture and reflectance can be insufficient against sophisticated attack instruments that mimic visual details. To address this limitation, physiological signals such as remote photoplethysmography (rPPG) provide a promising complementary modality. rPPG is a non-contact technique that estimates cardiovascular activity by detecting subtle, periodic color variations in facial skin caused by the cardiac cycle. Unlike genuine presentations, face spoofing attempts, particularly printed photos and video replays, lack authentic

physiological blood volume pulse (BVP) dynamics or exhibit inconsistent frequency responses due to display artifacts and motion noise.

In this paper, we propose rPPG-Enhanced Fed-StackFPAD, an extension of our federated framework that integrates rPPG-based physiological cues into the stacking process. We employ different supervised and unsupervised rPPG extraction methods to recover pulse-related signals. By extracting discriminative features from the time and frequency domains, we identify physiologically implausible patterns characteristic of presentation attacks. These cues are then utilized to train data-center-specific temporal models, whose outputs are incorporated into an extended stacking matrix alongside the ViT-based models.

The primary contributions of this work are as follows:

- We introduce rPPG-Enhanced Fed-StackFPAD, the first federated FPAD framework that enhances appearance-based ViT representations with rPPG-based physiological models via an extended stacking mechanism.
- We provide a comprehensive analysis of time and frequency domain rPPG features for distinguishing genuine physiological rhythms from spoofing presentations.
- We demonstrate through experimental evaluations that the fusion of spatial and physiological cues significantly improves cross-dataset generalization, outperforming existing SOTA federated methods.

The remainder of this paper is organized as follows: Section II provides a comprehensive literature review covering federated learning in FPAD and rPPG-based liveness detection for the FPAD problem. Section III details the proposed rPPG-Enhanced Fed-StackFPAD methodology, including the details of rPPG-based features and the extended stacking mechanism. Section IV presents the experimental setup and quantitative results. Section V provides a discussion on the complementary roles of physiological and appearance-based cues, along with the challenges of temporal modeling in FPAD. Finally, Section VI concludes the paper with key insights and future directions.

II. Related work

This section reviews existing literature on FPAD, with a particular focus on (i) FL–based FPAD frameworks, (ii) rPPG-based physiological liveness cues proposed for FPAD, and (iii) stacking-based ensemble strategies applied to improve FPAD robustness and generalization.

A. Federated learning in FPAD

Traditional deep learning-based FPAD models often suffer from performance degradation when deployed in unseen environments, a phenomenon known as domain shift. FL addresses this challenge by enabling decentralized model training across multiple data centers, preserving privacy by exchanging model weights rather than raw biometric data. While several optimization strategies like federated averaging (FedAvg) [7] exist, their application in the FPAD domain is relatively recent. FedPAD [8] pioneered FL for FPAD but focused primarily on basic optimization without explicitly addressing severe domain heterogeneity. Subsequent frameworks, such as FedGPAD [9], attempted to disentangle domain-specific and domain-invariant features but struggled with the extreme variance in real-world data distributions. Our previous work, Fed-StackFPAD, demonstrated that integrating stacking-based model aggregation within an FL framework significantly lowers the average HTER by mitigating the non-IID nature of distributed datasets. However, the Fed-StackFPAD framework exclusively utilizes ViT models that operate on spatial data and does not exploit useful temporal cues that could be derived from inter-frame variations for robust presentation attack detection. To bridge this gap and leverage dynamic information, the following section explores rPPG as an intrinsic physiological modality that provides the necessary temporal cues for robust liveness detection.

B. rPPG-based FPAD methods

rPPG has gained significant attention as an intrinsic liveness cue because it recovers cardiovascular signals from subtle skin color variations, which are physically blocked by spoofing instruments like 3D masks and printed photos. Early rPPG-based methods primarily relied on frequency-domain analysis, requiring long observation windows (10–12 seconds) to identify a stable heartbeat peak, which limits their usability in real-time applications.

To address this limitation, in [10], the authors proposed the temporal similarity of rPPG (TSrPPG) feature, which operates on a much shorter 1-second window by analyzing the rPPG waveform in the time domain. This approach leverages the principle that local rPPG signals from different facial regions of a genuine subject exhibit high similarity in terms of amplitude, gradient, and phase, whereas signals from masked faces are characterized by unstable environmental noise. To further enhance discriminability, a self-supervised local consistent learning strategy was introduced in [10], utilizing a decorrelation loss to explicitly reduce the correlation between masked faces and genuine physiological signals. While this method proved highly effective on 3D mask datasets, it also demonstrated that rPPG signals from background regions can serve as an effective reference for identifying non-genuine presentations.

Complementing these temporal cues, in [11], a motion-robust dual-stream network was introduced that combines spatiotemporal rPPG features with multiscale central difference convolutional (CDC) texture descriptors. Recognizing that rPPG signals are highly susceptible to head movements, a phase-driven attention mechanism was designed in [11] to guide the network to focus on facial regions with minimal motion interference by calculating phase differences between successive frames. However, both studies in [10] and [11] primarily focus on 3D mask presentation attacks, and their performance has not been extensively validated against sophisticated video replay attacks, which remain a significant evaluation gap.

Recent advancements have also focused on the security and evaluation standards of rPPG-based systems. In [12], the authors introduced the OR-PAD, the first FPAD dataset

to include synchronized physiological ground-truth (PPG and ECG) for real access recordings. This study exposed critical vulnerabilities by demonstrating how rPPG signal injection attacks, performed via modulated illumination or artificial pixel variations, can mimic genuine cardiac rhythms to bypass liveness detection. These findings highlight the need for robust evaluation across diverse rPPG methods and datasets beyond the specific 3D mask scenarios.

Despite these advancements, rPPG-only approaches remain highly susceptible to environmental noise and hardware artifacts, making them insufficient to fully overcome the challenges of domain shift and data heterogeneity in FPAD. Relying solely on physiological signals often leads to performance degradation when models encounter unseen acquisition conditions or novel attack instruments. To address these limitations, stacking-based ensemble methods provide a powerful alternative by leveraging the complementary decision boundaries of diverse models. By integrating appearance-based spatial representations with physiological temporal cues, stacking can achieve superior generalization, as explored in the next section.

C. Stacking-based ensemble learning

The application of stacking-based ensemble learning in the face presentation attack detection (FPAD) domain is relatively limited, particularly in centralized settings. In [13], authors addressed photo, video replay, and 3D mask attacks by leveraging a multi-modal approach that combined motion-based features, texture-based descriptors, and rPPG-based physiological features. These diverse representations were processed through a cascade of support vector machine (SVM) classifiers, where the final decision was made by a stacked meta-classifier. Similarly, in [14], a stacking-based ensemble framework was proposed to improve generalization by training separate base models on various spatiotemporal variants of the original data. This approach allowed the meta-classifier to exploit the complementary characteristics of the base models, enhancing the ability to learn diverse and discriminative representations.

Despite the potential of stacking to capture specialized decision surfaces, its adoption within privacy-preserving federated frameworks remained unexplored until the introduction of Fed-StackFPAD. Our previous work established that integrating a lightweight meta-classifier to fuse positive-class probabilities from a federated global model and data-center-specific appearance models significantly reduces the HTER by mitigating the impact of non-IID data distributions. However, as noted in the previous sections, this architecture relied primarily on spatial representations.

To address this limitation and leverage the synergy between appearance and physiology, the following section introduces the rPPG-Enhanced Fed-StackFPAD framework, which extends the stacking mechanism to incorporate temporal liveness cues derived from rPPG.

III. Proposed Methodology: rPPG-Enhanced Fed-StackFPAD

This section presents the proposed rPPG-Enhanced Fed-StackFPAD framework, which extends our previous federated stacking approach [5] by incorporating physiological rPPG cues as complementary temporal information for robust FPAD process.

A. rPPG background

rPPG is a non-contact physiological signal measurement technique that estimates cardiovascular activity directly from standard RGB video recordings. Unlike traditional contact-based photoplethysmography (PPG), rPPG extracts pulse-related information by analyzing subtle skin color variations observed on the facial surface. The underlying principle is based on the cardiac cycle: as blood volume in superficial vessels increases during systole, light absorption—particularly in the green wavelength—increases due to the optical properties of hemoglobin [15]. Conversely, reflectance increases during diastole as blood volume decreases. These periodic fluctuations manifest as a temporal signal known as the blood volume pulse (BVP). In genuine videos, it is expected that rPPG signals exhibit quasi-periodic and spatially coherent patterns across facial regions [10].

B. rPPG signal extraction methods

To capture a diverse range of physiological cues, we employ three distinct rPPG extraction approaches. All extraction methods are implemented using the rPPG-Toolbox framework [16]. An overview of the rPPG signal extraction pipeline, including classical and supervised neural approaches employed in this study, is illustrated in Fig. 2.

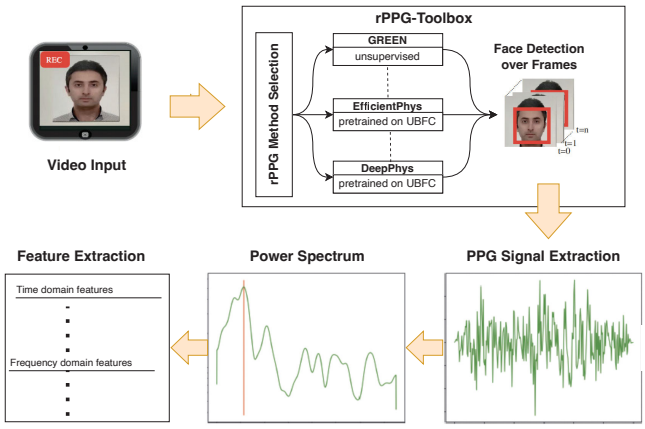

Fig. 2. Overview of the rPPG signal extraction workflow utilizing the rPPG-Toolbox [16]. Supervised neural models (EfficientPhys [17], DeepPhys [18]) utilize weights pretrained on UBFC-rPPG [19] to recover BVP signals in the absence of physiological labels in FPAD datasets

1) GREEN method (unsupervised): The GREEN method [15] is a classical unsupervised baseline that exploits the higher sensitivity of the green color channel to blood volume

variations. For each video frame, the spatially averaged green-channel intensity is extracted from a defined facial region Ω, forming the raw temporal signal $s(t)$:

$$s(t) = \frac{1}{|\Omega|} \sum_{(x,y)\in\Omega} G_t(x, y) \tag{1}$$

where $G_t(x, y)$ represents the green-channel intensity at time t. The resulting signal $s(t)$ has a length of T samples, where T corresponds to the number of video frames in the analyzed temporal window, which typically spans 5–10 seconds at the native frame rate, depending on the dataset-specific clip length and frame rate. This method requires no prior training and serves as a robust baseline for detecting the absence of authentic physiological rhythms.

2) EfficientPhys and DeepPhys (supervised): In addition to the unsupervised GREEN baseline, we employ supervised deep neural network architectures to estimate rPPG signals from facial videos. In this context, the output of each supervised model is treated as a temporal rPPG signal $s(t)$, analogous to the raw signal extracted by the GREEN method, but obtained through learned spatio-temporal representations.

EfficientPhys [17] is a lightweight convolutional neural network (CNN) designed for end-to-end BVP estimation, which implicitly models motion suppression and illumination normalization. DeepPhys [18] utilizes a convolutional attention-based architecture that explicitly decomposes the estimation process into an appearance stream capturing static skin properties and a motion stream modeling temporal variations, with an attention mechanism to suppress non-physiological disturbances.

A critical constraint in applying supervised rPPG signal extraction methods to FPAD is that standard FPAD benchmark datasets do not provide synchronized ground-truth physiological signals. Consequently, we employed pretrained models available in the rPPG-Toolbox framework, which were trained on the UBFC-rPPG dataset [19], and no additional fine-tuning was performed on FPAD data.

C. Signal preprocessing

All extracted temporal rPPG signals $s(t)$ obtained using the GREEN, EfficientPhys, and DeepPhys extraction methods undergo a standardized preprocessing pipeline to ensure comparability and to suppress environmental and acquisition-related artifacts:

1) **Linear detrending:** Applied to the raw signals to remove slow-varying illumination drifts.
2) **Band-pass filtering:** Signals are filtered within the physiological heart-rate range of 0.75–2.5 Hz (45–150 beats per minute (bpm)) to suppress high-frequency sensor noise and low-frequency motion artifacts.

After preprocessing, the resulting signal is denoted as $x(t)$ and represents the cleaned rPPG time series used for subsequent feature extraction.

D. Feature extraction from rPPG signals

Following preprocessing, discriminative features are computed from the preprocessed rPPG signal $x(t)$ in both the time and frequency domains to identify physiologically implausible patterns characteristic of presentation attacks.

1) Time-domain statistical features: Let $x(t)$ denote the preprocessed rPPG signal. The following statistical moments are computed to quantify signal variability, asymmetry, and peakedness [20], [21]:

$$\mu = \mathbb{E}[x(t)] \tag{2}$$

$$\sigma = \sqrt{\mathbb{E}[(x(t) - \mu)^2]} \tag{3}$$

$$\text{skewness} = \mathbb{E}\left[\left(\frac{x(t) - \mu}{\sigma}\right)^3\right] \tag{4}$$

$$\text{kurtosis} = \mathbb{E}\left[\left(\frac{x(t) - \mu}{\sigma}\right)^4\right] \tag{5}$$

2) Frequency-domain analysis: To characterize periodicity, the power spectral density (PSD) is estimated via a periodogram:

$$P(f) = \frac{1}{N}\left|\sum_{t=1}^{N} x(t)e^{-j2\pi ft}\right|^2 \tag{6}$$

From the resulting PSD, the dominant frequency (f_{peak}) is identified as the peak within the [0.75, 2.5] Hz band. We further compute band power ratios to measure energy concentration:

$$R_{\text{low}} = \frac{\int_{0.75}^{1.5} P(f)df}{\int_{0.75}^{2.5} P(f)df}, \quad R_{\text{high}} = \frac{\int_{1.5}^{2.5} P(f)df}{\int_{0.75}^{2.5} P(f)df} \tag{7}$$

Finally, the spectral flatness (Wiener entropy) is computed to distinguish tonal, physiological signals from noise-like behavior:

$$\text{SF} = \frac{\exp\left(\frac{1}{N}\sum_f \log P(f)\right)}{\frac{1}{N}\sum_f P(f)} \tag{8}$$

It is often observed that bona fide facial videos exhibit a dominant spectral component in the physiological frequency range, whereas face presentation attacks may yield flatter or multi-peaked spectra due to non-physiological artifacts, as illustrated in Fig. 3. These physiological tendencies motivate the extraction of frequency- and shape-based rPPG features for FPAD. While such spectral patterns are not guaranteed to hold under all acquisition conditions, they provide useful discriminative cues, as face presentation attacks often introduce non-physiological periodicities, frequency smearing, or flattened spectra that deviate from plausible cardiac rhythms.

E. Extended stacking with physiological cues

The proposed rPPG-Enhanced Fed-StackFPAD framework augments our previous federated stacking process [5] by incorporating data-center-specific rPPG models. For each data center k, a temporal SVM classifier M_k^{temp} is trained using the extracted rPPG features.

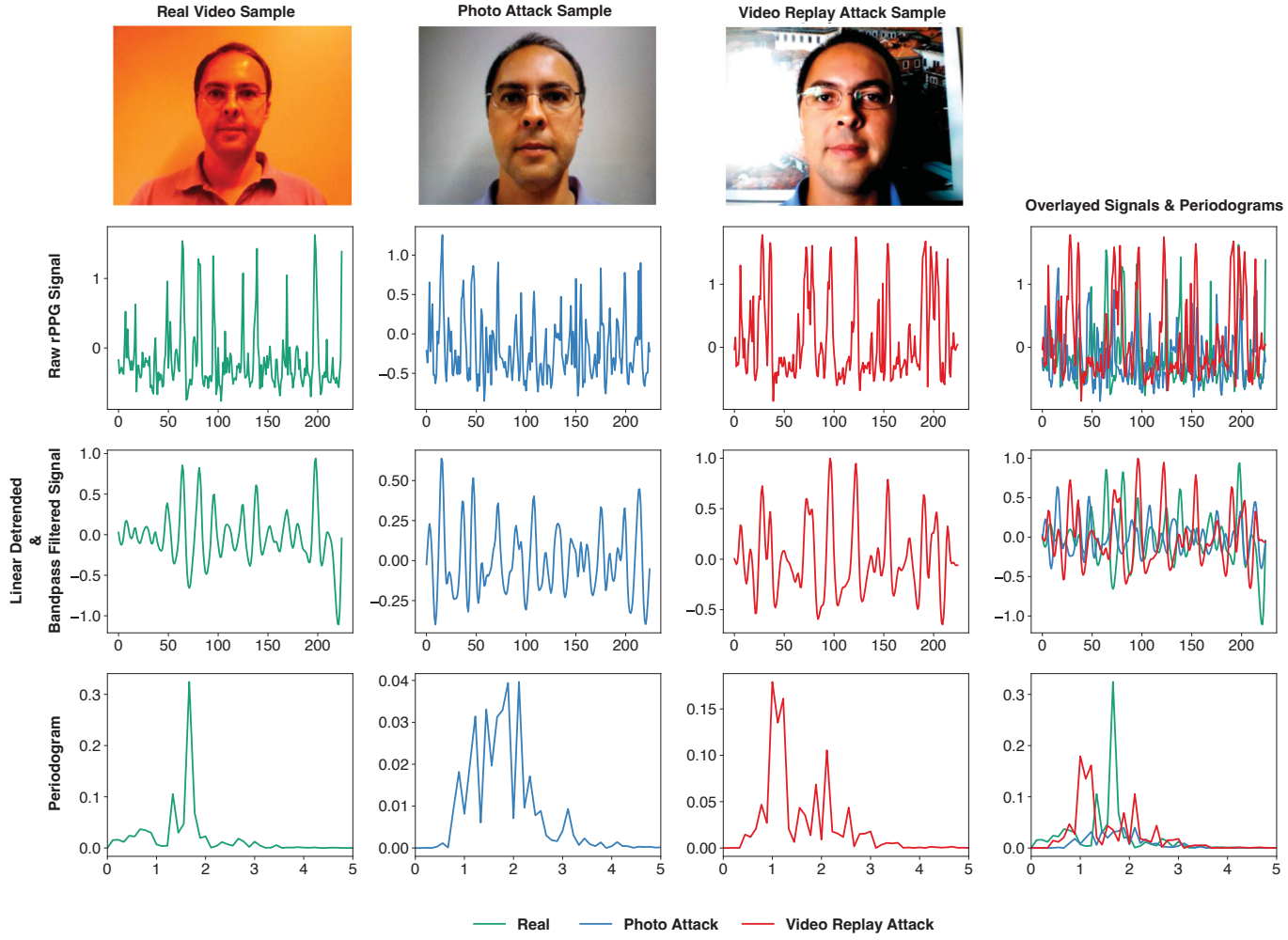

Fig. 3. Comparison of spectral characteristics. Real access video samples exhibit a single narrow-band peak, while presentation attacks yield flattened or multi-peaked spectra

During the stacking phase, prediction probabilities from the federated global model (M_G), the local appearance models (M_k^{ViT}), and the rPPG temporal models (M_k^{temp}) are concatenated column-wise to form an extended feature matrix $\mathcal{F}_{\text{ext}} \in \mathbb{R}^{m\times(2N+1)}$:

$$\mathcal{F}_{\text{ext}} = \begin{bmatrix} p^1_{\text{ViT},1} & p^1_{\text{temp},1} & \cdots & p^N_{\text{ViT},1} & p^N_{\text{temp},1} & p^{\text{global}}_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ p^1_{\text{ViT},m} & p^1_{\text{temp},m} & \cdots & p^N_{\text{ViT},m} & p^N_{\text{temp},m} & p^{\text{global}}_m \end{bmatrix}, \tag{9}$$

where

- m denotes the number of validation set samples at the federated central server,
- N is the number of participating data centers,
- $p^k_{\text{ViT},i}$ is the predicted bona fide probability for the i-th sample produced by the *local ViT-based appearance model* at data center k,
- $p^k_{\text{temp},i}$ denotes the corresponding probability predicted by the *rPPG-based local temporal model* at data center k,
- p^{global}_i represents the probability output of the *federated global appearance model* M_G for the i-th sample.

As illustrated in Fig. 4, prediction probabilities from the federated global model, local appearance models, and rPPG-based temporal classifiers are concatenated to form an extended feature representation for meta-level classification. A final SVM meta-classifier is trained on $\mathcal{F}_{\text{ext}}$ to exploit the complementary nature of spatial appearance cues and temporal physiological liveness information.

IV. Experimental Study

This section presents a comprehensive experimental evaluation of the proposed rPPG-Enhanced Fed-StackFPAD framework. We analyze the effectiveness of rPPG-based temporal cues both as standalone predictors and as auxiliary experts within a stacking-based federated learning setup. All experiments are conducted under strictly defined intra- and cross-dataset protocols to assess robustness against domain shifts.

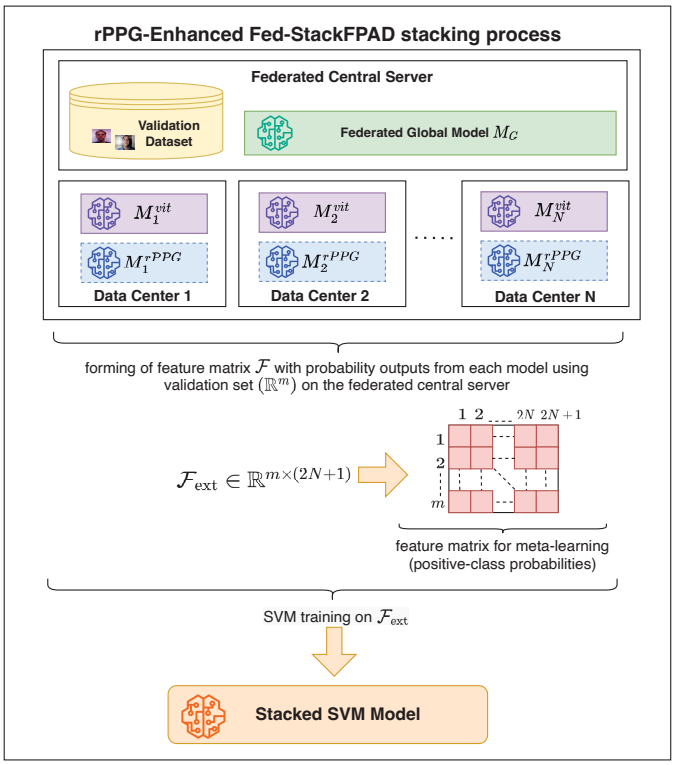

Fig. 4. Overview of the rPPG-enhanced stacking mechanism in the proposed Fed-StackFPAD framework. Each data center trains a local ViT-based appearance model and an rPPG-based temporal classifier, whose probability outputs are combined with the federated global model prediction and fed into a meta-level SVM for final decision making

A. Datasets

Experiments are conducted on four publicly available and widely adopted face presentation attack detection benchmarks. Key dataset characteristics, including the number of subjects, videos, attack instruments, and capture resolutions, are summarized below.

- **Idiap Replay-Attack** [1] consists of 50 subjects and 1,300 video samples. Videos are captured under both controlled and adverse illumination conditions using a fixed webcam setup, with a resolution of 320×240. The attack scenarios involve printed photographs and replayed videos displayed on electronic screens.
- **MSU-MFSD** [2] contains 35 subjects and 280 video samples acquired using different capture devices, including laptops and mobile phones. The dataset includes both photo and video replay attacks recorded under varying illumination and image quality conditions. Video resolutions range from 640×480 to 720×480, introducing moderate device- and quality-related domain shifts.
- **OULU-NPU** [3] is a large-scale benchmark comprising 55 subjects and 4,950 video samples. The dataset is designed to explicitly evaluate generalization and includes systematic variations in illumination conditions, background scenes, and camera devices. Videos are recorded at six different resolutions.

TABLE I. INTRA-DATASET FPAD PERFORMANCE OF RPPG-BASED MODELS

Dataset	GREEN		EfficientPhys		DeepPhys	
	EER(%)	HTER(%)	EER(%)	HTER(%)	EER(%)	HTER(%)
MSU-MFSD	40.00	–	42.92	–	37.08	–
Replay-Attack	0.00	1.25	30.50	29.75	17.83	13.12
OULU-NPU	23.01	24.51	33.59	35.88	30.65	35.17
CASIA-FASD	29.63	–	38.69	–	32.57	–
Average	23.16	12.88	36.43	32.82	29.53	24.15

- **CASIA-FASD** [4] comprises 50 subjects and 600 video samples recorded at two different spatial resolutions (640×480 and 1920×1080). The dataset includes a diverse set of presentation attack instruments, such as printed photos, warped photos, cut photos, and video replays. These samples are captured under varying image quality levels and environmental conditions, leading to substantial intra-dataset variability and pronounced domain shifts.

Together, these datasets cover a wide range of capture settings, attack instruments, and acquisition conditions, providing a rigorous testbed for evaluating the robustness and generalization capability of federated FPAD frameworks.

B. Evaluation metrics

Performance is evaluated using standard FPAD metrics. The equal error rate (EER) corresponds to the operating point where the false acceptance rate equals the false rejection rate. When a validation set is available, the decision threshold determined at the EER point is applied to the test set to compute the HTER. In addition, the area under the receiver operating characteristic (ROC) curve (AUC) is reported to measure threshold-independent discriminative capability. Lower values of EER and HTER and higher values of AUC indicate better performance.

C. Intra-Dataset evaluation with rPPG-based models

We first evaluate rPPG-based approaches under an intra-dataset protocol, where training and testing are performed on the same dataset. This setting assesses the standalone discriminative capability of rPPG cues for FPAD. The quantitative results of this evaluation are summarized in Table I.

We consider the unsupervised GREEN method as well as two supervised neural rPPG models (EfficientPhys and DeepPhys). For supervised models, no dataset-specific fine-tuning is performed, since standard FPAD benchmark datasets do not provide synchronized physiological ground-truth signals.

The results reveal a strong dataset dependency for rPPG-only models. While near-perfect performance is achieved on Replay-Attack with a HTER of 1.25%, generalization degrades substantially on more diverse datasets, indicating that rPPG cues alone are insufficient for robust and reliable FPAD and should be treated as complementary temporal information rather than standalone discriminative features.

D. Extended stacking with rPPG-based temporal models

We next evaluate the proposed rPPG-Enhanced Fed-StackFPAD framework, where rPPG-based temporal models

are incorporated as auxiliary experts within a stacking-based ensemble. In this setting, probability outputs from local ViT-based appearance models, rPPG-based temporal classifiers, and the federated global model are combined using a meta-level SVM. The quantitative results of this evaluation are reported in Table II.

Compared to rPPG classifiers trained on a single dataset (see Table I), FL with stacking using 3 different data centers consistently improves performance across all test datasets, demonstrating that rPPG cues provide useful complementary information when fused with appearance-based models.

E. Comparison with state-of-the-art federated FPAD methods

We further compare the proposed rPPG-Enhanced Fed-StackFPAD framework against existing SOTA federated FPAD approaches under cross-dataset evaluation protocols. The comparative performance results are summarized in Table III.

The results indicate that the proposed stacking-based federated framework consistently achieves lower error rates than existing federated FPAD methods. By incorporating rPPG-based temporal models as auxiliary experts, the framework effectively leverages complementary temporal information alongside appearance representations, leading to improved robustness under cross-dataset evaluation settings.

V. Discussion

The experimental results provide several important insights into the role of physiological temporal cues within federated FPAD. First, the intra-dataset evaluation of rPPG-only models confirms that physiological signals alone exhibit strong dataset dependency. As shown in Table I, near-perfect performance is achieved on Replay-Attack (HTER of 1.25% using the GREEN method), whereas error rates increase substantially on more diverse datasets such as OULU-NPU (HTER up to 35.88%) and CASIA-FASD, highlighting the sensitivity of rPPG extraction to capture conditions, illumination, and device characteristics.

When rPPG-based temporal models are integrated into the stacking-based federated framework, a consistent performance improvement is observed across all cross-dataset evaluation scenarios. As reported in Table II, the proposed rPPG-Enhanced Fed-StackFPAD achieves average HTER values in the range of 4.28–4.46% across different rPPG methods, representing a substantial reduction compared to standalone rPPG classifiers. These results indicate that even temporally weak or unstable predictors can contribute meaningful complementary information when fused with strong appearance-based models through stacking.

The comparison with SOTA federated FPAD methods further demonstrates the effectiveness of the proposed approach. As summarized in Table III, rPPG-Enhanced Fed-StackFPAD consistently outperforms classical federated baselines such as FedMA, FedPAD, and FedGPAD, particularly under severe cross-dataset shifts. For instance, in the O&C&M→I protocol, the proposed method achieves an HTER of 2.40% and an EER of 0.00%, outperforming existing federated approaches.

An important limitation revealed by these experiments concerns the lack of synchronized ground-truth physiological signals in standard FPAD datasets. Supervised rPPG models employed in this study rely on weights pretrained on generic rPPG datasets (e.g., UBFC-rPPG) and cannot be adapted to FPAD-specific characteristics. The availability of FPAD datasets with reliable ground-truth PPG signals would enable task-specific fine-tuning and potentially improve both standalone rPPG performance and its contribution within the stacking framework. Recent efforts such as OR-PAD [12] suggest that this is a promising but still restricted research direction.

Overall, the results indicate that rPPG cues should not be treated as a replacement for appearance-based FPAD models. Instead, their strengths are best realized when incorporated as auxiliary experts within a federated stacking architecture, where heterogeneous cues can be combined without raw data sharing.

VI. Conclusion

In this work, we investigated the integration of physiological temporal cues into federated FPAD by extending the Fed-StackFPAD framework with rPPG-based temporal models. Comprehensive experiments conducted on four widely used FPAD benchmarks demonstrate that rPPG signals, while insufficient as standalone predictors, provide complementary discriminative information when incorporated into a stacking-based federated learning framework.

The proposed rPPG-Enhanced Fed-StackFPAD consistently improves robustness under cross-dataset evaluation protocols, achieving lower HTER and EER values compared to existing SOTA federated FPAD methods. These results highlight the effectiveness of stacking-based ensemble learning in mitigating data heterogeneity across distributed data centers while preserving privacy.

From a broader perspective, the findings of this study suggest that temporal information should be exploited as a complementary modality rather than a direct substitute for appearance-based representations. The lack of ground-truth physiological annotations in current FPAD datasets remains a key limitation, and future datasets providing synchronized PPG signals would enable more effective supervision and adaptation of rPPG models for FPAD.

As future work, this framework can be further extended by incorporating richer temporal representations, particularly motion-centric cues such as optical flow and micro-motion analysis. Such features are less sensitive to physiological variability and have shown strong discriminative power in replay and video-based attack scenarios. Combining rPPG, optical flow, and potentially depth-based temporal cues within the proposed federated stacking architecture represents a promising direction for advancing robust and scalable FPAD method.

Acknowledgment

This work was supported by The Scientific and Technological Research Council of Türkiye (TÜBİTAK) under projects EEAG-122E025 and ARDEB 1002 [Grant no.: 124E292].

TABLE II. PERFORMANCE OF RPPG-ENHANCED FED-STACKFPAD WITH DATA CENTER-SPECIFIC RPPG MODELS. (O&C&I→M DENOTES THAT OULU-NPU, CASIA-FASD, AND MSU-MFSD ARE USED FOR TRAINING AT DATA CENTERS, WHILE REPLAY-ATTACK IS USED AS THE UNSEEN TEST DATASET AT THE SERVER)

	EfficientPhys			**GREEN**			**DeepPhys**		
Dataset	HTER(%)	EER(%)	AUC(%)	HTER(%)	EER(%)	AUC(%)	HTER(%)	EER(%)	AUC(%)
O&C&I→M	3.74	2.84	96.25	3.68	1.46	96.31	3.63	3.23	96.36
O&M&I→C	6.69	6.65	93.30	6.95	6.39	93.04	6.62	6.08	93.37
O&C&M→I	2.50	2.02	97.49	2.40	0.00	97.60	3.09	1.60	96.90
I&C&M→O	4.29	3.17	95.70	4.07	3.19	95.92	4.50	3.48	95.49
Average	4.31	3.67	95.69	**4.28**	**2.76**	**95.72**	4.46	3.60	95.53

TABLE III. COMPARISON WITH STATE-OF-THE-ART FEDERATED FPAD FRAMEWORKS. (O&C&I→M DENOTES THAT OULU-NPU, CASIA-FASD, AND MSU-MFSD ARE USED FOR TRAINING AT DATA CENTERS, WHILE REPLAY-ATTACK IS USED AS THE UNSEEN TEST DATASET AT THE SERVER)

Method	O&C&I→M			O&M&I→C			O&C&M→I			I&C&M→O		
	HTER(%)	EER(%)	AUC(%)	HTER(%)	EER(%)	AUC(%)	HTER(%)	EER(%)	AUC(%)	HTER(%)	EER(%)	AUC(%)
FedMA [9], [22]	29.68	25.85	84.68	31.17	30.78	78.81	24.82	26.75	73.05	30.95	25.15	84.28
FedPAD [8]	19.45	17.43	90.24	42.27	36.95	70.49	32.53	26.54	73.58	34.44	34.45	71.74
FedGPAD [9]	12.73	13.36	91.25	28.69	27.55	80.58	10.97	11.11	95.34	21.95	17.91	89.85
Fed-StackFPAD [5]	4.07	2.12	95.92	7.91	**4.90**	92.08	2.92	3.43	97.07	4.50	3.25	95.49
rPPG-Enhanced Fed-StackFPAD (*GREEN method*)	**3.68**	**1.46**	**96.31**	**6.95**	6.39	**93.04**	**2.40**	**0.00**	**97.60**	**4.07**	**3.19**	**95.92**

The authors acknowledge the use of AI-assisted tools based on LLMs for grammar and writing refinement in certain parts of this manuscript. The sentences rephrased by the AI were reviewed and edited by the authors to ensure their accuracy and relevance.

REFERENCES

[1] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group (BIOSIG)*. IEEE, 8 2012. [Online]. Available: https://ieeexplore.ieee.org/document/6313548

[2] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, pp. 746–761, 4 2015. [Online]. Available: http://ieeexplore.ieee.org/document/7031384/

[3] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "Oulu-npu: A mobile face presentation attack database with real-world variations," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 5 2017, pp. 612–618. [Online]. Available: http://ieeexplore.ieee.org/document/7961798/

[4] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *2012 5th IAPR International Conference on Biometrics (ICB)*. IEEE, 3 2012, pp. 26–31. [Online]. Available: http://ieeexplore.ieee.org/document/6199754/

[5] M. F. Gündoğar, Ç. E. Erdem, and Ö. Korçak, "Fed-stackfpad: Federated learning for face presentation attack detection with stacking to tackle data heterogeneity," *IEEE Access*, vol. 13, pp. 190 354–190 370, 2025.

[6] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[7] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282, 2 2016. [Online]. Available: http://arxiv.org/abs/1602.05629

[8] R. Shao, P. Perera, P. C. Yuen, and V. M. Patel, "Federated face presentation attack detection," *arXiv preprint arXiv:2005.14638*, 5 2020. [Online]. Available: http://arxiv.org/abs/2005.14638

[9] ——, "Federated generalized face presentation attack detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, pp. 103–116, 1 2024. [Online]. Available: https://ieeexplore.ieee.org/document/9780603/

[10] S.-Q. Liu, X. Lan, and P. C. Yuen, "Learning temporal similarity of remote photoplethysmography for fast 3d mask face presentation attack detection," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3195–3210, 2022.

[11] R. Sun, X. Yu, H. Feng, F. Wang, and X. Zhang, "Motion-robust mask face presentation attack detection via dual-stream texture-rppg network," *The Visual Computer*, vol. 41, no. 7, pp. 4517–4532, 2025.

[12] M. Savic and G. Zhao, "Oulu remote-photoplethysmography presentation attacks database (or-pad)," *International Journal of Computer Vision*, vol. 134, no. 1, p. 25, 2025. [Online]. Available: https://doi.org/10.1007/s11263-025-02588-z

[13] M. F. Gündoğar and Ç. E. Erdem, "Presentation attack detection for face recognition using remote photoplethysmography and cascaded fusion," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 29, pp. 3240–3258, 11 2021. [Online]. Available: https://journals.tubitak.gov.tr/elektrik/vol29/iss7/22

[14] U. Muhammad, J. Laaksonen, D. Romaissa Beddiar, and M. Oussalah, "Domain generalization via ensemble stacking for face presentation attack detection," *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5759–5782, 2024.

[15] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light." *Optics express*, vol. 16 26, pp. 21 434–45, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:7782171

[16] X. Liu, G. Narayanswamy, A. Paruchuri, X. Zhang, J. Tang, Y. Zhang, S. Sengupta, S. Patel, Y. Wang, and D. McDuff, "rppg-toolbox: Deep remote ppg toolbox," *Advances in Neural Information Processing Systems*, vol. 36, 2024. [Online]. Available: http://arxiv.org/abs/2210.00716

[17] X. Liu, B. Hill, Z. Jiang, S. Patel, and D. McDuff, "Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 4997–5006.

[18] W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 349–365.

[19] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognition Letters*, vol. 124, pp.

82–90, 2019, award Winning Papers from the 23rd International Conference on Pattern Recognition (ICPR). [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865517303860

[20] R. Krishnan, B. Natarajan, and S. Warren, "Two-stage approach for detection and reduction of motion artifacts in photoplethysmographic data," *IEEE transactions on biomedical engineering*, vol. 57, no. 8, pp. 1867–1876, 2010.

[21] N. Selvaraj, Y. Mendelson, K. H. Shelley, D. G. Silverman, and K. H. Chon, "Statistical approach for the detection of motion/noise artifacts in photoplethysmogram," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 4972–4975.

[22] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," *arXiv preprint arXiv:2002.06440*, 2020.